Classifying Anomalous Members in a Collection of Multivariate Time Series Data Using Large Deviations Principle: An Application to COVID-19 Data

Sreelekha Guggilam[1,2], Varun Chandola[1,2], and Abani K. Patra[3]
[1] Computational Data Science & Eng., University at Buffalo (SUNY)
[2] Computer Science & Eng., University at Buffalo (SUNY)
[3] Data Intensive Studies Center, Tufts University
sreelekh@buffalo.edu,chandola@buffalo.edu,abani.patra@tufts.edu

Abstract. Anomaly detection for time series data is often aimed at identifying extreme behaviors within an individual time series. However, identifying extreme trends relative to a collection of other time series is of significant interest, like in the fields of public health policy, social justice and pandemic propagation. We propose an algorithm that can scale to large collections of time series data using the concepts from the theory of *large deviations*. Exploiting the ability of the algorithm to scale to high-dimensional data, we propose an online anomaly detection method to identify anomalies in a collection of multivariate time series. We demonstrate the applicability of the proposed *Large Deviations Anomaly Detection* (LAD) algorithm in identifying counties in the United States with anomalous trends in terms of COVID-19 related cases and deaths. Several of the identified anomalous counties correlate with counties with documented poor response to the COVID pandemic.

Keywords: Large deviations, Anomaly detection, High-dimensional data, Multivariate time series, Time series database

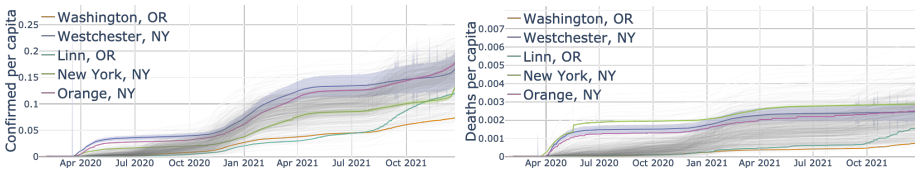
1 Introduction

Anomaly detection has been extensively studied over many decades across many domains [5] but remains difficult for comparisons across time series. This problem is critical to study policy responses in pandemic propagation, economics, social justice, climate change adaptation to name a few e.g. studying anomalous COVID-19 infection data trends across various countries, states or counties could identify successful public policies. Usual approaches to monitoring individual time series [16] and identifying sudden outbreaks or significant causal events cannot be used to detect gradual divergence or drift. In this paper, we propose a new anomaly detection algorithm *Large deviations Anomaly Detection* (LAD), for large/high-dimensional data and multivariate time series data. LAD uses the rate function from *large deviations theory* (LDT) [24] to deduce anomaly scores for identifying anomalies. Core ideas for the algorithm are inspired from an LDT projection theorem that allows better handling of high dimensional data. Unlike most high dimensional anomaly detection models, LAD does not use feature selection or dimensionality reduction, which makes it ideal to study multiple time series in an online mode. LAD model naturally segregates the anomalies at each time step while enabling comparison of multiple multivariate time series. Key advances of the novel LAD algorithm reported here are:

[3] An introductory pre-print version available as Guggilam et al. [11].

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

1. *Large deviations Anomaly Detection* (LAD) algorithm is a scalable LDP based method, for scoring based anomaly detection.
2. LAD model can analyze large and high dimensional datasets without additional dimensionality reduction increasing accuracy and reducing cost.
3. Online extension of LAD can detect anomalies across many multivariate time series using an evolving anomaly score for each tracking developing behavior.
4. An empirical study of publicly available anomaly detection benchmark datasets to analyze robustness and performance on high dimensional and large datasets.
5. A detailed analysis of COVID-19 trends for US counties where we identify counties with anomalous behavior (See Figure 1 for an illustration).



(a) Total Confirmed Cases **(b)** Total Deaths

Fig. 1: Top 5 anomalous counties in in USA identified by the LAD algorithm based on time-series data consisting of cumulative COVID-19 per-capita infections and deaths. The time-series for the non-anomalous counties are plotted (light-gray) in the background for reference. For the counties in New York, significant rise during early 2021 in confirmed cases (*left*) and high death rates, is detected. Washington and Linn County in Oregon are anomalous primarily due to steady low rates of infection.

2 Related Work

A large body of research exists on studying anomalies in high dimensional data [3]. Many anomaly detection algorithms use dimensionality reduction techniques as a pre-processing step to anomaly detection. However, many high dimensional anomalies can only be detected in high dimensional problem settings and dimensionality reduction in such settings can lead to false negatives. Many methods exist that identify anomalies on high-dimensional data without dimensional reduction or feature selection, e.g. by using distance metrics. *Elliptic Envelope* (EE) [21] fits an ellipse around data centers by fitting a robust covariance estimates. *Isolation Forest* (I-Forest) [15] uses recursive partitioning by random feature selection and isolating outlier observations. *k nearest neighbor outlier detection* (kNN) [18] uses distance from nearest neighbor to get anomaly scores. *local outlier factor* (LOF) [4] uses deviation in local densities with respect to its neighbors to detect anomalies. *k-means--* [7] method uses distance from nearest cluster centers to jointly perform clustering and anomaly detection. *Concentration Free Outlier Factor* (CFOF) [2] uses a "reverse nearest neighbor-based score" which measures the number of nearest neighbors required for a point to have a set proportion of data within its envelope. In particular, methods like I-Forest and CFOF are targeted towards anomaly detection in high dimensional datasets. However, they are not tailored for evolving data.

