Sequential Model-Free Anomaly Detection for Big Data Streams

Mehmet Necip Kurt
Department of Electrical Engineering
Columbia University
New York, NY
m.n.kurt@columbia.edu

Yasin Yılmaz
Department of Electrical Engineering
University of South Florida
Tampa, FL
yasiny@usf.edu

Xiaodong Wang
Department of Electrical Engineering
Columbia University
New York, NY
wangx@ee.columbia.edu

***Abstract*—We study sequential anomaly detection for big data streams where the nominal and anomalous high-dimensional probabilistic data models are unknown. We propose a model-free solution approach in that we firstly compute a set of univariate summary statistics from a nominal dataset in an offline phase where the summary statistics are useful to distinguish anomalous data from nominal data. We then evaluate whether the online summary statistics deviate from the nominal case via a cumulative sum-like detector. Our experiments with real-world data illustrate the advantages of the proposed detector in early and reliable anomaly detection in big data settings compared to the existing alternatives.**

***Index Terms*—Big data, sequential anomaly detection, model-free, cumulative sum (CUSUM), summary statistic.**

I. Introduction

Early anomaly detection has a critical importance for safe and reliable operation of many modern large-scale systems such as the power networks and the Internet of Things (IoT) networks that produce big data streams. Anomalies often correspond to changes in the underlying statistical properties of the observed processes. To detect such changes, the framework of quickest detection [1], [2] is quite suitable, where the statistical inference about the monitored process is typically done through observations acquired sequentially over time and the goal is to detect the changes as soon as possible after they occur while limiting the risk of false alarm.

The well-known quickest detection algorithms are model-based: they require either the exact knowledge or estimates of the probability density functions (pdfs) of the observed data stream for both the pre- and post-change cases [1]–[3]. On the other hand, it is usually difficult to model or intractable to estimate the high-dimensional multivariate pdfs. Moreover, it is quite difficult to model all possible types of anomalies [4], [5]. To overcome such difficulties, we propose to extract useful univariate summary statistics from the observed data stream and perform the anomaly detection task based on the summary statistics, through which we also aim to make more efficient use of limited computational resources and to speed up the algorithms, that is especially required in time-sensitive online settings.

Summary statistics should be well informative to (statistically) distinguish anomalous data from nominal (non-anomalous) data, and its computation should be simple to allow for real-time processing. In this paper, we consider two alternative summary statistics: (i) if the observed nominal data has a low intrinsic dimensionality, firstly learning a representative low-dimensional submanifold for the nominal data and then computing a statistic that shows how much the incoming data stream deviates from the nominal submanifold; (ii) in the general case, learning an acceptance region for the nominal data via the Geometric Entropy Minimization (GEM) [6], [7] and then computing a nearest neighbor (NN) statistic that shows how much the incoming data stream is away from the acceptance region.

Anomaly detection schemes based on parametric models are vulnerable to model mismatch that limits their applicability. For instance, it is common to fit a Gaussian or Gaussian mixture model to the observed data or the data after dimensionality reduction [8], [9] and to assume Gaussian noise or residual terms, see e.g., [10]. Such parametric approaches are powerful only if the observed data perfectly matches with the presumed model. On the other hand, nonparametric (model-free) techniques are data-driven and hence robust to the data model mismatch. Moreover, in high-dimensional settings, the lack of parametric models is common and complicated parameter-laden algorithms generally result in low performance, overfitting, and bias towards particular anomaly types [11]. Hence, in this paper, we do not make parametric model assumptions for the observed big data stream nor for the summary statistics.

Conventional anomaly detection schemes ignore the temporal relation between anomalous data points and make sample-by-sample decisions [8], [9]. Such schemes are essentially outlier detectors that are vulnerable to false alarms since it is possible to observe non-persistent random outliers in a regular system operation due to e.g., heavy-tailed random noise processes. On the other hand, if a system produces persistent outliers, then this may indicate an actual anomaly. Hence, we define an anomaly as persistent outliers and from the observed data stream, we propose to accumulate statistical evidence for anomaly over time, similarly to the accumulation of log-likelihood ratios (LLRs) in the well-known cumulative sum (CUSUM) algorithm for change detection [12]. With the goal of making a reliable decision, we declare an anomaly only if we have a strong evidence for that. The sequential decision making based on the accumulated evidence also enables the detection of small but persistent changes, that would be missed by outlier detectors.

