SEGMENTED CONFIDENCE SEQUENCES AND MULTI-SCALE ADAPTIVE CONFIDENCE SEGMENTS FOR ANOMALY DETECTION IN NONSTATIONARY TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

As time series data become increasingly prevalent in domains such as manufacturing, IT, and infrastructure monitoring, anomaly detection must adapt to nonstationary environments where statistical properties shift over time. Traditional static thresholds are easily rendered obsolete by regime shifts, concept drift, or multi-scale changes. To address these challenges, we introduce and empirically evaluate two novel adaptive thresholding frameworks: Segmented Confidence Sequences (SCS) and Multi-Scale Adaptive Confidence Segments (MACS). Both leverage statistical online learning and segmentation principles for local, contextually sensitive adaptation, maintaining guarantees on false alarm rates even under evolving distributions. Our experiments across Wafer Manufacturing benchmark datasets show significant F1-score improvement compared to traditional percentile and rolling quantile approaches. This work demonstrates that robust, statistically principled adaptive thresholds enable reliable, interpretable, and timely detection of diverse real-world anomalies.

1 INTRODUCTION

Time series data are ubiquitous across modern applications, from industrial process monitoring and predictive maintenance to financial markets and sensor-driven systems. Detecting anomalies—unusual patterns or behaviors that deviate from expected trends—is crucial for preventing faults, reducing risk, and ensuring operational reliability (Chandola et al., 2009). Unlike static datasets, time series often exhibit evolving behavior, including trends, seasonality, and abrupt regime shifts, making anomaly detection a particularly challenging problem.

In recent years, researchers have developed advanced techniques that go beyond simple static thresholds. Approaches such as robust moving windows, online quantile estimation, and confidence sequence theory have emerged to provide more adaptive and statistically principled anomaly detection (Howard et al., 2021; Wang et al., 2023). These methods aim to balance computational efficiency with real-time adaptability, enabling detection systems to respond to changing data dynamics.

However, existing adaptive thresholding methods often struggle when data exhibit multiple temporal scales or sudden regime shifts. Fixed-window or global percentile-based strategies may either fail to capture local variations, leading to missed anomalies, or produce excessive false positives when the baseline drifts (Benidis et al., 2022). This highlights the need for a thresholding framework that can simultaneously adapt to both abrupt and gradual changes in data distribution.

To address these challenges, we contribute two novel frameworks for adaptive thresholding:

- Segmented Confidence Sequences (SCS): segments time series by regime, maintaining distinct confidence-based bounds per segment to adapt to local statistics.
- Multi-Scale Adaptive Confidence Segments (MACS): adapts detection simultaneously at multiple window lengths, enabling the detection of both rapid bursts and slow regime changes.

- Comprehensive experiments supporting statistically significant improvements over traditional percentile or fixed adaptive thresholds.

2 Related Work

2.1 Static and Traditional Thresholding

Early approaches used fixed global thresholds—often prescribed as mean $\pm\, k\sigma$ or a static quantile—assuming stationarity and i.i.d. observations (Chandola et al., 2009). These methods fail under concept drift or dynamic variance and are prone to false positives in practical systems (Blázquez-García et al., 2021). Percentile-based approaches, such as the 99th percentile threshold, adjust for heavy tails but still falter under persistent distributional drift or nonstationarity (Genton et al., 2021). Extreme Value Theory (EVT) and Peak-Over-Threshold (POT) models empirical tails beyond a high threshold but assume quasi-stationarity (Genton et al., 2021).

2.2 Sliding Windows, Rolling Quantiles

Adaptive methods using sliding windows recalculate thresholds over a recent window—updating the mean or quantile in an online manner (Aggarwal, 2015). EWMA improves rapid adaptation to trends or regime switches, but window size determines sensitivity and is often hard to tune (Blázquez-García et al., 2021). Nonparametric models reduce reliance on distributional assumptions (Rousseeuw et al., 2020).