Many score based anomaly detection algorithms have been designed to classify anomalies within individual time series. For instance, Twitter Ad Vec [14]

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

are unsupervised study deviations from the data. Numenta[1] uses prediction errors to classify anomalies. Relative Entropy [25] compares entropy to identify anomalous observations. However, these algorithms are limited to studying only individual time series and not easily extended to an entire database of time series. Recently, large deviations theory has been widely applied in the fields of climate models [8], statistical mechanics [23], among others. Specially for analysis of time series, the theory of large deviations has proven to be of great interest over recent decades [17]. However, these methods are data specific, and often study individual time series. In most settings, real time detection of anomalies is needed to dispatch necessary preventive measures for damage control. Such problem formulation requires collectively monitoring a high dimensional time series database to identify anomalies in real time. While, the task of detecting anomalous time series in a collection of time series has been studied in the past [13], most of these works have focused on univariate time series and have not shown to scale to long time series data or provide limited explanation on why the identified trends are anomalous. Our proposed method addresses this.

3 Large Deviation Principle

Large deviations theory provides techniques to derive the probability of rare events[4] that have an asymptotically exact exponential approximation[9, 24]. The key concept of this theory is the Large Deviations Principle (LDP). The principle describes the exponential decay of the probabilities for the mean of random variables. To implement LDP on data with known distributions, it is important to decipher the rate function $\mathcal{I}$. Cramer's Theorem provides the relation between $\mathcal{I}$ and the logarithmic moment generating function Λ[5].

Theorem 1 (Cramer's Theorem). *Let $X_1, X_2, \dots X_n$ be a sequence of iid real random variables with finite logarithmic moment generating function, e.g. $\Lambda(t) < \infty$ for all $t \in \mathbb{R}$. Then the law for the empirical average satisfies the large deviations principle with rate $\epsilon = 1/n$ and rate function given by $\mathcal{I}(x) := \sup_{t\in\mathbb{R}} (tx - \Lambda(t)) \quad \forall t \in \mathbb{R}$.*

Thus, we get, $\lim_{n\to\infty} \frac{1}{n} \log\left(P\left(\sum_{i=1}^{n} X_i \geq nx\right)\right) = -\mathcal{I}(x), \quad \forall x > E[X_1]$. For more complex distributions, identifying the rate function using logarithmic moment generating function can be challenging. Many methods like contraction principle and exponential tilting exist that extend rate functions from one topological space that satisfies LDP to the topological spaces of interest[9]. For our work, we are interested in the Dawson-Gärtner Projective LDP, that generates the rate function using nested family of projections.

Theorem 2. *Dawson-Gärtner Projective LDP: Let $\{\pi^N\}_{N\in\mathbb{N}}$ be a nested family of projections acting on $\mathcal{X}$ s.t. $\cup_{N\in\mathbb{N}}\pi^N$ is the identity. Let $\mathcal{X}^N = \pi^N\mathcal{X}$ and $\mu_\epsilon^N = \mu_0 \circ (\pi^N)^{-1}, N \in \mathbb{N}$. If $\forall N \in \mathcal{N}$, the family $\{\mu_\epsilon^N\}_{\epsilon>0}$ satisfies the LDP on $\mathcal{X}^N$ with rate function $\mathcal{I}^N$, then $\{\mu_\epsilon\}_{\epsilon>0}$ satisfies the LDP with rate function I given by, $\mathcal{I}(x) = sup_{N\in\mathbb{N}}\mathcal{I}^N(\pi^N x) \quad x \in \mathcal{X}$.Since $\mathcal{I}^N(y) =$*

[4] In our context, these rare events include outlier/anomalous behaviors.

[5] The logarithmic moment generating function of a random variable X is defined as $\Lambda(t) = \log E[\exp(tX)]$.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

$inf_{\{x \in \mathcal{X} | \pi^N(x)=y\}} \mathcal{I}(x)$, $y \in \mathcal{Y}$, the supremum defining $\mathcal{I}$ is monotone in N because projections are nested.

The theorem allows extending the rate function from a lower to higher projection space. The implementation of this theorem in LAD model is seen in Section 4.

4 Methodology

Consider the case of multivariate time series data. Let $\{\mathbf{t_n}\}_{\mathbf{n=1}}^{\mathbf{N}}$ be a set of multivariate time series datasets where $\mathbf{t_n} = (\mathbf{t_{n,1}}, \ldots, \mathbf{t_{n,T}})$ is a time series of length T and each $\mathbf{t_{n,t}}$ has d attributes. The motivation is to identify anomalous $\mathbf{t_n}$ that diverge significantly from the non-anomalous counter parts at a given or multiple time steps. The main challenge is to design a score for individual time series that evolves in a temporal setting as well as enables tracking the initial time of deviation as well as the scale of deviation from the normal trend. As shown in following sections, our model addresses the problem through the use of rate functions derived from large deviations principle. We use the Dawson-Gärtner Projective LDP (See Section 4.2) for projecting the rate function function to a low dimensional setting while preserving anomalous instances. The extension to temporal data (See Section 4.3) is done by collectively studying each time series data as one observation.

4.1 Large Deviations for Anomaly Detection

Our approach uses a direct implementation of LDP to derive the rate function values for each observation. As the theory focuses on extremely rare events, the raw probabilities associated with them are usually very small [9, 24]. However, the LDP provides a rate function that is used as a scoring metric for LAD.

Consider a dataset X of size n. Let $\mathbf{a} = \{\mathbf{a_1}, \ldots, \mathbf{a_n}\}$ and $\mathbf{I} = \{\mathbf{I_1}, \ldots, \mathbf{I_n}\}$ be anomaly score and anomaly label vectors for the observations respectively such that $a_i \in [0, 1]$ and $I_i \in \{0, 1\}$ $\forall i \in \{1, 2, \ldots, n\}$. By large deviations principle, we know that for a given dataset X of size n, $P(\bar{X} = p) \approx e^{-n\mathcal{I}(p)}$. Assuming that the underlying data is standard Gaussian distribution with mean 0 and variance 1, we can use the rate function for Gaussian data where $\mathcal{I}(p) = \frac{p^2}{2}$. Then the resulting probability that the sample mean is p is given by $P(\bar{X} = p) \approx e^{-n\frac{p^2}{2}}$. Now, in presence of an anomalous observation x_a, the sample mean is shifted by approximately x_a/n for large n. Thus, the probability of the shifted mean being the true mean is given by $P(\bar{X} = x_a/n) \approx e^{-\frac{x_a^2}{2n}}$. However, for large n and $|x_a| << 1$, the above probabilities decay exponentially which significantly reduces their effectiveness for anomaly detection. Thus, we use $\frac{x_a^2}{2n}$ as anomaly score for our model. Thus generalizing this, the anomaly score for each individual observation is given by $a_i = n\mathcal{I}(x_i) \quad \forall i \in \{1, 2, \ldots, n\}$.