II. Problem Description

We observe a big data stream, particularly, at each time t we obtain $\mathbf{x}_t \in \mathbb{R}^p$ where $p \gg 1$ and the data points are independent and identically distributed (i.i.d.) over time. Suppose that an abrupt anomaly happens in the observed process at an unknown time τ, called the change-point, and continues thereafter. That is, the process is under regular operating conditions up to time τ and then its underlying statistical properties suddenly change at time τ due to an anomaly. Denoting the pdfs of $\mathbf{x}_t$ under regular (pre-change) and anomalous (post-change) conditions as $f_0^{\mathbf{x}}$ and $f_1^{\mathbf{x}} \neq f_0^{\mathbf{x}}$, respectively, we have

$$\mathbf{x}_t \sim \begin{cases} f_0^{\mathbf{x}}, & \text{if } t < \tau \\ f_1^{\mathbf{x}}, & \text{if } t \geq \tau. \end{cases}$$

We aim to detect the changes as quickly as possible after they occur. The framework of quickest detection well matches with this purpose. A well-known problem formulation in the quickest detection framework is the minimax problem where the goal is to minimize the worst-case detection delay subject to false alarm constraints [13]. If both $f_0^{\mathbf{x}}$ and $f_1^{\mathbf{x}}$ are known, then the CUSUM algorithm is the optimal solution to the minimax problem [14]. Let

$$\ell_t \triangleq \log\left(\frac{f_1^{\mathbf{x}}(\mathbf{x}_t)}{f_0^{\mathbf{x}}(\mathbf{x}_t)}\right)$$

denote the LLR at time t. In the CUSUM algorithm, the LLR is considered as the statistical evidence for change at a time and the LLRs are accumulated over time. If the accumulated evidence exceeds a predefined threshold, then a change is declared. Denoting the CUSUM decision statistic at time t by g_t and the decision threshold by h, the CUSUM algorithm is given by

$$\begin{aligned} \Gamma &= \inf\{t : g_t \geq h\}, \\ g_t &= \max\{0, g_{t-1} + \ell_t\}, \end{aligned} \tag{1}$$

where Γ denote the stopping time at which a change is declared and $g_0 = 0$.

Since it is practically difficult to model all types of anomalies, $f_1^{\mathbf{x}}$ needs to be assumed unknown for a general anomaly detection problem. In that case, if only $f_0^{\mathbf{x}}$ is known and also has a parametric form, slight deviations from the parameters of $f_0^{\mathbf{x}}$ can be detected using a generalized CUSUM algorithm [2], [15], [16]. However, in general, it might be difficult to model or estimate the high-dimensional multivariate nominal pdf $f_0^{\mathbf{x}}$. Hence, in this study, we assume that both $f_0^{\mathbf{x}}$ and $f_1^{\mathbf{x}}$ are unknown. We propose to use an alternative technique in that we extract useful univariate summary statistics from the observed high-dimensional data stream and perform the anomaly detection task in a single-dimensional space based on the extracted summary statistics, as detailed below.

III. Proposed Solution Approach

Firstly, we assume that there is an available set of nominal data points $\mathcal{X} \triangleq \{\mathbf{x}_i : i = 1, 2, \ldots, N\}$, that are free of anomaly. Using $\mathcal{X}$, we aim to extract univariate baseline statistics that summarize the regular system operation such that the summary statistics corresponding to anomalous data deviate from the baseline statistics. To this end, summary statistics should be well informative to distinguish anomalous conditions from the regular operating conditions.