2.3 Model-Based and Learning Approaches

Forecasting-model-based detection fits ARIMA or seasonal decomposition, then tests for outliers in the residuals (Brockwell & Davis, 2016). Neural networks, autoencoders, and reinforcement learners learn context-sensitive anomaly scores but may lack explicit error guarantees or require extensive labeled data (Benidis et al., 2022; Ahmad et al., 2017; Xue et al., 2023).

2.4 Confidence Sequences for Online Adaptation

Confidence sequences (CS) give time-uniform intervals guaranteeing error rate control under arbitrary stopping (Howard et al., 2021). Recent algorithms maintain confidence bounds for quantiles or means, enabling adaptive anomaly scoring robust to drift, heavy tails, or outliers (Wang et al., 2023). Applying CS-based threshold selection to streaming anomaly detection is an active area (Howard et al., 2021; Sun et al., 2024).

2.5 Segmentation-Based Local Thresholding

Segmentation—by APCA or clustering—brings statistical homogeneity to threshold estimation, allowing each regime to have locally fitted adaptive rules (Keogh et al., 2001; Aghabozorgi et al., 2015). Clustering on summary features captures regime changes, but decision boundaries within segments remain underexplored.

3 Methods

We focus on two novel, unsupervised adaptive thresholding strategies for streaming time series: Segmented Confidence Sequences (SCS) and Multi-Scale Adaptive Confidence Segments (MACS).

3.1 Segmented Confidence Sequences (SCS)

SCS first performs time series segmentation using either Adaptive Piecewise Constant Approximation (APCA), which iteratively splits at points that minimize reconstruction error, or feature-based k-means clustering. Each segment is assumed locally stationary. Within each segment, an independent confidence sequence is maintained for anomaly score thresholds using Hoeffding's inequality. Segment-specific anomaly flags are triggered if new scores violate bounds.

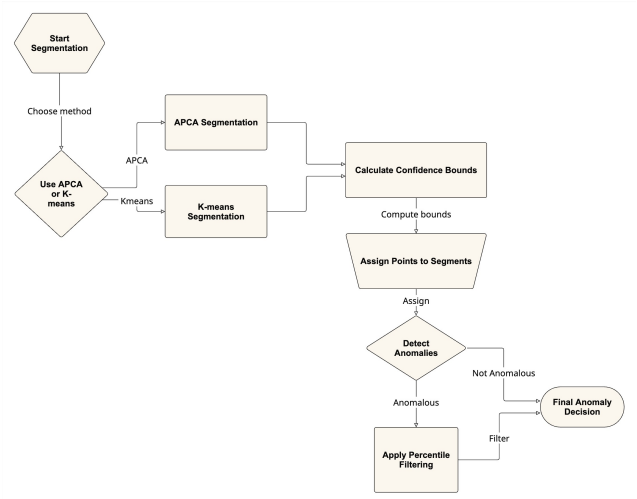

Figure 1: Illustration of the SCS flow.

APCA searches for split points minimizing total reconstruction error:

$$\texttt{total\_error} = \texttt{left\_error} + \texttt{right\_error} \tag{1}$$

$$\texttt{left\_error} = \sum (x_i - \bar{x}_{\text{left}})^2 \tag{2}$$

$$\texttt{right\_error} = \sum (x_j - \bar{x}_{\text{right}})^2 \tag{3}$$

Splitting continues until segment length is minimal or no improvement remains. For flat regions, APCA defaults to fixed segmentation:

$$\max(200, \left\lfloor \frac{n}{15} \right\rfloor) \tag{4}$$

Candidate splits are accepted if reconstruction error decreases by preset ratios.

Alternatively, SCS uses k-means segmentation on sliding window features (mean, std, median, skewness), normalized by standard scaler. Multidimensional input is averaged into one dimension.