4.2 LDP for High Dimensional Data

High dimensional data pose significant challenges to anomaly detection. Presence of redundant or irrelevant features act as noise making anomaly detection difficult. However, dimensionality reduction can impact anomalies that arise from less significant features of the datasets. To address this, we use the Dawson-Gärtner Projective theorem in LAD model to compute the rate function for

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

high dimensional data. The theorem records the maximum value across all projections which preserves the anomaly score making it optimal to detect anomalies in high dimensional data. The model algorithm is presented in Algorithm 1.

Algorithm 1: Algorithm 1: LAD Model

Input: Dataset X of size (n, d), number of iterations N_{iter}, threshold th.
Output: Anomaly score $\mathbf{a}$
Initialization: Set initial anomaly score and labels $\mathbf{a}$ and $\mathbf{I}$ to zero vectors and, entropy matrix $E = 0_{(n,d)}$ where $0_{(n,d)}$ is a zero matrix of size (n, d).

for *each* $s \rightarrow 1$ **to** N_{iter} **do**
1. Subset $X_{sub} = X[I_i == 0]$
2. $X_{normalized}[:, d_i] = \frac{X[:,d_i] - X_{sub}[:,d_i]}{cov(X_{sub}[:,d_i])}, \forall d_i \in \{1, \ldots, d\}$
3. $E[i, :] = -X_{normalized}[i]^2/2n, \forall i$
4. $a_i = -max(E[i, :])$
5. $\mathbf{a} = \frac{\mathbf{a} - \mathbf{min(a)}}{\mathbf{max(a)} - \mathbf{min(a)}}$
6. $th = min(th, quantile(\mathbf{a}, \mathbf{0.95})$
7. $I_i = 1$ if $a_i > th, \forall i$

4.3 LAD for Time Series Data

Broadly, time series anomalies can be categorized to two groups [6]: (1) **Divergent trends/Process anomalies**: Time series with divergent trends that last for significant time periods fall into this group. Here, one can argue that generative process of such time series could be different from the rest of the non-anomalous counterparts, and (2) **Subsequence anomalies**: Such time series have temporally sudden fluctuations or deviations from expected behavior which can be deemed as anomalous. These anomalies occur as a subsequence of sudden spikes or fatigues in a time series of relatively non-anomalous trend. The online extension of the LAD model is designed to capture anomalous behavior at each time step. Based on the mode of analysis of the temporal anomaly scores, one can identify both divergent trends and subsequence anomalies. In this paper, we focus on the divergent trends (or process anomalies). In particular, we try to look at the anomalous trends in COVID-19 cases and deaths in US counties. Studies to collectively identify divergent trends and subsequence anomalies is being considered as a prospective future work.

In this section, we present an extension of the LAD model to multivariate time series data where we preserve the dependency temporal and across different features of the time series. Thus, as shown in Algorithm 2, a horizontal stacking of the data is performed. This allows collective study of temporal and non-temporal features. To preserve temporal dependency, the anomaly scores and labels are carried on to next time step where the labels are then re-evaluated.

As long term anomalies are of interest, time series with temporally longer anomalous behaviors are ranked more anomalous. The overall time series anomaly score A_n for each time series $\mathbf{t_n}$ can be computed as $A_n = \frac{\sum_{t=1}^{T} I[n,t]}{T} \quad \forall n$. For a database of time series with varying lengths, the time series anomaly score is computed by normalizing with respective lengths. Similarly, the method can be extended to studying anomalies within an individual time series by breaking the series into a database of sub-sequences of a time series extracted via a

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

Algorithm 2: Algorithm 2: LAD for Time series anomaly detection

Input: Time series dataset $\{\mathbf{t_n}\}_{\mathbf{n=1}}^{\mathbf{N}}$ of size (N, T, d), number of iterations N_{iter}, threshold th, window w.
Output: An array of temporal anomaly scores $\mathbf{a}$, an array of temporal anomaly labels I
Initialization: Set initial anomaly score and labels $\mathbf{a}$ and $\mathbf{I}$ to zero matrices of size (N, T) and, entropy matrix E to a zero matrix of size (N, T, d).

for *each* $t \rightarrow 1$ **to** T **do**
  $X = hstack(\bar{t_{n,t}})$ where $\bar{t_{n,t}} = \{t_{n,t-w}, \ldots t_{n,t}\}$
  $I[i,t] = I[i,t-1]$
  $\mathbf{a}[:,\mathbf{t}] = \mathbf{a}[:,\mathbf{t-1}]$
  for *each* $s \rightarrow 1$ **to** N_{iter} **do**
    1. Subset non-anomalous time series $X_{sub} = \{X[i,:] | I[i,t] == 0, \forall i\}$
    2. $X_{normalized}[:,d_i] = \frac{X[:,d_i] - \bar{X_{sub}[:,d_i]}}{cov(X_{sub}[:,d_i])}, \forall d_i \in \{1, 2, \ldots, d * w\}$
    3. $E[i,:] = -X_{normalized}[i]^2/2n, \forall i$
    4. $\mathbf{a}[\mathbf{i},\mathbf{t}] = -\mathbf{max}(\mathbf{E}[\mathbf{i},:])$
    5. $\mathbf{a}[:,\mathbf{t}] = \frac{\mathbf{a}[:,\mathbf{t}] - \mathbf{min}(\mathbf{a}[:,\mathbf{t}])}{\mathbf{max}(\mathbf{a}[:,\mathbf{t}]) - \mathbf{min}(\mathbf{a}[:,\mathbf{t}])}$
    6. $th = min(th, quantile(\mathbf{a}[:,\mathbf{t}], \mathbf{0.95})$
    7. $I[i,t] = 1$ if $\mathbf{a}[\mathbf{i},\mathbf{t}] > \mathbf{th}, \forall \mathbf{i}$

sliding window. It must be noted that this approach allows for a retrospective classification of anomalies.