Let the summary statistic corresponding to $\mathbf{x}_t$ be denoted by d_t. Since the statistical properties of $\mathbf{x}_t$ changes at time τ, we assume that the statistical properties of d_t also changes at τ. Denoting the nominal and anomalous pdfs of d_t as f_0^d and $f_1^d \neq f_0^d$, respectively, we then have

$$d_t \sim \begin{cases} f_0^d, & \text{if } t < \tau \\ f_1^d, & \text{if } t \geq \tau, \end{cases}$$

where we assume that f_0^d and f_1^d are both unknown. Nonetheless, extracting a set of nominal summary statistics from $\mathcal{X}$ and using this set as i.i.d. realizations of the nominal pdf f_0^d, we can form an empirical distribution function (edf) of the nominal summary statistics that estimates the nominal cumulative distribution function (cdf) F_0^d of d_t. Then, based on the nominal edf of the summary statistics, for an incoming data point $\mathbf{x}_t$ at time t and its corresponding summary statistic d_t, we can estimate the corresponding p-value, denoted with p_t. In statistical outlier detection, a data point $\mathbf{x}_t$ is considered as an outlier with respect to the level of α if its p-value is less than α, i.e., $p_t < \alpha$. Let

$$s_t \triangleq \log\left(\frac{\alpha}{p_t}\right). \tag{2}$$

Then, for an outlier $\mathbf{x}_t$, we have $s_t > 0$ and similarly, for a non-outlier $\mathbf{x}_t$, we have $s_t \leq 0$.

Under regular system operation, we may observe random non-persistent outliers due to e.g., high-level random system noise. However, if a system produces persistent outliers, then this may indicate an actual anomaly. Hence, we can model anomalies as persistent outliers. Considering s_t in (2) as a positive/negative statistical evidence for anomaly at time t, we can accumulate s_t's over time and obtain an accumulated evidence for anomaly. We can then declare an anomaly only if we have a strong (reliable) evidence supporting an anomaly. This gives rise to the following CUSUM-like anomaly detection algorithm where we replace the LLR ℓ_t in the CUSUM algorithm (see (1)) with s_t:

$$\begin{aligned} \Gamma &= \inf\{t : g_t \geq h\}, \\ g_t &= \max\{0, g_{t-1} + s_t\}, \end{aligned} \tag{3}$$

where $g_0 = 0$.

In the following section, we present derivations of the proposed summary statistics. Then, in Sec. V, we explain the estimation of the tail probability p_t (and hence s_t) based on the set of nominal summary statistics, that results in the final proposed detection algorithm.

IV. Summary Statistics

In this section, we firstly explain our methodology to derive summary statistics for a general high-dimensional data stream. We then explain the derivation of summary statistics in a

special case where the observed data exhibit a low intrinsic dimensionality.

A. GEM-based Summary Statistics

Given a nominal dataset $\mathcal{X}$ and a chosen significance level of α, the GEM method [6] determines an acceptance region $\mathcal{A}$ for the nominal data based on the asymptotic theory of random Euclidean graphs such that if a data point falls outside $\mathcal{A}$, it is considered as an outlier with respect to the level α, otherwise considered as a non-outlier. The GEM method is based on the NN statistics that capture the local interactions between data points governed by the underlying statistical properties of the observed data stream.

A computationally efficient GEM method presented in [7] is based on bipartite kNN graphs (BP-GEM). The BP-GEM method asymptotically achieves the minimum entropy set, i.e., the most compact acceptance region for the nominal data. In this method, firstly $\mathcal{X}$ is uniformly randomly partitioned into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ with sizes N_1 and $N_2 = N - N_1$, respectively. Then, for each data point $\mathbf{x}_j \in \mathcal{S}_2$, the kNNs of $\mathbf{x}_j$ among the set $\mathcal{S}_1$ are determined. Denoting the Euclidean distance of $\mathbf{x}_j$ to its ith NN in $\mathcal{S}_1$ by $e_j(i)$, the sum of distances of $\mathbf{x}_j$ to its kNNs can be written as follows:

$$d_j \triangleq \sum_{i=1}^{k} e_j(i).$$

After computing $\{d_j : \mathbf{x}_j \in \mathcal{S}_2\}$, d_j's are sorted in ascending order and the $(1-\alpha)$ fraction of $\mathbf{x}_j$'s in $\mathcal{S}_2$ corresponding to the smallest $(1-\alpha)$ fraction of d_j's form the acceptance region $\mathcal{A}$. Then, for a new data point $\mathbf{x}_t$, if its sum of distances to its kNNs among $\mathcal{S}_1$, denoted with d_t, is greater than the smallest $(1-\alpha)$ fraction of d_j's, i.e.,

$$\frac{\sum_{\mathbf{x}_j \in \mathcal{S}_2} \mathbb{1}\{d_t > d_j\}}{N_2} > 1-\alpha,$$

then $\mathbf{x}_t$ is considered as an outlier with respect to the level of α, where $\mathbb{1}\{\cdot\}$ denotes an indicator function.