Within each resulting segment, regardless of the segmentation method, SCS maintains an independent confidence sequence for thresholding anomaly scores. These bounds are derived using Hoeffding-style inequalities Howard et al. (2021) and are parameterized by the local standard deviation of the segment's scores. The width of the confidence bound is initially set as:

$$\texttt{bound\_width} = 1.5 \times \texttt{std\_score} \tag{5}$$

Bounds adjust with confidence level. The final interval is:

$$\texttt{lower\_bound} = \bar{x} - \texttt{bound\_width} \tag{6}$$

$$\texttt{upper\_bound} = \bar{x} + \texttt{bound\_width} \tag{7}$$

Anomalies must violate local bounds and a global percentile filter.

To summarize, the algorithm flow is outlined below:

- **Segmentation Phase**: Apply APCA or K-means to identify regime boundaries
- **Bound Calculation**: Compute confidence bounds for each segment independently
- **Point Assignment**: Dynamically assign incoming points to their corresponding segment
- **Anomaly Detection**: Compare each point to segment-specific thresholds
- **Filtering**: Apply percentile-based filtering for conservativeness

3.2 Multi-Scale Adaptive Confidence Segments (MACS)

MACS is designed to capture anomalies occurring at different temporal resolutions by maintaining multiple rolling windows of varying lengths. MACS maintains multiple rolling windows of varying

lengths in parallel (short, medium, long). Each scale independently maintains confidence bounds:

$$\texttt{bound\_width} = 1.5 \times \texttt{std\_score} \tag{8}$$

It is then scaled according to the desired confidence level. Specifically, the bound width is increased by 20% for high-confidence settings ($> 95\%$) and decreased by 20% for low-confidence settings ($< 90\%$).

MACS uses an attention mechanism based on local variance, with weights:

- High variance (> 0.7): [0.6, 0.3, 0.1]
- Medium variance (> 0.3): [0.2, 0.6, 0.2]
- Low variance (≤ 0.3): [0.1, 0.3, 0.6]

These weights are used to compute a combined confidence bound as a weighted sum across scales:

$$\texttt{combined\_bound} = \sum_{i=1}^{3} \texttt{weight}_i \cdot \texttt{bound}_i \tag{9}$$

In addition to confidence sequences, MACS performs regime change detection using a CUSUM-like procedure based on rolling statistics. It tracks both the rolling mean and standard deviation over the long window. A regime change is flagged if the normalized change in mean exceeds 2.0, or if the change in standard deviation exceeds 1.5, defined respectively as:

$$\texttt{mean\_change} = \frac{\bar{x}_{\text{current}} - \bar{x}_{\text{historical}}}{\texttt{std}_{\text{historical}} + 10^{-8}} \tag{10}$$

$$\texttt{std\_change} = \frac{\texttt{std}_{\text{current}} - \texttt{std}_{\text{historical}}}{\texttt{std}_{\text{historical}} + 10^{-8}} \tag{11}$$

A threshold violation counting mechanism flags a point as anomalous if it exceeds at least two out of three individual scale-specific thresholds.

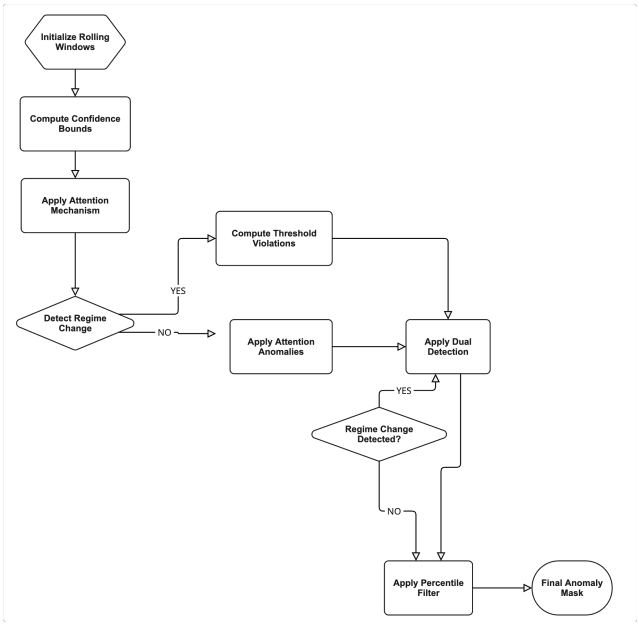

Figure 2: Illustration of the MACS flow.

To summarize, the algorithm flow is outlined below:

- **Multi-Scale Analysis**: Calculate confidence bounds at three temporal scales

- **Attention Calculation**: Compute local variance and determine attention weights
- **Bound Combination**: Apply attention mechanism to combine multi-scale bounds
- **Regime Detection**: Identify statistical regime changes using CUSUM-like logic
- **Dual Detection**: Apply both threshold violation counting and attention-weighted bounds
- **Regime-Aware Decision**: Combine detection methods based on regime state
- **Filtering**: Apply percentile-based filtering for conservativeness

4 Experimental Results

We evaluated SCS and MACS against traditional and adaptive methods on public datasets containing ground-truth anomaly labels. Metrics include confusion matrix, accuracy, precision, recall, and F1 relative to baseline. Evaluations span July 5th–31st, 2025.

4.1 Experiment and Dataset Description

4.1.1 Baseline: Traditional Percentile Thresholding

Our reference method follows the classic p-percentile rule.

1. **Reconstruction-error vector**
   Let x'_t be the output of the diffusion auto-encoder at time t and x_t the original series window. We compute the point-wise L2 residual:
   $$r_t = \|x_t - x'_t\|_2 \tag{12}$$
2. **Threshold selection**
   A global cut-off: the 99th percentile of the residual distribution on the training split
3. **Decision rule**
   A time stamp is labelled anomalous iff $r_t > \theta$.

Table 1: Overview of evaluated datasets

Name of Dataset	Source & Scope	Anomaly Labels
Wafer Manufacturing	151 traces from fabrication sensors	Pass/fail ground truth
CalIt2	People-count sensor, UC-Irvine CalIt2	High footfall event labels
Google Cloud Platform	30 KPI traces from NVIDIA DGX	Curated incident tickets
Mars Science Laboratory	Curiosity rover telemetry	Event labels
Server Machine Dataset	5-week production trace, 38 KPIs	Point-level anomaly labels
CPU-KPI	Seasonal CPU utilisation KPI	Partial point labels

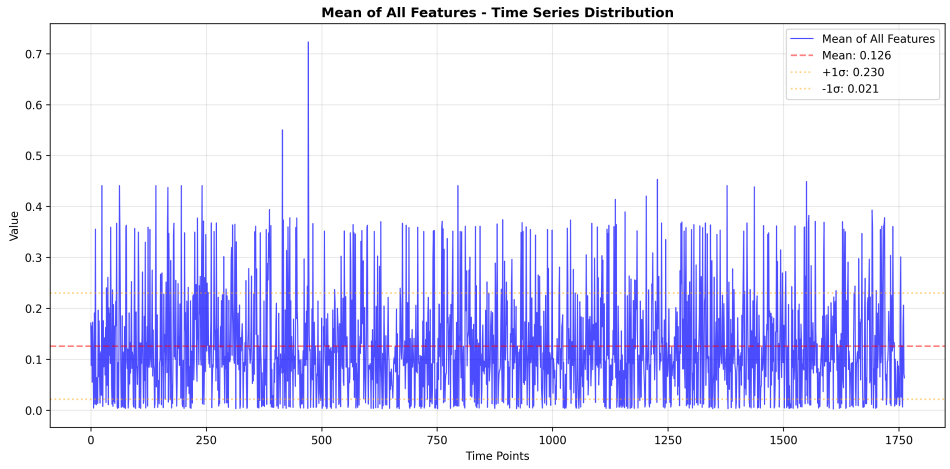

Figure 3: Wafer Manufacturing dataset distribution.

4.1.2 HYPER-PARAMETERS AND VARIANTS

- Confidence level $1 - \alpha$ for adaptive confidence sequences: $\{0.05, 0.01\}$.
- Segmentation for SCS: Adaptive Piecewise Constant Approximation (APCA) vs. k-means on residual variance.
- Baseline: fixed 99% percentile rule described above.

4.1.3 EVALUATION PROTOCOL

For every dataset we compute:

- Confusion-matrix counts (TP, FP, TN, FN)
- Change in Accuracy, Precision, Recall, F1 compared to baseline
- Proportional improvement over the baseline, calculated as:

$$\frac{\texttt{new\_method} - \texttt{traditional\_method}}{\texttt{traditional\_method}} \tag{13}$$

4.2 QUANTITATIVE COMPARISON

Key results on Wafer dataset:

Table 2: Performance delta on Wafer Manufacturing dataset

Method	Δ Acc.	Δ Prec.	Δ Recall	Δ F1
SCS APCA ($\alpha = 0.99$)	-0.0422	-0.3282	3.9952	1.9074
SCS KMEANS ($\alpha = 0.99$)	-0.0260	-0.3999	1.6643	0.9262
MACS Multi-Scale ($\alpha = 0.99$)	-0.0279	-0.1890	3.9952	2.1705
SCS APCA ($\alpha = 0.95$)	-0.0830	-0.4290	6.1595	2.1289
SCS KMEANS ($\alpha = 0.95$)	-0.0545	-0.4656	3.3286	1.4148
MACS Multi-Scale ($\alpha = 0.95$)	-0.0638	-0.3651	5.6595	2.2349

Figure 4: Results for Wafer Manufacturing dataset, $\alpha = 0.99$.

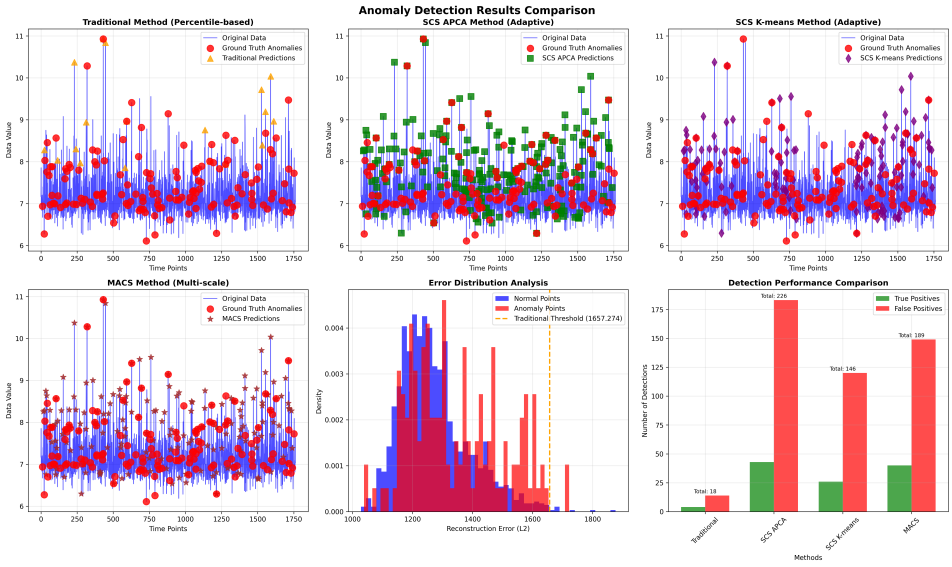

Figure 5: Results for Wafer Manufacturing dataset, $\alpha = 0.95$.

Table 3: Anomaly count comparison

Method	TP	TN	FP	FN
Baseline (99th percentile)	6	1608	12	137
SCS APCA ($\alpha = 0.99$)	30	1516	104	113
SCS KMEANS ($\alpha = 0.99$)	16	1556	64	127
MACS Multi-Scale ($\alpha = 0.99$)	30	1539	81	113
SCS APCA ($\alpha = 0.95$)	43	1437	183	100
SCS KMEANS ($\alpha = 0.95$)	26	1500	120	117
MACS Multi-Scale ($\alpha = 0.95$)	40	1471	149	103

4.3 Detailed Analysis

Both SCS and MACS show significant performance improvements over the traditional static percentile thresholding approach. The F1-score of both SCS and MACS with a confidence level of $\alpha = 0.99$ increases approximately twice compared to the baseline. When the confidence level is $\alpha = 0.95$, recall improves substantially, leading to an over two times increase in F1-score relative to the baseline, even at the cost of a moderate decline in precision.

The success of both approaches lies in their ability to localize statistical estimation. SCS adapts quickly to changes by segmenting the time series into regions with approximately stationary behavior, which allows for tight confidence bounds within each region. MACS, on the other hand, incorporates temporal diversity through rolling windows at multiple resolutions and adaptive attention weighting, enabling it to respond to anomalies that manifest at different time scales.

Finally, while removing the percentile filter maximizes recall and F1-score, this setting may not always be optimal. In noisy environments or when false positives carry significant cost, reintroducing percentile filtering may be desirable to balance interpretability with operational reliability.

5 Discussion

Our empirical findings reinforce the known limitations of static thresholding techniques such as global percentiles and rolling quantiles when applied to nonstationary time series data. These traditional approaches fail to account for dynamic distributional shifts, leading to poor recall and under-detection of relevant anomalies Howard et al. (2021). In contrast, the proposed SCS and MACS methods substantially improve performance by incorporating structural and temporal adaptivity. Specifically, they address evolving data behavior through segmentation (SCS) and multi-scale

temporal analysis (MACS), yielding significant F1-score gains with only modest reductions in precision.

SCS is particularly well-suited to settings characterized by abrupt regime shifts and piecewise stationarity, where local adaptation via segmentation captures the changing statistical properties of the signal. Its regime-specific confidence sequences offer interpretable bounds and fast detection of contextual outliers. MACS, on the other hand, is more flexible across a wider range of temporal patterns. By leveraging multiple rolling windows and variance-sensitive attention mechanisms, MACS generalizes across both fast transients and slow drifts. This makes it especially effective in environments with layered or multi-scale anomaly behavior, such as bursty network activity or gradual process degradation Wang et al. (2023).

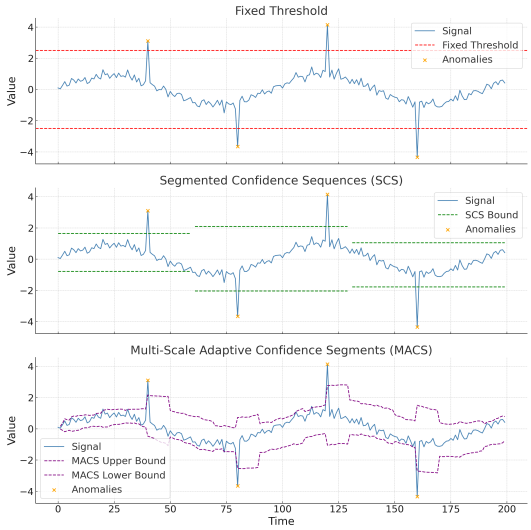

Figure 6: Illustration of different thresholding strategies.

A key advantage of both approaches lies in their model-free, unsupervised nature. Unlike many machine learning-based anomaly detectors, which often rely on labeled anomaly instances for training and hyperparameter tuning, SCS and MACS operate without supervision and retain explicit control over false alarm rates through statistically principled confidence sequences. This is crucial in high-stakes domains such as manufacturing, infrastructure monitoring, or cybersecurity, where excessive false positives can desensitize operators and degrade trust in automated systems Chandola et al. (2009); Ahmad et al. (2017).

Despite these advantages, our work also highlights some significant limitations and open challenges. The performance of SCS is sensitive to the structure of the time series. In datasets that are highly stationary or exhibit noisy, unstructured behavior, segmentation may fail to produce meaningful partitions. Similarly, while MACS benefits from its multi-scale architecture, its effectiveness hinges on the appropriate calibration of attention weights and confidence levels – parameters that may need tuning depending on the domain and noise profile.

An important direction for future work is the development of robust online segmentation algorithms capable of operating under adversarial conditions or extreme nonstationarity. This includes detecting latent regime transitions that are subtle, overlapping, or induced by external interventions. Additionally, while this study used fixed window sizes for MACS, there is potential in exploring adaptive window scaling or learned attention mechanisms that adjust over time based on predictive uncertainty or performance feedback.

6 Conclusion

Adaptive thresholding is vital for anomaly detection in nonstationary series. We introduce and evaluate SCS and MACS, integrating online confidence sequences and localized adaptation. Results show statistically principled, interpretable, high-performing anomaly detection in complex benchmarks.

Both frameworks enable flexible precision-recall trade-offs. Future work will extend to multivariate, correlated input streams, and inference-based scoring.

A Pseudocode for Segmented Confidence Sequences (SCS)

```
# Pseudocode for SCS adaptive thresholding

# Input: time_series, window_size, confidence_level, n_segments, segmentation_method
if segmentation_method == "APCA":
    segments = APCA_segment(time_series, n_segments)
elif segmentation_method == "k-means":
    segments = kmeans_segment(time_series, n_segments)

# Initialize confidence sequence per segment
for segment in segments:
    scores = compute_anomaly_scores(segment)
    conf_bounds = init_confidence_sequence(scores, confidence_level)

for new_point in stream:
    assigned_segment = assign_to_segment(new_point, segments)
    update(assigned_segment, new_point)
    if is_anomalous(new_point, assigned_segment.conf_bounds):
        flag_anomaly(new_point)
```

B Pseudocode for Multi-Scale Adaptive Confidence Segments (MACS)

```
# Pseudocode for MACS

# Input: time_series, short_window, medium_window, long_window, confidence_level
scales = [short_window, medium_window, long_window]
for scale in scales:
    window_scores[scale] = initialize_window(scale)
    conf_bounds[scale] = init_confidence_sequence(window_scores[scale], confidence_level)

for new_point in stream:
    for scale in scales:
        window_scores[scale].add(new_point)
        update_confidence_sequence(window_scores[scale], confidence_level)
    violation_count = sum(is_anomalous(new_point, conf_bounds[scale]) for scale in scales)
    if violation_count >= threshold:
        flag_anomaly(new_point)
```

C Pipeline Structure

1. Input: Time Series Data
2. Preprocessing: Remove seasonality/trend if needed
3. Segmentation: APCA/k-means (SCS), multi-scale windows (MACS)
4. Adaptive Thresholding: update segment/scale confidence sequences
5. Detection: composite anomaly filtering

References

Charu C. Aggarwal. *Outlier Analysis*. Springer, 2nd edition, 2015.

Seyed Amin Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16–38, 2015.

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

Konstantinos Benidis, Yoshua Bengio, Marc Blais, et al. Machine learning for time series forecasting: challenges and opportunities. *Proceedings of the IEEE*, 110(5):656–678, 2022.

A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3):1–33, 2021.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer, 2nd edition, 2016.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

Marc G. Genton, Yuguo Chen, and William Kleiber. Statistical methods for outlier detection. *Annual Review of Statistics and Its Application*, 8:297–321, 2021.

Steven R. Howard, Aaditya Ramdas, Jasjeet Sekhon, et al. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 18:1–45, 2021.

Eamonn Keogh, Kaushik Chakrabarti, Michael Pazzani, and Sharad Mehrotra. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 151–162, 2001.

Peter J. Rousseeuw, Mia Hubert, and Wesley Schmitt. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1380, 2020.

Sophia Sun, Aaditya Ramdas, and Jing Lei. Online adaptive anomaly thresholding with confidence sequences. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Jinlin Wang, Aaditya Ramdas, and Jing Lei. Robust and adaptive confidence sequences for heavy-tailed data. *Journal of the American Statistical Association*, 2023. To appear.

Yao Xue, Lingfei Wu, Pin-Yu Chen, and Bo Li. Adt: Agent-based dynamic thresholding for anomaly detection. In *Proceedings of the Adaptive and Learning Agents Workshop (ALA 2023)*, 2023.