5 Experiments

In this section, we evaluate the performance of the LAD algorithm on multi-aspect datasets. The following experiments have been conducted to study the model: 1) Anomaly Detection Performance: LAD's ability to detect real-world anomalies as compared to state-of-the-art anomaly detection models is evaluated using the ground truth labels. 2)Handling Large Data: Scalability of the LAD model on large datasets (high observation count or high dimensionality) are studied. 3) COVID-19 Time Series Data.

5.1 Datasets

We consider a variety of publicly available benchmark data sets from Outlier Detection DataSets ODDS [19] (See Tables 1) for the experimental evaluation. For anomaly detection within individual time series, we study univariate time series data from Numenta Benchmark Datasets [6]. Additionally, for the time series data, we use COVID-19 deaths and confirmed cases for US counties from John Hopkins COIVD-19 Data Repository [10]. The country level global data for COVID-19 trends was taken from the Our World in Data Repository [20].

5.2 Baseline Methods and Parameter Initialization

As described in Section 4, LAD falls under unsupervised learning regime targeted for high dimensional data, we do not compare with supervised algorithms. For this we consider *Elliptic Envelope* (EE) [21], *Isolation Forest* (I-Forest) [15][7],

[6] http://numenta.com/press/numenta-anomaly-benchmark-nab-evaluates-anomaly-detection-techniques.htm

[7] I-Forest model returns both anomaly scores and anomaly labels though we only present classification model since they outperforms score based schemes.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

Table 1: Description of the benchmark data used for evaluation of the anomaly detection for high dimensional/large sample datasets and time series. N - number of instances, d - number of attributes and a - fraction of known anomalies in the data set.

Name	N	d	a
HTTP	567498	3	0.39%
MNIST	7603	100	9.207%
Arrhythmia	452	274	14.602%
Shuttle	49097	9	7.151%
Letter	1600	32	6.25%
Musk	3062	166	3.168%
Optdigits	5216	64	2.876%
Satellite Img.	6435	36	31.639%
Speech	3686	400	1.655%
SMTP	95156	3	0.032%
Satellite Img.-2	5803	36	1.224%
Forest Cover	286048	10	0.96%
KDD99	620098	29	29 0.17%

(a) High Dimensional and Large Sample Datasets

Dataset	N	a
EC2 CPU UTILIZATION 825CC2	4032	0.09%
EC2 NETWORK IN 257A54	4032	0.1%
EC2 CPU UTILIZATION 5F5533	4032	0.1%
EC2 CPU UTILIZATION AC20CD	4032	0.1%
EC2 CPU UTILIZATION 24AE8D	4032	0.1%
SPEED 7578	1127	0.1%
SPEED 6005	2500	0.1%
OCCUPANCY 6005	2380	0.1%
SPEED T4013	2495	0.1%
ART LOAD BALANCER SPIKES	4032	0.1%
EXCHANGE-3 CPM RESULTS	1538	0.1%
EXCHANGE-4 CPM RESULTS	1643	0.1%
TWITTER VOLUME KO	15851	0.1%
TWITTER VOLUME CVS	15853	0.1%
TWITTER VOLUME CRM	15902	0.1%
MACHINE TEMP. SYS. FAILURE	22695	0.1%
EC2 REQ. LATENCY SYS. FAILURE	4032	0.09%
CPU UTIL. ASG MISCONFIG.	18050	0.08%

(b) Benchmark Time Series

local outlier factor (LOF) [4], and *Concentration Free Outlier Factor* CFOF [2]. The CFOF and LOF models assign an anomaly score for each observation, while remaining methods provide an anomaly label. As above mentioned methods are parametric, we investigated a range of values for each parameter, and report the best results. For Isolation Forest, Elliptic Envelope and CFOF, the contamination value is set to the true proportion of anomalies in the dataset. To study anomaly detection in time series, the LAD model is compared with other score based time series anomaly detection algorithms like Twitter AD Vec (TAV) [14], Skyline [22], Earthgecko Skyline (E.Skyline) [8], Numenta [1], Relative Entropy (RE) [25] , Random Cut Forest (RCF) [12], Windowed Gaussian (WG). The LAD model relies on a threshold value to classify observations with scores the value as strictly anomalous. Though this value is iteratively updated, an initial value is required by the algorithm. In this paper, the initial threshold value for the experiment is set to 0.95 for all datasets. All the methods for anomaly detection benchmark datasets are implemented in Python and all experiments were conducted on a 2.7 GHz Quad-Core Intel Core i7 processor with a 16 GB.

5.3 Evaluation Metrics

As LAD is an score based algorithm, we study the ROC curves by comparing the True Positive Rate (TPR) and False Positive Rate (FPR), across various thresholds. The final ROC-AUC (Area under the ROC curve) is reported for evaluation. For anomaly detection within individual time series, we use the F-measure as the evaluation metric to study the overall performance of the model. Since all the models return anomaly scores, thresholds were used to classify observations as anomalous vs non-anomalous. Threshold was set to be the maximum score in the truly non-anomalous data for each model and the observations with scores higher than set threshold were labelled anomalous. This is to ensure that the model is able to distinguish anomalies from the rest of the data. For time series database anomaly detection, we present the final outliers and study ~~their deviation~~s from normal baselines under different model settings.