If $\mathbf{x}_t$ is an outlier, then it falls outside the acceptance region $\mathcal{A}$, i.e., the corresponding NN statistic d_t takes a higher value compared to non-outliers. Moreover, if the observed data stream persistently fall outside the acceptance region, or equivalently if we persistently observe high NN statistics over time, then this may indicate an anomaly. Hence, we can use the GEM-based NN statistic as a summary statistic to distinguish anomalous data from nominal data. Then, we can use $\{d_j : \mathbf{x}_j \in \mathcal{S}_2\}$ as a set of GEM-based nominal summary statistics.

B. Summary Statistics for High-Dimensional Data Exhibiting Low Intrinsic Dimensionality

In many practical applications, observed big data exhibits a low intrinsic dimensionality and hence it can be well represented in a lower-dimensional subspace. In such cases, we can model the data as follows:

$$\mathbf{x}_t = \mathbf{y}_t + \mathbf{r}_t,$$

where $\mathbf{y}_t$ is the representation of $\mathbf{x}_t$ in a submanifold and $\mathbf{r}_t$ is the residual term, i.e., the departure of $\mathbf{x}_t$ from the submanifold, mostly consisting of noise.

Suppose that we learn a submanifold that the nominal data are embedded in. Since the learned manifold is mainly representative for the nominal data, anomalous data points are expected to deviate from the nominal submanifold and hence the magnitude of the residual term, i.e., $\|\mathbf{r}_t\|_2$, is expected to take higher values for anomalous data compared to nominal data. Hence, the magnitude of the residual term can be used as a summary statistic to distinguish anomalous data. Given a nominal dataset $\mathcal{X}$, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two subsets of $\mathcal{X}$, i.e., $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{X}$, with sizes N_1 and N_2, respectively, where $N_1, N_2 \leq N$. Firstly, using $\mathcal{S}_1$, we can determine a representative submanifold that the nominal data are embedded in. Then, using $\mathcal{S}_2$, we can compute the magnitude of the residual terms, i.e., $\{\|\mathbf{r}_j\|_2 : \mathbf{x}_j \in \mathcal{S}_2\}$, that can be used as a set of nominal summary statistics.

There are various methods to determine the underlying submanifold, among which the PCA is well known model-free method for learning a linear submanifold, called the principal subspace [17, Sec. 12.1]. In the PCA method, denoting $\bar{\mathbf{x}}$ as the sample mean, i.e.,

$$\bar{\mathbf{x}} \triangleq \frac{1}{N_1} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \mathbf{x}_i$$

and $\mathbf{Q}$ as the sample data covariance matrix, i.e.,

$$\mathbf{Q} \triangleq \frac{1}{N_1} \sum_{\mathbf{x}_i \in \mathcal{S}_1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}, \tag{4}$$

firstly, the eigenvalues $\{\lambda_j : j = 1, 2, \ldots, p\}$ and the eigenvectors $\{\mathbf{v}_j : j = 1, 2, \ldots, p\}$ of $\mathbf{Q}$ are computed. Then, the dimensionality of the submanifold, r, can be determined based on the desired fraction of data variance retained in the submanifold, given by

$$\gamma \triangleq \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \leq 1, \tag{5}$$

where the r-dimensional principal subspace is spanned by the orthonormal eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_r$ corresponding to the r largest eigenvalues λ_1, λ_2, ..., λ_r of $\mathbf{Q}$. Let $\mathbf{V} \triangleq [\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_r]$. The residual term for $\mathbf{x}_t$ can then be computed as follows:

$$\mathbf{r}_t = (\mathbf{I}_p - \mathbf{V}\mathbf{V}^{\mathrm{T}})(\mathbf{x}_t - \bar{\mathbf{x}}), \tag{6}$$

where $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ is an identity matrix.