[8] https://github.com/earthgecko/skylin

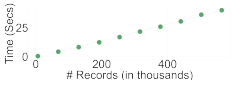
ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

5.4 Anomaly Detection Performance

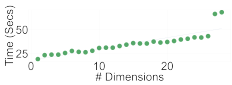
Table 2 shows the performance of LOF, I-Forest, EE, CFOF and LAD on anomaly detection benchmark datasets. Due to relatively large run-time[9], CFOF results are shown for datasets with samples less than 10k. For all the listed algorithms, results for best parameter settings are reported. The proposed LAD model outperforms other methods on most data sets. For larger and high dimensional datasets, it can be seen from Table 2 that the LAD model outperforms all the models in most settings.[10] It was interesting to note that the LAD model, despite being non-parametric (for a non-temporal setting), had a comparable if not better performance as compared to the LOF, EE, I-Forest and CFOF where multiple parameter setting were tested to derive the best fitting model. To study the LAD model's computational effectiveness, we study the computation time and scaling of LAD model on large and high dimensional datasets. Figure 2a shows the scalability of LAD with respect to the number of records against the time needed to run on the first k records of the KDD-99 dataset. Each record has 29 dimensions. Figure 2b shows the scalability of LAD with respect to the number of dimensions (linear-scale). We plot the time needed to run on the first 1, 2, ..., 29 dimensions of the KDD-99 dataset. The results confirm the linear scalability of LAD with number of records as well as number of dimensions.

Table 2: Comparing LAD with existing anomaly detection algorithms for large/ high dimensional datasets using ROC-AUC as the evaluation metric.

Data	LOF	I-Forest	EE	CFOF	LAD
SHUTTLE	0.52	0.98	0.96	-	**0.99**
SATIMAGE-2	0.57	0.95	0.96	0.70	**0.99**
SATIMAGE	0.51	0.64	**0.65**	0.55	0.6
KDD99	0.51	0.85	0.54	-	**1.0**
ARRHYTHMIA	0.61	0.67	0.7	0.56	**0.71**
OPTDIGITS	0.51	**0.52**	-	0.49	0.48
LETTER	0.54	0.54	0.6	**0.90**	0.6
MUSK	0.5	**0.96**	**0.96**	0.49	**0.96**
HTTP	0.47	0.95	0.95	-	**1.0**
MNIST	0.5	0.61	0.65	0.75	**0.87**
COVER	0.51	0.63	0.52	-	**0.96**
SMTP	**0.84**	0.83	0.83	-	0.82
SPEECH	0.5	**0.53**	0.51	0.47	0.47



(a) LAD scales linearly with the number of records for KDD-99 data



(b) LAD scales linearly with the number of dimensions in KDD-99 data.

Fig. 2: LAD Model Scaling on Large and High Dimensional Data

5.5 Anomaly Detection in Individual Time Series

In Table 3, we compare the performance of the LAD model as compared to other score-based algorithms. In particular, it can be seen that LAD model with window length of 100 has the best anomaly detection performance as compared other methods in most datasets.

5.6 Anomaly Detection in Time Series Data

This section presents the results of LAD model on COVID-19 time series data at the US county level. Multiple settings were used to understand the data: 1.

[9] The CFOF model is computationally expensive and its use is primarily for high-dimensional data. We restrict results to datasets with <10K observations.

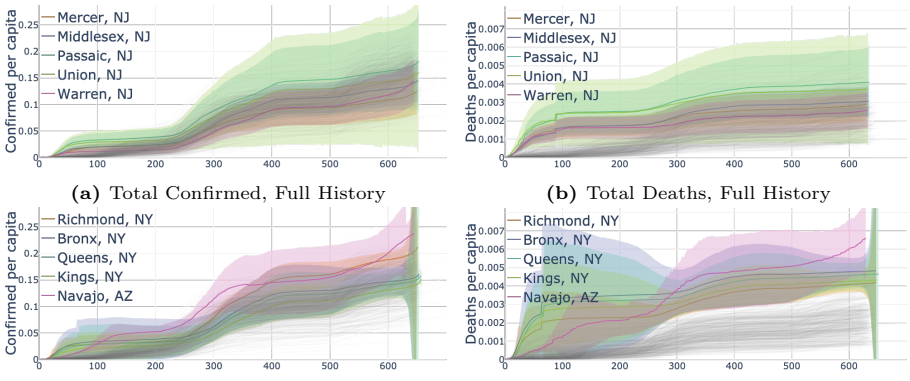
[10] Lowest AUC values for the LAD model are observed for Speech and Optdigits data where multiple true clusters are noted.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

Table 3: Comparing LAD with existing anomaly detection algorithms for time series datasets using F-measure as the evaluation metric.

Data	WL=10	WL=50	WL=100	TAV	Skyline	E.Skyline	Numenta	RE	RCF	WG
EC2 CPU UTIL. 825CC2	0.0	0.1	0.37	0.16	**0.45**	0.16	0.03	0.05	0.13	0.19
EC2 NETWORK IN 257A54	0.14	0.25	**0.33**	0.03	0.04	0.18	0.02	0.01	0.03	0.02
EC2 CPU UTIL. 5F5533	0.14	0.36	**0.57**	0.18	0.03	0.18	0.01	0.03	0.04	0.0
EC2 CPU UTIL. AC20CD	0.0	0.31	**0.33**	0.03	0.02	0.01	0.01	0.03	0.0	0.11
EC2 CPU UTIL. 24AE8D	0.09	0.12	**0.59**	0.01	0.01	0.0	0.0	0.0	0.0	0.01
SPEED 7578	0.26	0.29	**0.54**	0.19	0.08	0.05	0.05	0.08	0.02	0.17
SPEED 6005	0.15	**0.59**	**0.59**	0.04	0.11	0.11	0.03	0.04	0.04	0.01
OCCUPANCY 6005	0.08	0.29	**0.5**	0.01	0.01	0.01	0.01	0.01	0.01	0.0
SPEED T4013	0.27	**0.88**	0.45	0.15	0.16	0.02	0.04	0.03	0.13	0.14
ART LOAD BALANCER SPIKES	0.08	**0.16**	0.15	0.02	0.01	0.0	0.0	0.01	0.0	0.08
EXCHANGE-3 CPM RESULTS	0.0	0.4	**0.77**	0.01	0.01	0.01	0.01	0.03	0.01	0.01
EXCHANGE-4 CPM RESULTS	**0.21**	0.21	0.17	0.02	0.04	0.04	0.05	0.19	0.05	0.05
TWITTER VOL. KO	0.01	0.06	**0.11**	0.01	0.01	0.0	0.01	0.0	0.0	0.03
TWITTER VOL. CVS	0.04	0.06	**0.12**	0.03	0.01	0.01	0.01	0.01	0.01	0.03
TWITTER VOL. CRM	0.01	0.06	**0.11**	0.03	0.01	0.0	0.0	0.01	0.01	0.01
MACHINE TEMP SYS. FAIL.	0.02	0.04	0.08	**0.18**	0.03	0.01	0.0	0.02	0.03	0.0
EC2 REQ LATENCY SYS. FAIL.	0.2	**0.62**	0.35	0.15	0.04	0.15	0.02	0.15	0.03	0.02
CPU UTIL ASG MISCONFIG.	0.03	0.24	**0.83**	0.04	0.0	0.0	0.0	0.02	0.0	0.0

Deaths and confirmed case trends were considered for analysis. 2. Daily New vs Total Counts: Both total cases as well daily new cases were analyzed. 3. Complete history vs One Time Step: Two versions of the model were studied where data from previous time steps were and were not considered. By this, we tried to distinguish the impact of the history of the time series on identifying anomalous trends. 4. Univariate vs Multivariate Time Series data: To further understand the LAD model, the deaths and case trends were studied individually as a univariate time series as well as collectively in a multivariate time series data setting. 5. Time Series of Uniform vs Varying Lengths: Finally, all the above analyses were conducted on time series data with varying lengths. Here, for each county level time series, the time of first event was considered as initial time step to objectively study the relative temporal changes in trends. To bring all the counts to a baseline, the total counts in each time series were scaled to the respective county population. Missing information was replaced with zeros and counties with population less than 50k were eliminated from the study.



(a) Total Confirmed, Full History

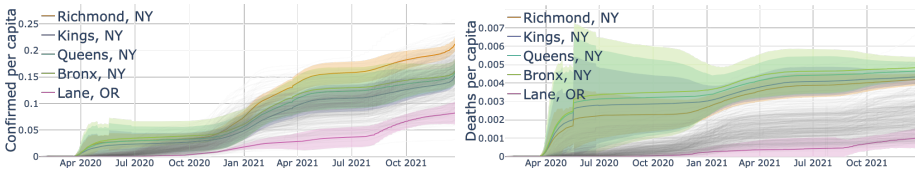
(b) Total Deaths, Full History

(c) Total Confirmed, One Time Step

(d) Total Deaths, One Time Step

Fig. 3: Top 5 Counties with Anomalous Trends : Varying lengths, Total Counts, Multivariate Time Series

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10



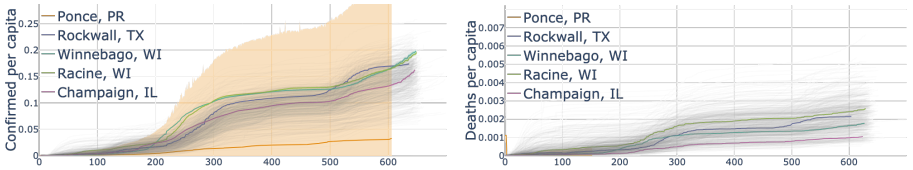
(a) Total Confirmed, One Time Step **(b)** Total Deaths, One Time Step

Fig. 4: Top 5 Counties with Anomalous Trends : Uniform lengths, Total Counts, Multivariate Time Series

5.7 Discoveries: US COVID-19 Trends

In this section, we look at the the daily new case and deaths in US counties trends in start of 2021. To rank the counties, anomaly scores between Jan 1 - Mar 1 2021 were considered.

Complete history vs One Time Step The full history setting considers the complete history of the time series and is aimed to capture most deviant trends over time. The one time step (or any smaller window) setting is more suitable to study deviations within the specific window. As we target long term deviating trends, the one time step setting returns trends that have stayed most deviant throughout the entire time range. This can be seen in Figures 3 and 4 where the one time step setting returns trends that have stayed deviant almost throughout the duration while the full history setting is able to capture significantly higher overall deviations from normal trends and therefore higher anomaly score. For instance, counties like Mercer(NJ), Union (NJ), that had extensive testing conducted[11] were captured in the one time step model as seen in Figure 3c and 3d. Similarly, counties in NY observed a peak in early 2021 [12], which was not captured as anomalous in the one time step model as seen in Figures 1a and 1b.



(a) New Confirmed, Full History **(b)** New Deaths, Full History

Fig. 5: Top 5 Counties with Anomalous Trends : Varying lengths, Daily New Counts, Multivariate Time Series

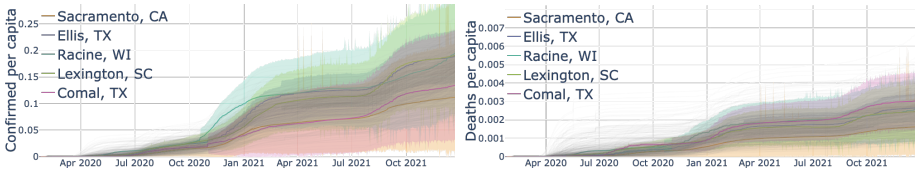
Univariate vs Multivariate Time series In Figures 3, 4, 5 and 6 we see the anomalous trends in multivariate time series, where total confirmed cases and deaths were collectively evaluated for anomaly detection. For instance, despite the near-normal trends in confirmed cases, Kings, Queens and Bronx (NY)[13]

[11] https://www.nj.com/coronavirus/2021/12/more-covid-testing-sites-opening-as-cases-climb-here-are-9-places-to-go.html