To obtain the PCA-based nominal summary statistics, firstly, using $\mathcal{S}_1$, we can compute $\mathbf{Q}$ based on (4), and then its eigenvalues and eigenvectors. Then, for a chosen γ (see (5)), we can determine r and the corresponding $\mathbf{V}$. Finally, using $\mathcal{S}_2$ and (6), we can compute $\{\|\mathbf{r}_j\|_2 : \mathbf{x}_j \in \mathcal{S}_2\}$, that forms a set of nominal PCA-based summary statistics.

V. Sequential Model-Free Anomaly Detection

For outliers, both of the proposed summary statistics, d_t and $\|\mathbf{r}_t\|_2$, take higher values compared to non-outliers (see

Sec. IV). Hence, outliers in fact correspond to the right tail events based on the nominal pdf of the summary statistics. Let us specifically consider d_t. In case the knowledge of the nominal pdf of d_t, i.e., f_0^d, is available, we would compute the corresponding right tail probability as follows:

$$p_t = \int_{d_t}^{\infty} f_0^d(z)dz = 1 - F_0^d(d_t),$$

where F_0^d is the cdf of d_t. If $p_t < \alpha$, we can then consider d_t (correspondingly $\mathbf{x}_t$) as an outlier with respect to the significance level α.

In our problem, although we do not have the knowledge of f_0^d (and F_0^d), using a set of i.i.d. realizations of the nominal summary statistics, we can obtain an edf that estimates F_0^d. Let $\{d_j : \mathbf{x}_j \in \mathcal{S}_2\}$ be the set of nominal summary statistics. Then, the corresponding edf is given by

$$\hat{F}_{0,N_2}^d(z) \triangleq \frac{1}{N_2} \sum_{\mathbf{x}_j \in \mathcal{S}_2} \mathbb{1}\{d_j \leq z\}.$$

Moreover, by the Glivenko-Cantelli theorem, $\hat{F}_{0,N_2}^d$ pointwise almost surely converges to the actual cdf F_0^d as $N_2 \to \infty$ [18]. Then, we can estimate p_t based on $\hat{F}_{0,N_2}^d$ as follows:

$$\begin{aligned} \hat{p}_t &= 1 - \hat{F}_{0,N_2}^d(d_t) \\ &= \frac{1}{N_2} \sum_{\mathbf{x}_j \in \mathcal{S}_2} \mathbb{1}\{d_j > d_t\}. \end{aligned} \tag{7}$$

That is, $\hat{p}_t$ is simply the fraction of the nominal summary statistics $\{d_j : \mathbf{x}_j \in S_2\}$ greater than d_t. If $\hat{p}_t < \alpha$, then we can consider $\mathbf{x}_t$ as an outlier with respect to the level of α.

Let

$$\hat{s}_t \triangleq \log\left(\frac{\alpha}{\hat{p}_t}\right).$$

Notice that for an outlier $\mathbf{x}_t$ with respect to a level of α, we have $\hat{s}_t > 0$ and similarly, for a non-outlier $\mathbf{x}_t$, we have $\hat{s}_t \leq 0$. Then, by replacing $\hat{s}_t$ with s_t in (3), we propose the following model-free CUSUM-like anomaly detection algorithm:

$$\begin{aligned} \Gamma &= \inf\{t : g_t \geq h\}, \\ g_t &= \max\{0, g_{t-1} + \hat{s}_t\}, \end{aligned}$$

where $g_0 = 0$[1].

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed detection schemes using a human physical activity dataset. We choose $\alpha = 0.2$ and for all the proposed and benchmark tests, we obtain the tradeoff curves between the average detection delay, $\mathbb{E}_\tau\left[(\Gamma - \tau)^+\right]$, and the average false alarm period, $\mathbb{E}_\infty[\Gamma]$, by varying the test thresholds h. In computing the detection delays, we assume that anomalies happen at $\tau = 1$, that corresponds to the worst-case detection delay for the proposed algorithms since the decision statistic g_t is equal to zero just before the anomalies happen (recall that $g_0 = 0$). We use the following benchmark algorithms: Information Theoretic Multivariate Change Detection (ITMCD) algorithm presented in [19], NN-based online change detection algorithm presented in [20] and the QuantTree algorithm presented in [21].