[12] https://www.newsday.com/news/health/coronavirus/coronavirus-long-island-deaths-vaccinations-1.50200404

[13] https://www.nbcnewyork.com/news/coronavirus/nyc-mask-mandate-indoors-an-option-if-needed-mayor-says-as-23-nations-report-omicron/3428102/

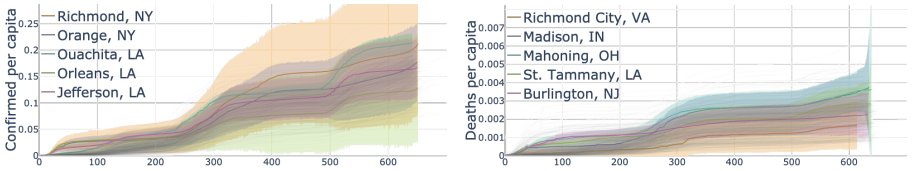
ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10



(a) New Confirmed, Full History **(b)** New Deaths, Full History

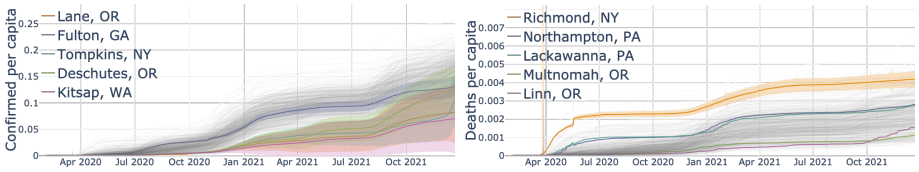
Fig. 6: Top 5 Counties with Anomalous Trends : Uniform lengths, Daily New Counts, Multivariate Time Series

in Figures 3c- 3d, were identified anomalous due to their the deviant death trends which significantly contributed to the anomaly scores. This setting enables identification of time-series with at least one deviating feature.



(a) Total Confirmed, Full History **(b)** Total Deaths, Full History

Fig. 7: Top 5 Counties with Anomalous Trends: Varying lengths, Total counts



(a) Total Confirmed, Full History **(b)** Total Deaths, Full History

Fig. 8: Top 5 Counties with Anomalous Trends: Uniform lengths, Total counts

Daily New vs Total Counts Figures 4 and 6, show anomalous trends in multivariate time series for total and daily new counts respectively. It can be seen that the anomaly score is relatively more erratic for new case trends as the data for new case and death counts is more erratic leading to fluctuating normal average and non-smooth anomaly scores. Similar behavior can be seen across Figures 3 and 5. The LAD model on the daily new counts data was able to capture the escalation in Racine, Wisconsin in Figure 6a and 6b during late 2020 when multiple meatpacking were tied to COVID-19 cases[14].

Uniform Length vs Varying Length Time Series The US county cases and deaths data consists of time series of uniform lengths. However, not all counties have events recorded in the early stages. Thus, studying the non-synchronized database creates a bias against counties with early reported cases. Also, counties with longer reporting on trends or earlier outbreaks tend to be associated with higher

[14] https://www.jsonline.com/story/news/2020/11/25/meatpacking-plants-tied-more-covid-19-cases-than-known-new-bussiness-outbreak-data-shows/6376197002/

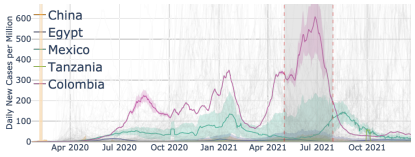
ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

anomaly scores towards the most recent data due to lack of equally long time series. This can be seen in Figures 4 where counties like Lane, Oregon that was flagged anomalous due to distinctively low cases due to later outbreak of the pandemic much after many counties in NY, unlike in Figures 3 which reports counties in NY with an early start as highly anomalous in the later stages[15].

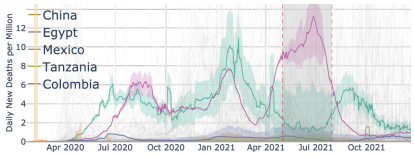
5.8 Global Trends and Emergence of Other COVID-19 Variants

Coronavirus Pandemic (COVID-19) Data from Our World in Data [20] for countries with population more than 5 million was used for the analysis. Trends in the daily new deaths per million and confirmed cases per million (7 day rolling average, right-aligned), biweekly growth rates in deaths and confirmed cases and case fatality rates were considered collectively as multivariate time series. Two end dates were studied to analyze the onset of the Delta and Omicron variants.

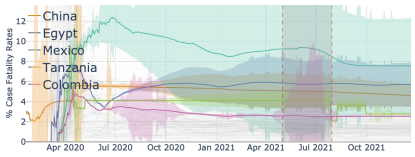
Delta Variant To rank the trends post the incidence of the Delta variant (See Figures 9a-9c), we considered behaviors during the 90 day period May 1 2021 - July 29 2021. China, Egypt, Mexico, Tanzania and Columbia were found most anomalous. In particular, China and Mexico had low per capita weekly average deaths and confirmed cases. However, the case fatality rates were consistently high [16] indicating need for additional investigation to understand the root cause which may be under-reporting or reporting issues or presence of a new variant.



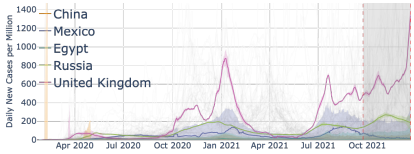
(a) Delta Variant: Daily new cases per million people (rolling 7-day average)



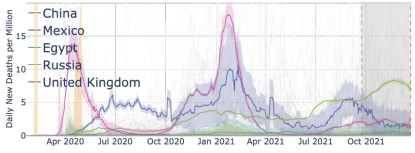
(b) Delta Variant: Daily new deaths per million people (rolling 7-day average)



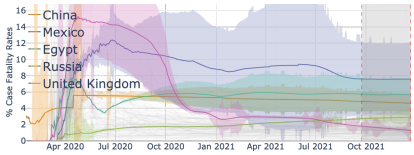
(c) Delta Variant: Case Fatality Rates



(d) Omicron Variant: Daily new cases per million people (rolling 7-day average)



(e) Omicron Variant: Daily new deaths per million people (rolling 7-day average)



(f) Omicron Variant: Case Fatality Rates

Fig. 9: Top 5 Anomalous Country-level Trends: Delta and Omicron Variants

[15] https://time.com/5812569/covid-19-new-york-morgues/

[16] https://www.marketwatch.com/story/new-daily-covid-19-cases-and-deaths-spike-to-6-week-highs-as-delta-variant-spreads-rapidly-11625673956

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

Omicron Variant To study the Omicron variant, we looked at the 90 day period data September 23 2021 - December 21 2021 (See Figures 9d-9f). UK, China have the most anomalous trends. Egypt, UK and Russia also have high anomaly scores [17]. However, in Egypt and Russia, the surge in cases was not due to the Omicron variant but due to earlier COVID wave that coincides with the it[18].

6 Conclusion

In this paper, we propose LAD, a novel scoring algorithm for anomaly detection in large/high-dimensional data. The algorithm successfully handles high dimensions by implementing large deviation theory. Our contributions include reestablishing the advantages of large deviations theory to large and high dimensional datasets. We present an online extension of the model aimed to identify anomalous time series in a multivariate time series data. The model shows vast potential in scalability and performance against baseline methods. The online LAD returns a temporally evolving score for each time series that allows us to study the deviations in trends relative to the complete time series database.

A potential extension to the model could include anomalous event detection for each individual time series. Another possible future work could be extending the model to enable anomaly detection in multi-modal datasets. Additionally, the online LAD model could be enhanced to use temporally weighted scores prioritizing recent events.

7 Acknowledgements

The authors would like to acknowledge University at Buffalo Center for Computational Research for computing resources and financial support of the National Science Foundation Grant numbers NSF/OAC 1339765 and NSF/DMS 1621853
.

References

[1] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing* (2017), 134–147.
[2] F. Angiulli. 2020. CFOF: a concentration free measure for anomaly detection. *ACM TKDD* 14, 1 (2020), 1–53.
[3] F. Angiulli and C. Pizzuti. 2002. Fast outlier detection in high dimensional spaces. In *Eur. Conf. on principles of data mining and knowledge discovery*. Springer, 15–27.
[4] M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of 2000 ACM SIGMOD International Conference on Management of Data*. 93–104.
[5] V. Chandola, A. Banerjee, and V. Kumar. 2009. Anomaly detection: A survey. *Comput. Surveys* 41, 3 (2009).

[17] https://www.cnn.com/2021/12/13/uk/uk-omicron-infections-tidal-wave-gbr-intl/index.html
[18] https://www.egyptindependent.com/egypt-has-not-passed-the-peak-of-the-covid-19-fourth-wave/, https://tass.com/society/1370957

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10

[6] V. Chandola, D. Cheboli, and V. Kumar. 2009. *Detecting Anomalies in a Timeseries Database.* Technical Report 09-004. University of Minnesota, Computer Science Dept.
[7] Sanjay Chawla and Aristides Gionis. 2013. k-means–: A unified approach to clustering and outlier detection. In *SDM*.
[8] G. Dematteis, T. Grafke, and E. Vanden-Eijnden. 2018. Rogue waves and large deviations in deep sea. *PNAS* 115, 5 (2018), 855–860.
[9] F. Den Hollander. 2008. *Large deviations.* Vol. 14. AMS.
[10] E. Dong, H. Du, and L. Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* 20, 5 (2020), 533–534.
[11] S. Guggilam, V. Chandola, and A. Patra. 2021. Anomaly Detection for High-Dimensional Data Using Large Deviations Principle. *arXiv preprint arXiv:2109.13698* (2021).
[12] S. Guha, N. Mishra, G. Roy, and O. Schrijvers. 2016. Robust random cut forest based anomaly detection on streams. In *ICML*. PMLR, 2712–2721.
[13] Hajar Homayouni, Indrakshi Ray, Sudipto Ghosh, Shlok Gondalia, and Michael G Kahn. 2021. Anomaly Detection in COVID-19 Time-Series Data. *SN Computer Science* 2, 4 (2021), 1–17.
[14] Arun Kejariwal. 2015. Introducing practical and robust anomaly detection in a time series. *Twitter Engineering Blog. Web* 15 (2015).
[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM TKDD* 6, 1 (2012), 1–39.
[16] M. Maleki, M. Mahmoudi, D. Wraith, and K. Pho. 2020. Time series modelling to forecast the confirmed and recovered cases of COVID-19. *Travel medicine and infectious disease* 37 (2020), 101742.
[17] T. Mikosch and O. Wintenberger. 2016. A large deviations approach to limit theory for heavy-tailed time series. *Prob. Theory and Related Fields* 166, 1 (2016), 233–269.
[18] S. Ramaswamy, R. Rastogi, and K. Shim. 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD Intl. Conf. on Management of data*. ACM Press, 427–438.
[19] Shebuti Rayana. 2016. ODDS Library. http://odds.cs.stonybrook.edu
[20] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser. 2020. Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
[21] P. Rousseeuw and K. Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 3 (1999), 212–223.
[22] A Stanway. 2013. Etsy skyline. https://github.com/etsy/skyline. *Online Code Repos* (2013).
[23] Hugo Touchette. 2009. The large deviation approach to statistical mechanics. *Physics Reports* 478, 1-3 (2009), 1–69.
[24] SR Srinivasa Varadhan. 1984. *Large deviations and applications.* SIAM.
[25] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, and K. Schwan. 2011. Statistical techniques for online anomaly detection in data centers. In *12th IFIP/IEEE Intl. Symposium on Integrated Network Management (IM 2011) and Workshops*. IEEE, 385–392.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08751-6_10