The Human Activities and Postural Transitions (HAPT) dataset [22] obtained from the UCI Machine Learning Repository [23] contain data for six physical activities: sitting, standing, laying, walking, walking upstairs, and walking downstairs. The first three, i.e., sitting, standing, and laying, are static and the remaining three are dynamic activities. We divide the given dataset into two parts based on the given activity labels such that the first part of the dataset contains data for static activities and the second part contains data for dynamic activities. We detect changes from a static to a dynamic activity where each data point is 561-dimensional. We hence consider the static activities as the pre-change (nominal) state and the dynamic activities as the post-change (anomalous) state.

We firstly uniformly select 2500 data points from the set of data points corresponding to static activities and using the PCA method, we obtain the eigenvalues of the corresponding sample data covariance matrix, as shown in descending order in Fig. 1. We observe through Fig. 1 that the nominal data exhibit a low intrinsic dimensionality. We then choose the minimum desired γ as 0.99. Accordingly, we choose $r = 115$ and retain approximately $\gamma = 0.9903$ fraction of the data variance in the 115-dimensional principal subspace. Then, for the entire set of static activities ($\mathcal{S}_2 = \mathcal{X}$), we compute the PCA-based nominal summary statistics that form the histogram shown in Fig. 2.

In cases where the observed data stream exhibits a low intrinsic dimensionality, we can employ the proposed GEM-based detection scheme after dimensionality reduction for time efficiency. That is, after obtaining the matrix $\mathbf{V}$ via the PCA, each data point in the nominal training set, $\mathbf{x}_i \in \mathcal{X}$, and also each sequentially available data point, $\mathbf{x}_t$, can be projected onto a r-dimensional space as $\mathbf{V}^{\mathrm{T}}\mathbf{x}_i$ and $\mathbf{V}^{\mathrm{T}}\mathbf{x}_t$, respectively. We then employ the GEM-based detector over the projected data, where we uniformly choose $\mathcal{S}_1$ and $\mathcal{S}_2$ with sizes $N_1 = 1000$ and $N_2 = 4738$, respectively. Fig. 3 shows that the proposed algorithms perform superior than the benchmark algorithms. In the figure, we use an asterisk for the GEM-based detector to emphasize that it is employed based on the projected data.

VII. Conclusions

In this paper, we have proposed data-driven sequential anomaly detection schemes for big data streams. The proposed schemes are reliable, effective, and scalable. Moreover, they are widely applicable in a variety of applications as we do not make unrealistic data model assumptions. We have considered both the special case where the observed data stream has a low intrinsic dimensionality and the general case. In both cases, we have proposed to extract and monitor univariate summary statistics from the observed big data streams, where

[1]In case where $\sum_{\mathbf{x}_j \in \mathcal{S}_2} \mathbb{1}\{d_j > d_t\} = 0$, we have $\hat{p}_t = 0$ (see (7)), and hence $g_t = \infty$. In this case, a small nonzero value, e.g., $1/N_2$, can be assigned to $\hat{p}_t$ in order to prevent the decision statistic to raise to infinity due to a single outlier. This modification can be useful to reduce the false alarm rate especially in the small-sample settings (small N_2).

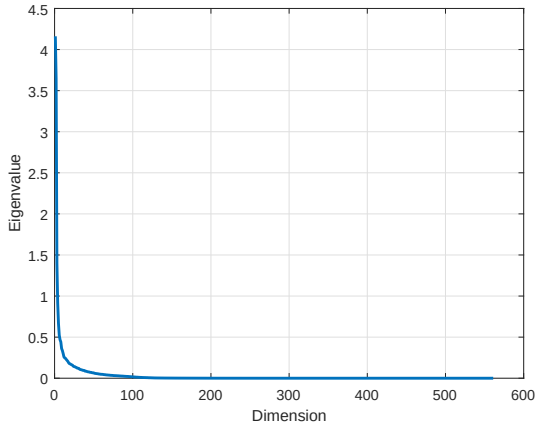

Fig. 1. Eigenvalues of the sample data covariance matrix for a representative set of static activities in the HAPT dataset.

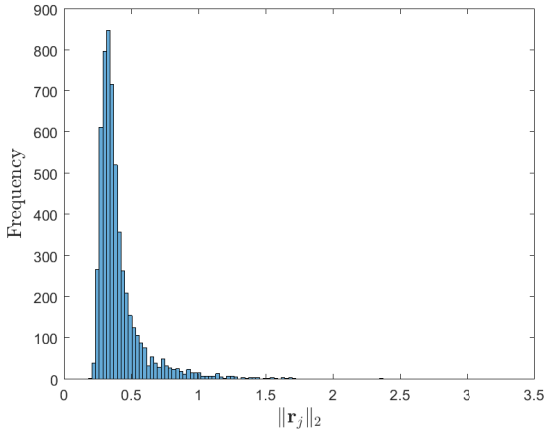

Fig. 2. PCA-based nominal summary statistics for static activities in the HAPT dataset.

the summary statistics are useful to distinguish anomalous data from nominal data. We have proposed a low-complexity CUSUM-like anomaly detection algorithm that makes use of the extracted summary statistics. Simulations with real-world data demonstrate the effectiveness of the proposed schemes in quick and accurate anomaly detection.

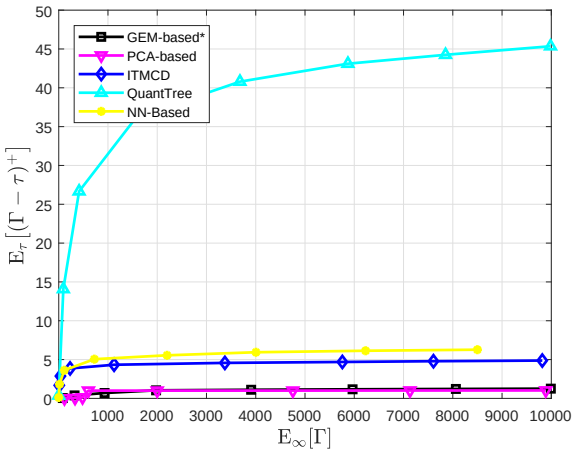

Fig. 3. Average detection delay vs. average false alarm period for detecting changes in human physical activities.

REFERENCES

[1] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2008.

[2] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[3] M. N. Kurt and X. Wang, "Multisensor sequential change detection with unknown change propagation pattern," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 3, pp. 1498–1518, June 2019.

[4] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: A reinforcement learning approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5174–5185, Sep. 2019.

[5] M. N. Kurt, Y. Yilmaz, and X. Wang, "Secure distributed dynamic state estimation in wide-area smart grids," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2019.

[6] A. O. Hero, III, "Geometric entropy minimization (gem) for anomaly detection and localization," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, pp. 585–592.

[7] K. Srichanran and A. O. Hero, "Efficient anomaly detection using bipartite k-nn graphs," in *Advances in Neural Information Processing Systems*, 2011, pp. 478–486.

[8] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[9] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[10] Y. Xie, J. Huang, and R. Willett, "Change-point detection for high-dimensional time series with missing data," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 12–27, 2013.

[11] R. Laxhammar and G. Falkman, "Online learning and sequential anomaly detection in trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1158–1173, June 2014.

[12] Y. Yilmaz, "Online nonparametric anomaly detection based on geometric entropy minimization," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 3010–3014.

[13] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Statist.*, vol. 42, no. 6, pp. 1897–1908, 1971.

[14] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, 1986.

[15] M. N. Kurt, Y. Yilmaz, and X. Wang, "Distributed quickest detection of cyber-attacks in smart grid," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2015–2030, Aug. 2018.

[16] ——, "Real-time detection of hybrid and stealthy cyber-attacks in smart grid," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 498–513, Feb 2019.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[18] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 1998, vol. 3.

[19] L. Faivishevsky, "Information theoretic multivariate change detection for multisensory information processing in internet of things," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6250–6254.

[20] H. Chen, "Sequential change-point detection based on nearest neighbors," *Ann. Statist.*, vol. 47, no. 3, pp. 1381–1407, 2019.

[21] G. Boracchi, D. Carrera, C. Cervellera, and D. Macciò, "Quanttree: Histograms for change detection in multivariate data streams," in *ICML*, 2018.

[22] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, 2016.

[23] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml