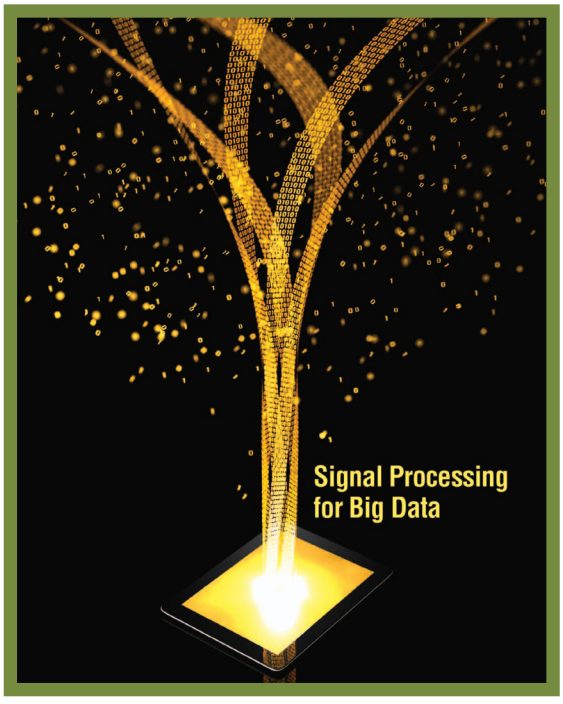
Ali Tajer, Venugopal V. Veeravalli, and H. Vincent Poor

Outlying Sequence Detection in Large Data Sets



A data-driven approach

Outliers refer to observations that do not conform to the expected patterns in high-dimensional data sets. When such outliers signify risks (e.g., in fraud detection) or opportunities (e.g., in spectrum sensing), harnessing the costs associated with the risks or missed opportunities necessitates mechanisms that can identify them effectively. Designing such mechanisms involves striking an appropriate balance between reliability and cost of sensing, as two opposing performance measures, where improving one tends to penalize the other. This article poses and analyzes outlying sequence detection in a hypothesis testing framework under different outlier recovery objectives and different degrees of knowledge about the underlying statistics of the outliers.

Digital Object Identifier 10.1109/MSP.2014.2329428
Date of publication: 19 August 2014

INTRODUCTION

MOTIVATION

Advances in data acquisition and high-dimensional information processing are rapidly transforming various technological, social, and economic domains, including the Internet, telecommunication, energy grids, social networks, and the health industries, to name a few. Empowered by these advances, such domains are evolving into

1053-5888/14©2014IEEE

complex networked platforms in which high-dimensional and complex data is routinely generated, communicated, stored, and processed for various monitoring, inference, and resource management purposes. Due to the inherent scale of data and complexity of the processes involved, the challenges associated with capturing, curating, searching, and sharing the information are also expected to grow well into the future. Hence, benefiting from the full extent of such enabling technologies is feasible only when appropriate measures are implemented that address these growing challenges while recognizing constraints pertinent to the physical limits of the application domains of interest.

Analyzing large-scale and complex data sets involves multifaceted phases, each of which introduces its own set of challenges. These phases include data acquisition and storage, information extraction, data aggregation, data modeling, and query processing, and the associated challenges include information heterogeneity, processing timeliness, data security and privacy, and human interactions. By capitalizing on the promises of data-driven information processing theories for understanding and addressing these challenges, this article focuses on a particular class of challenges related to information extraction and its associated timeliness requirements.

Extracting information and knowledge from data sets has been studied extensively over the past decade through developing powerful data mining and statistical learning methods. These methods are primarily focused on discovering (inferring) patterns in data sets and have widespread applications. In addition to the ongoing developments in discovering patterns in large data sets, there has also been a growing interest in uncovering outlying observations, which are observations that do not conform to expected patterns in large data sets. Such outlying observations generally refer to observations that are significantly different from the other data set constituents. While defining and identifying outliers are subjective exercises, outlier observations are often abstracted as deviations in the nature of a data set population and are considered to be caused by transient disruptions during data acquisition due to, for instance, a malfunctioning measurement apparatus, noisy data transmission media, or abrupt changes in the nature or behavior of the population. There exists a rich literature on outlier detection for the setting in which outliers are candidates for aberrant data that lead to biased or incorrect inferences. The general approach to cope with outliers in such circumstances is to clean up the data prior to modeling and performing the attendant statistical analysis [1]. Relevant outlier detection methods can be categorized under different taxonomies, the major ones being univariate versus multivariate methods and parametric versus nonparametric methods. Some popular approaches for such outlier detection approaches include Pierce's criterion [2], Chauvenet's criterion [3], and Dixon's test [4].

In contrast to the aforementioned notion of outlier detection that aims to render disturbance-free data, a less-investigated aspect of identifying outliers pertains to searching for rare and at the same time significant anomalies that do not conform to expected patterns and are often manifested as opportunities to be exploited (arising, e.g., in spectrum sensing) or risks to be ameliorated (e.g., network intrusion or fraud detection). In these settings, we can consider the outlying sequence detection problem as one in which a large number of sequences are being monitored simultaneously and the goal is to choose a small subset of sequences that are outliers. We refer to such problems as outlying sequence detection problems to distinguish them from the setting described in the previous paragraph in which a few outlier observations are winnowed out from a single set of data.

Detecting the outliers, especially in large data sets, is often very time-sensitive due to the transient nature of the opportunities that are attractive only when detected quickly, or due to the substantial costs that risks can incur if not managed swiftly. In this article, we focus on the fundamental problems in quick detection of outliers while recognizing different system- and physical-level constraints imposed by various contexts.

BACKGROUND

Outlier detection has immediate application in a broad range of contexts in which large volumes of data are constantly generated and processed. Some of these contexts and their application domains will be reviewed briefly in the section "Application Domains." While outlier observations in all contexts conform in representing unusual changes of the behavior of the underlying physical phenomena over one or more dimensions (e.g., time or space), the broad diversity in the range of the relevant applications necessitates diverse formulations that are customized to capture the specifics of each application domain. The remainder of this subsection focuses on reviewing some of the widespread models for abstracting the outlier detection problem in large data sets. The three major components for modeling the outliers and abstracting the outlier detection problem are the level of available information about the normal and outlying data streams, the type of the outliers, and the figure of merit for identifying the outliers. A comprehensive review of all such abstractions can be found in [5] and [6].

SUPERVISION LEVEL

Availability of information about the models for the data streams governs the modes and approaches for performing outlier detection in large data sets. Specifically, the existing approaches to outlier detection can be broadly categorized into four classes: supervised, semisupervised, unsupervised, and universal approaches, which are distinguished based on the availability of information about the structure of the data streams.

- *Supervised*: In the presence of prior information about the data streams (often acquired through training data) the models of both normal and abnormal (outlying) observations are known, which enables supervised outlier detection. These approaches are appropriate for static data or data models that evolve slowly enough so that tracking and learning the changes in the model are viable. In the statistics and computer science literature, the class of supervised outlier detection is studied extensively under classification-based approaches [7], [8], neural networks [9]–[11], Elman networks [12], naïve Bayes, and support vector machines [13].

- *Semisupervised*: In many practical circumstances acquiring models for both normal and outlying data streams is often infeasible. Based on the availability of information about a model for either normal or outlying sequences, semisupervised outlier detection approaches are developed, which capitalize on the known structure of normal (outlying) data streams to be robust against uncertainty about the structure of outlying (normal) data streams. While there exist scenarios that assume availability of information about the outlying sequences and lack of information about normal data streams [9], [14], these scenarios do not often arise. This is primarily due to the fact that outliers typically have an unpredictable nature and designing learning algorithms that can cover all possible outlying events is difficult. On the other hand, normal behavior is often well defined and thus it is more viable to construct models for normal data streams. Hence, in the majority of the existing literature on semisupervised outlier detection, the normal data streams are assumed to have known models while those of the outliers are unknown.
- *Unsupervised*: Under this category no assumption is made about models for the normal or outlying data streams and, instead, some other assumptions (e.g., parametric) are made about the models. In these approaches the normal observation are those that share a pattern occurring frequently and the outliers are those with rare and distinct patterns. Some representative unsupervised approaches include discriminative approaches [15]–[19], parametric approaches [18], [20]–[24], and online analytical processing (OLAP) approaches [25].
- *Completely universal*: Unlike in the supervised, semisupervised and unsupervised approaches, in the completely universal approach, no training data is available for either the typical or outlier distributions. As we discuss in the section "Universal Outlying Sequence Detection," it is possible to construct decision rules under this completely universal setting, with only the assumption that the typical and outlier distributions are different.

TYPES OF OUTLIERS

A pivotal step toward formulating any outlier detection approach is an abstraction for modeling the outliers. Here we review some of the more common categories of outliers, which are distinguished based on their composition and their relevance to normal observations.

- *Outlying points within a data stream*: This type of outlier occurs in circumstances when we are dealing with one data stream (often modeled as a time series) and one or more isolated elements of the stream do not conform to the common pattern of the data stream. Depending on whether the objective is to perform real-time or in-retrospect (offline) outlier detection, there are two different types of detection procedures. In real-time scenarios, the existing approaches often dynamically provide forecasts for the upcoming observations and, upon collecting the actual observations, a similarity measure between the actual observations and their forecast is computed. This measure determines whether the observation deviates from the expected pattern, and consequently whether it is an outlier or a normal observation [26]. In the offline outlier detection approaches, on the other hand, one popular approach is to cast the outlier detection problem as an in-retrospect change point detection problem [27].
- *Outlying subsequences within a data stream*: In contrast to outlying points, which appear sporadically and in isolation in one data stream, outlying subsequences appear in the form of consecutive outlying points. Similar to outlying points, detecting such outliers can be studied under real-time and offline settings. For the former there exist a body of window-based prediction approaches that form similarity measures for identifying outlying subsequences, and, for the latter, in-retrospect change point detection approaches are applicable.
- *Outlying data streams*: The previous two types of outliers occur within a data stream. Outlying data streams occur when we are given a large group of data streams, most of which follow a common pattern, but a few of which do not conform to this common pattern. Hence, there is no notion of outliers occurring within a stream anymore, but rather, each entire data stream is either normal or outlying. In such circumstances, the objective is to identify a group of sequences that exhibit behaviors different from the common pattern. This setting has been studied extensively in the statistics literature in which several approaches based on autoregression, moving average, and cumulative sum tests have been proposed with details reviewed in depth in [1], [5], and [28].

DECISION MECHANISM

Upon designing the information-gathering process and collecting observations, there are two broad schemes for forming a decision on individual observations or sets of observations and categorizing them as normal or outlying. In one approach, often termed the *labeling technique*, a binary decision about each individual observation is made. The outcome in this approach is a classification of the observations into two sets. The advantage of this approach is its accuracy in labeling every observation with a decision, while its drawback is that when the data volume increases, forming an accurate decision for every single observation is computationally prohibitive. In an alternative approach, often referred to as a *sorting technique*, each observation receives a score that indicates the likelihood of that observation being an outlier. The advantage of this technique is that it is less stringent in reaching an accurate decision for all observations in favor of enhancing the speed of the detection procedure, which makes it more suited for analyzing large data sets. The drawback of this approach, on the other hand, is that there should be a supplementary mechanism deciding about a threshold on the scores to delineate the normal and outlying regions.

APPLICATION DOMAINS

Different combinations of the different types of outliers, supervision level, and decision mechanisms (and other details reviewing, which is not relevant to the scope of this article)

create different abstractions for the outlier detection problem, each of which is relevant in certain application domains. Specifically, there exist a wide range of applications in which large volumes of data are constantly generated and the goal is to search for features or to identify anomalies that signify risks or opportunities. These goals can often be cast as outlier detection where the nature of the outliers, supervision level, the attendant decision mechanism, and other assumptions and constraints collectively formulate the underlying outlier detection problem. Examples of the application domains that involve detecting outliers in large data sets include credit card fraud detection [29], clinical trials [30], high-frequency trading [31], voting irregularity analysis [32], spectrum sensing [33], network intrusion [34], severe weather prediction [35], and seismic data analysis [36]. In this subsection we review a few application domains in which the problem of outlying sequence detection has important physical implications.

NETWORK INTRUSION

Network intrusion detection refers to detecting malicious penetrations to data networks. Intrusions exhibit behaviors different from the normal patterns in the network and the measurements associated with them can be modeled as outliers. The major impediment for identifying intrusions in this setting is the large volume of data, which makes the intrusion detection process computationally costly and time-consuming, while agile response to the presence of the intruders is crucial as any delay in detecting them leads to recovery costs for the system. Intrusions can often be modeled as outlying subsequences or sequences for which an observation model is unknown and, consequently, semisupervised or unsupervised approaches are best suited for identifying them. A comprehensive review of the literature on outlier detection approaches for network intrusion detection is available in [37].

FRAUD DETECTION

Fraud detection, which is the practice of identifying deliberately unlawful gains, is widely deployed by commercial entities including financial institutions, telecommunication companies, and insurance agencies. The pivotal step in designing fraud detection algorithms is creating profiles for usage activities of legitimate users and flagging any activity deviant from these profiles as a potential fraud. Hence, fraudulent activities can be modeled as outlying activities that should be identified swiftly to minimize the associated financial losses. A survey of different outlier detection approaches suited for credit card, mobile phone, insurance claim, and insider trading fraud detection is available in [37].

SPECTRUM SENSING

Wireless connectivity is ubiquitous and is constantly growing in scale and complexity to cope with the existing demands (e.g., data communication and sensor networks) and to accommodate the emerging ones (e.g., wireless health and smart grids). All such enabling technologies are viable at the expense of increasing demands for radio spectrum, which is the major commodity in the wireless industry. As reported by the U.S. Federal Communications Commission (FCC), exclusive spectrum access rights lead to underutilization of the spectrum. Driven by this observation and the urgency for higher spectral efficiency, future spectrum access policies are envisioned to provide the flexibility of dynamically granting spectrum access to unlicensed wireless services when the spectrum is underutilized by the license-holding services. Under such envisioned spectrum access policies, unlicensed services compete to make use of shared spectrum opportunities. The underutilized segments of the spectrum, hence, will not be as abundant as they otherwise should be and such reduction in their availability becomes even more severe as wireless sensing and networking grows in size and services. Hence, spectrum holes across wideband spectrum can be modeled as outliers in terms of their occupancy status and the problem of spectrum sensing in congested wideband spectrum can be abstracted as an outlier detection problem [33].

ENVIRONMENTAL MONITORING

The applications of outlier detection in environmental monitoring are multifaceted. Different forms of outlier detection are being used across the globe, e.g., for determining locations with constantly different temperatures from their neighbors, discovering drought areas, positioning fertility loss areas, and detecting hurricanes. A detailed overview of these application domains is available in [6].

DATA-ADAPTIVE OUTLYING SEQUENCE DETECTION

We introduce a general dichotomous hypothesis testing model for the outlying sequence detection problem of interest. This will be a unifying theme for investigating the problem under different settings. In this dichotomous model, we assume that the data set consists of M data streams, each being either a typical or an outlying sequence. Typical sequences exhibit identical statistical behavior, with which the outliers do not comply by exhibiting arbitrarily different known or unknown behaviors. The data volume increases as the number of data streams M increases, and in this article the focus is placed on high-dimensional data by performing the analysis in the asymptote of large values of M (i.e., $M \to \infty$). Furthermore, to emphasize the rarity of the outliers, we assume that the number of outliers grows sublinearly as M increases.

The above dichotomous model is adopted to mainly focus the attention on the discrepancy between the outliers and the typical observations and can be generalized to models that involve multiple statistical behaviors for the typical sequences. Each data stream generates independent and identically distributed (i.i.d.) real observations $\{Y_1^{(i)}, Y_2^{(i)}, \ldots\}$ obeying one of the two models

$$\begin{aligned} \mathrm{H}_0 &: \; Y_t^{(i)} \sim F, \;\; t = 1, 2, \ldots \\ \mathrm{H}_1 &: \; Y_t^{(i)} \not\sim F, \;\; t = 1, 2, \ldots, \end{aligned} \tag{1}$$

where F denotes a cumulative distribution function (cdf), modeling the statistical behavior of the typical sequences. Designing an

optimal outlying sequence detector rests fundamentally on delineating the inherent interplay between two opposing performance measures, one being the frequency of erroneous decisions and the other being the cost of sensing (e.g., the number of measurements taken). To this end, we consider the most general structure for the information-gathering process, which either sequentially, or based on a prespecified rule selects and takes measurements from a subset of the data streams at each time. By denoting the subset of data streams selected at time t by $\mathcal{L}_t \subseteq \{1, \ldots, M\}$, upon collecting the measurements at time t, the outlier detection process takes one of the following actions:

1) *Observation:* due to lack of sufficient information making any decision is deferred and the same set of data streams is retained for more scrutiny, i.e., $\mathcal{L}_{t+1} = \mathcal{L}_t$

2) *Exploration:* the information accumulated is insufficient to identify the outliers, but provides partial information that is sufficient for updating the set of data streams that should be measured more carefully, or possibly ruling out some of the data streams as typical ones, i.e., $\mathcal{L}_t \to \mathcal{L}_{t+1}$

3) *Detection:* the information gathering process is terminated and the outliers are identified.

The stopping time of the procedure, i.e., the time after which detection is performed, is denoted by τ. Furthermore, a switching function $\psi : \{1, \ldots, \tau\} \to \{0, 1\}$ is devised to distinguish between observation and exploration actions at time t. The switch is set to $\psi(t) = 0$ if it is decided in favor of performing observation at time t, while $\psi(t) = 1$ indicates a decision in favor of performing exploration. The sequential information-gathering procedure is uniquely determined by its stopping time τ, the sequence of switching functions $\bar{\psi}_\tau = [\psi(1), \ldots, \psi(\tau)]$, and the ordered collection $\overline{\mathcal{L}}_\tau \triangleq \{\mathcal{L}_1, \ldots, \mathcal{L}_{\tau-1}\}$.

The quality of the ultimate decision, which is the output of the detection action, is captured by the frequency of erroneous decisions. To formalize the dependence of such decision quality, on the given set of stopping time τ, switching sequence $\bar{\psi}_\tau$, and observation order $\overline{\mathcal{L}}_\tau$, we denote the frequency of erroneous decisions by $\mathrm{P}_M(\tau, \bar{\psi}_\tau, \overline{\mathcal{L}}_\tau)$. An optimal outlying sequence detection approach can be characterized as a strategy that optimizes a desired balance between this decision quality and the aggregate cost of sensing $\sum_{t=1}^{\tau} |\mathcal{L}_t|$, which incorporates the stopping time and the number of samples taken during the exploration cycles. Such a balance often can be cast as minimizing one of these measures, within a desired constraint on the other, e.g.,

$$\begin{aligned} &\min_{\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau} \ \mathbb{E}\left[\sum_{t=1}^{\tau} |\mathcal{L}_t|\right] \\ &\text{s.t.} \qquad \mathrm{P}_M(\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau) \le \rho, \end{aligned} \tag{2}$$

where ρ controls the decision reliability. In the following sections, we discuss several important topics under which the outlying sequence detection problem has different interpretations and can be cast as a balance between these measures.

Obtaining the optimal strategies for observation, exploration, and detection that strike a desired balance between decision quality and cost of sensing, in its most general form, is an open problem. By imposing certain structures on data or sampling models, however, one can delineate optimal strategies. In the remainder of the article, we discuss different outlying sequence detection approaches with different structures ranging from fully sequential to fully prespecified sampling strategies, and different objectives, ranging from identifying only one outlier to identifying all.

DATA-ADAPTIVE SAMPLING

In this section, we concretize the generic outlying sequence detection problem by focusing the attention on the closed-loop (adaptive) aspects of the sampling process. The extent of data adaptivity of the data-gathering process leads to a wide range of structures for the outlying sequence detection problem. Adaptivity is embedded in the sequential selection of the subset of data streams to be measured at each time, i.e., $\{\mathcal{L}_1, \ldots, \mathcal{L}_\tau\}$. Besides adaptivity in sensing, identifying the outliers can also be performed in either sequential or nonsequential fashion, where in the former the data collected is processed altogether to identify the outliers, whereas in the latter one could identify and remove an outlier and then search for other outliers among the remaining data streams.

QUICKEST SEARCH FOR ALL OUTLIERS

When the objective is to identify all outliers with minimum expected number of aggregate measurements and subject to controlled reliability, the problem is equivalent to forming a decision about the underlying model of all the sequences. Hence, the optimal sampling and decision-making problem can be decomposed into M independent hypothesis testing problems corresponding to the M sequences. The optimal solution to these latter subproblems is the sequential probability ratio test (SPRT), which minimizes the expected number of measurements required for forming a decision for each sequence with prespecified reliability [38] when the underlying distributions for normal and outlying observations are known.

These independent SPRTs can be performed either in parallel or sequentially. When performed in parallel, the sampling procedure is initiated by setting $\mathcal{L}_t = \{1, \ldots, M\}$ and after taking measurements at time t, the set $\mathcal{L}_t$ is refined by discarding the indices of the sequences for which their associated SPRT has reached a decision. In contrast, when the SPRTs are performed sequentially, the sampling strategy focuses on the sequences one at a time. While being effective in forming accurate decisions for individual data streams, performing independent SPRTs becomes computationally prohibitive as the size of the data set grows, and is not a suitable approach for large data sets.

QUICK SEARCH FOR A SUBSET OF OUTLIERS

In certain scenarios, one might be interested in recovering only a fraction of the outliers, especially when the outliers represent rare opportunities of interest, while in certain other scenarios, especially when the outliers model risks, it is imperative to identify all of the outliers.

Shifting the objective from recovering all the outliers to identifying only a fraction of them allows for missing some of the outliers in favor of quickly identifying the fraction of interest. Under this objective, performing SPRTs on all sequences is clearly not optimal as it tends to identify all sequences and does not take advantage of the more relaxed objective. Such a shift of objective and its ensuing flexibility leads to significant reduction in the sensing cost and the delay in reaching a decision.

Obtaining optimal structures of such sequential and data-adaptive experimental designs and finding associated nontrivial performance bounds for the such design are open for most scenarios, with some exceptions discussed in [39]–[41]. Nevertheless, by imposing certain structures on the refinement action, one can ascertain certain optimality properties with provable gains over nonadaptive approaches.

In this subsection, we focus on a specific structure studied in [33] and [42]–[45,] which consists of consecutive rounds of observations and exploration actions, followed by consecutive cycles of observations and satisfies certain optimality properties [45]. Driven by the premise that the outliers (anomalies) occur rarely, this adaptive structure starts by spending the sampling resources conservatively, and as more information about different data streams is accumulated, the sensing resources are progressively allocated to the data streams that are more likely outliers. The central motivation for such progressive allocation of the sensing resources is that while conservative (rough) observations are not accurate enough to identify the outliers, they can be informative enough to discard a considerable fraction of the typical streams. Consecutive cycles of rough observations and exploration, therefore, lead to substantial reduction in the search space, which facilitates using the sensing resources more effectively. Careful design of the exploration actions and the number of exploration actions, can provide sufficient guarantees that the discarded data streams are almost surely typical ones.

In this approach, more specifically, the sampling strategy is initiated by including all the streams for sampling and K consecutive cycles of exploration are performed, where K is determined by the amount of sampling resources and the fraction of the outliers one seeks to identify. The detailed steps of this procedure for identifying T outliers are provided in Table 1. In this procedure the exploration actions are designed such that at least T data streams will be retained after the exploration cycles for the final detection decision.

To assess adaptation gains, we formalize the adaptive experimental design problem as the minimizer of the decision quality under a hard constraint on the sampling budget, i.e.,

$$\mathcal{P}_M(S) \triangleq \begin{cases} \inf_{\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau} & \mathrm{P}_M(\tau,\overline{\psi}_\tau,\overline{\mathcal{L}}_\tau) \\ \text{s.t.} & \frac{1}{M}\sum_{t=1}^{\tau}|\mathcal{L}_t| \le S, \end{cases} \tag{3}$$

where S controls the sensing budget. Addressing the sensing problem in this setting sheds light on the ratio of the sensing resources to be allocated to the observation and exploration actions.

To assess the gains of adaptation we investigate the following two settings in which the typical distribution F is Gaussian with known mean and variance and the outliers are also Gaussian with either different mean or different variance values. Specifically, sequence i is generated according to $\mathcal{N}(\mu_i, \sigma_i^2)$. If sequence i is a typical sequence then $\mu_i = \mu$ and $\sigma_i = \sigma$, where μ and σ are known, and if it is an outlier sequence we consider two settings:

$$\begin{aligned} &\text{mean testing:} && \mu_i \neq \mu, \ \text{and} \ \ \sigma_i = \sigma \\ &\text{variance testing:} && \mu_i = \mu, \ \text{and} \ \ \sigma_i \neq \sigma. \end{aligned} \tag{4}$$

By defining $\overline{F}$ as the outlier cdf that exhibits the smallest Kullback–Leibler (KL) divergence from F, a necessary and sufficient condition for $\mathcal{P}_M(S) \xrightarrow{M\to\infty} 0$ to successively identify a small fraction of the outliers is presented in the following theorem. Here, a small fraction refers to a fraction that grows with M at a rate dominated by the growth rate of $M\theta_M$, where θ_M is the probability that a stream is an outlier.

[TABLE 1] THE ADAPTIVE OUTLYING SEQUENCE DETECTION ALGORITHM.

1) SET $\mathcal{L}_1 = \{1, \ldots, M\}$
2) FOR $t = 1$ TO K
3) TAKE ONE SAMPLE FROM EACH STREAM IN $\mathcal{L}_t$
4) SET $\beta_t = (1-\zeta)(|\mathcal{L}_t| - T)$ FOR A PRESPECIFIED CONSTANT $0 < \zeta < 1$
5) DISCARD β_t STREAMS THAT ARE MOST LIKELY TYPICAL
6) END FOR
7) SET $s = \left|\left(S - \sum_{t=1}^{K}|\mathcal{L}_t|\right)/|\mathcal{L}_K|\right|$
8) TAKE s SAMPLES FROM THE SURVIVING STREAMS
9) OUTPUT THE T SEQUENCES THAT ARE LEAST LIKELY TYPICAL

THEOREM 1

The decision error probability $\mathcal{P}_M(S)$ tends to zero in the asymptote of large M if and only if [43]

$$\text{mean testing:} \quad \frac{D(F\,\|\,\overline{F})}{\ln M} > \frac{(1-\sqrt{\varepsilon_M})^2}{\hat{S}+K}, \tag{5}$$

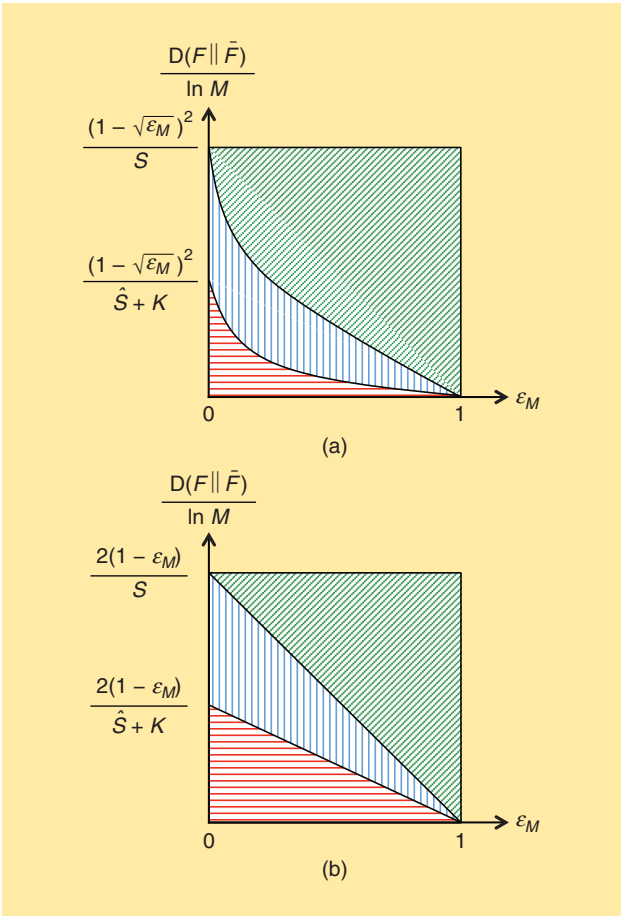
$$\text{variance testing:} \quad \frac{D(F\,\|\,\overline{F})}{\ln M} > \frac{2(1-\varepsilon_M)}{\hat{S}+K}, \tag{6}$$

where $D(\cdot\|\cdot)$ denotes the KL divergence, and $\hat{S}$ is a constant independent of M and determined by the constraints on the cost of sensing ($\hat{S} \approx S \cdot \zeta^{-K}$). Also $\varepsilon_M \in (0,1)$ is defined as

$$\varepsilon_M \triangleq \frac{\ln M\theta_M}{\ln M}, \tag{7}$$

where θ_M is the prior probability that a stream is an outlier.

The necessary and sufficient conditions on $D(F\,\|\,\overline{F})$ in Theorem 1 partition the (D, ε_M) plane into two regions separated by sharp boundaries, as shown in Figure 1. This figure also compares the regions over which the adaptive and nonadaptive procedures are guaranteed to make error-free decisions. Specifically, the diagonally shaded region is the region in which both schemes succeed to detect the T outliers. In the vertically dashed region, however, only the adaptive procedure succeeds and the nonadaptive procedure makes an erroneous decision almost surely, and finally both schemes fail in the horizontally shaded region. It is observed that, depending on the choice of S, the detectability region corresponding to the adaptive



[FIG1] $D(F\|\bar{F})$ versus the prior likelihood ε_M for Gaussian distributions with (a) a different mean and (b) different variance values.

procedure can be substantially larger than that corresponding to the nonadaptive procedure.

It is noteworthy that as long as the objective is to identify a small but prominent fraction of the outliers, the conditions given in (5) and (6) do not depend on the exact number of streams to be identified. This is due to the asymptotic nature of the results, which is dominantly shaped by the regime of interest (small fraction) and the precise number of the outliers has a vanishing effect as M grows. More general necessary and sufficient conditions for identifying any desired fraction of the outliers and with arbitrary distributions for the typical and outlier data streams are provided in [46].

QUICKEST SEARCH FOR ONE OUTLIER

In this subsection we discuss a special scenario of partial recovery of the outliers, in which the objective is to identify only one outlier. While the optimal sequential strategy for solving this problem, as discussed in the section "Data-Adaptive Sampling," is known, by imposing reasonable structures in sensing, some optimality properties can be ensured as $M \to \infty$. Specifically, when the sampling strategy is constrained to

1) observe only one data stream at a time, i.e., $|\mathcal{L}_t| = 1$ for all $t \in \{1, \ldots, \tau\}$
2) once a data stream is discarded after an exploration action, it will be discarded permanently, and the next stream to be examined will be selected randomly from the ones that remain
3) outliers have identical distributions denoted by $\bar{F}$,

the quickest search for detecting an outlier can be restated as

$$\begin{aligned} &\min_{\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau} \ \mathbb{E}[\tau] \\ &\text{s.t.} \qquad \mathrm{P}_M(\tau, \overline{\psi}_\tau, \overline{\mathcal{L}}_\tau) \le \rho. \end{aligned} \tag{8}$$

The sequential and data-adaptive sampling strategy that optimizes the above tradeoff between the average number of measurements and the decision quality (false alarm probability) is the cumulative sum (CUSUM) test [47]. In this test, one of the sequences is selected at random and measurements are taken from this sequence sequentially. After taking each sample and given all the information accumulated, the likelihood that the sequence under scrutiny is an outlier is updated. If this likelihood exceeds a certain threshold π_U, the sequence is declared an outlier; if it falls below a certain threshold π_L it is discarded permanently and another sequences will be selected to test; and if the likelihood remains within the interval $[\pi_L, \pi_U]$ another sample is taken from the same sequence. By defining π_t as the likelihood that the sequence observed at time t is an outlier given the information accumulated up to time t, the details of the optimal sampling strategy are presented in Table 2 with its optimality established by the following theorem.

THEOREM 2

The optimal stopping time for the quickest search problem in (8) is [48]

$$\tau = \inf\{t : \pi_t > \pi_U\},$$

and the optimal sampling strategy at time t switches to a new sequence if $\pi_t < \pi_L$. The thresholds π_L and π_U are determined uniquely as functions of ζ and the observation cdfs.

GROUP SAMPLING

Motivated by the insights gained from partial recovery of outliers, i.e., rough measurements can be sufficient for eliminating a substantial fraction of the typical streams through the exploration process, we next discuss the idea of group sampling, aiming at basing some of the decisions on even rougher measurements. Group sampling is facilitated by the possibility of taking samples that are combined measurements from multiple sequences. The ultimate objective of such measurements is to expedite the process of exploration and reduce the dimension of the search space with fewer measurements.

A central principle in designing the observation action in the previous section was that at any given time t, one measurement is taken from each data stream included in $\mathcal{L}_t$. In this section, in contrast, we consider two types of samples: coarse and fine samples, which bear information with different qualities. Coarse samples are constructed by linearly combining simultaneous measurements from a group of data streams. While such coarse

measurements are not informative for identifying the outliers, they can often be informative enough to discard a group of typical data streams altogether, especially when M is very large and the outliers occur very rarely. When such coarse samples are not sufficient to discard a block, or the block is deemed to contain an outlier, then the data streams constituting the blocks are measured individually via fine samples to refine the information about the status of the data streams within the block. Inclusion of coarse measurements reduces the required sampling budget.

Taking such coarse samples in some applications has a natural interpretation. For instance, in wideband spectrum sensing in which the majority of the channels are occupied and a mobile radio is interested in identifying rare spectrum opportunities (abstracted as outliers), due to the broadcast nature of the wireless channels, any measurement taken by the interested party is a linear superposition of the measurements that it can take from the channels individually via appropriate filtration.

For this purpose, we divide the data streams into blocks of size ℓ and take one sample that is a linear combination of ℓ measurements from the data streams. Such block sampling has, broadly, a twofold effect. On one hand, it takes only one sample for accumulating information about the ℓ sequences and is substantially smaller than the resources needed by the existing approaches that devote at least one sample to each sequence. On the other hand, one combined and aggregated sample is less informative about the status of the individual sequences in comparison to having ℓ different samples. To benefit from the advantage (reduction in sampling rate) and avoid its undesired effects (inaccurate information) these combined samples are used only to obtain some rough confidence about whether the block of data streams include outliers. When a block is deemed to include only typical data streams the entire block is discarded. Alternatively, if the block is deemed to include an outlier, then the block is retained for further scrutiny through more refined (fine) measurements.

QUICK SEARCH FOR A SUBSET OF OUTLIERS VIA GROUP SAMPLING

We define $r \triangleq M/\ell$ to be the number of blocks and without loss of generality we define

$$\mathcal{G}_i \triangleq \{(i-1)\ell + 1, \ldots, i\ell\} \tag{9}$$

to be the set of the data streams grouped in the ith block for $i \in \{1, \ldots, r\}$. With the ultimate objective of identifying T outliers the proposed sampling procedure is initiated by taking coarse samples from all groups $\mathcal{G}_1, \ldots, \mathcal{G}_r$. Based on these coarse observations a fraction of the groups that are least likely to contain outliers are discarded and the rest are retained for more accurate scrutiny. Repeating this procedure successively refines the search support and progressively focuses the observations on the more promising blocks. More specifically, at each time the sampling procedure selects a subset of the blocks $\{\mathcal{G}_1, \ldots, \mathcal{G}_r\}$ and takes one coarse sample from each of these blocks. Upon collecting these measurements, it takes one of the following actions:

[TABLE 2] THE QUICKEST SEARCH FOR ONE OUTLIER.

```
1)   INITIALIZE t = 0, φ_1 = 1, π_L, AND π_U
2)   t ← t + 1
3)   SET L_t = {φ_t}
4)   TAKE ONE SAMPLE FROM L_t
5)   UPDATE π_t
6)   IF φ_t ≤ π_L
7)     φ_{t+1} = φ_t + 1; GO TO 2
8)   ELSE IF π_L < π_t < π_U
9)     φ_{t+1} = φ_t; GO TO 2
10)  END IF
11)  SET τ = t; OUTPUT THE SEQUENCE φ_t
```

- *Observation:* Following the spirit of the generic observation action defined earlier, this action is taken in case of lack of sufficient confidence for deciding whether the blocks under scrutiny contain outliers.
- *Exploration:* There is sufficient confidence that some of the blocks are very unlikely to contain an outlier; discard a portion of the groups with the highest likelihoods of containing only typical data streams. This step can be designed similarly to the adaptive sampling procedure in Table 1.
- *Coarse sampling termination:* There is sufficient confidence that the blocks retained contain outliers; stop coarse sampling and start taking fine samples and perform SPRTs on individual sequences until an outlier is identified. If, after performing SPRTs on all sequences in the block, none is identified as an outlier, the sampling procedure resets by moving to the next block and starts taking coarse samples.

After terminating coarse sampling, the retained data streams contain a substantially more condensed proportion of outliers to typical data streams. When the block length $\ell > 1$ and the exploration action are designed carefully, while enjoying the same sensing budgets, adaptive group sampling yields a more reduced dimension for the search space compared with the adaptive procedure of the section "Quick Search for a Subset of Outliers" (i.e., $\ell = 1$). Similar to the mean and variance testing problems for partial recovery of the outliers presented in the section "Quick Search for a Subset of Outliers," the following theorem presents a necessary and sufficient condition for $\mathcal{P}_M(S) \xrightarrow{M\to\infty} 0$, for $\mathcal{P}_M(S)$ defined in (3), to successively identify a small fraction of the outliers. $\overline{F}$ denotes the outlier cdf that minimizes the KL divergence from F.

THEOREM 3

For fixed block size ℓ, the decision error probability $\mathcal{P}_M(S)$ tends to zero in the asymptote of large M if and only if [49]

$$\text{mean testing:} \quad \frac{D(F \,\|\, \overline{F})}{\ln M} > \frac{(1-\sqrt{\varepsilon_M})^2}{\ell(\hat{S}+K)},$$

$$\text{variance testing:} \quad \frac{D(F \,\|\, \overline{F})}{\ln M} > \frac{2(1-\varepsilon_M)}{\ell(\hat{S}+K)},$$

where $\hat{S}$ and $\varepsilon_M \in (0,1)$ are defined in Theorem 1.

This result indicates that as $M \to \infty$, the region of outliers that are undetectable by the adaptive procedure delineated by (5) and

(6) and depicted in Figure 1 is further shrunk by a factor of ℓ through group sampling.

QUICKEST SEARCH FOR ONE OUTLIER VIA GROUP SAMPLING

Similarly to the partial outlier recovery scenario, the quickest search approach of the section "Quickest Search for One Outlier" for identifying one outlier can be further extended by accommodating group sampling into the sampling strategy.

In the simplest scenario, the sequences can be bundled into groups of size $\ell = 2$ and the combined measurements taken will be the sum of two independent samples from each sequence. This leads to three possibilities for the distribution of the combined measurement. The sampling strategy is initiated by selecting a bundle at random and taking a mixed measurement from that sample and follows, in spirit, the same steps as the quickest search procedure in the section "Quickest Search for One Outlier." Specifically, when there is sufficient confidence that the group does not contain an outlier, the block is discarded; when there is a lack of confidence for making any reliable inference about the block, one more mixed sample is taken; and when there exists sufficient confidence that the block contains an outlier, taking combined measurements is terminated, and then the sequences contained in the block are examined individually to identify an outlier.

Designing the optimal sampling strategy involves characterizing two optimal stopping times, one corresponding to the terminal time of taking combined measurements, and the second one corresponding to reaching a decision for individual sequences after taking combined measurements is terminated. An effective approach for identifying these stopping times is proposed in [50], where a CUSUM test is applied to the sequence blocks to find a promising block, and then SPRTs are applied on the individual sequences to reach decisions about their underlying distributions.

UNIVERSAL OUTLYING SEQUENCE DETECTION

Depending on the underlying application, the underlying statistical models of the data streams might or might not be known. Whether the distributions of both typical and outlier sequences are known, only one is known, or both are unknown, outlier detection approaches can take drastically different structures. Representative examples are spectrum sensing in congested wideband channels as a case in which both distributions can be known (spectrum holes are the outliers) and fraud detection as a case in which either the outlier (fraud) or both distributions are unknown. When the statistics are fully known strategies that balance the interplay among different measures optimally can be characterized optimality according to the abstraction given in (2). These optimal strategies can be shown to be exponentially consistent and all the observation, exploration, and detection actions have likelihood-ratio-like structures [43].

When there exist uncertainties associated with the descriptions of the statistical models, the outlying sequence detection problem is related to general composite hypothesis testing problems, for which the generalized likelihood principle, which exhibits certain asymptotic optimality properties [51]–[53], is a popular solution. Universal outlying sequence detection is also closely related to homogeneity testing and classification [51], [54]–[58]. In homogeneity testing, one wishes to decide whether or not two samples come from the same probability law. In classification problems, a set of test data is classified to one of multiple streams of training data with distinct labels.

In this section, we investigate the effects of uncertainties about the statistics of the outliers and discuss a universal approach for identifying outliers in which, besides the premise that the outliers follow a distribution distinct from that governing the typical data streams, no knowledge of their statistics is assumed [59]. To focus the attention on the effects of unknown statistics, we mainly consider a simple setting in which it assumed that

1) only one data stream is an outlier and the remaining $M-1$ ones are typical
2) we have access to n samples from each data stream
3) the samples belong to a finite set $\mathcal{Y}$.

Under the hypothesis that the ith coordinate is the outlier, the joint distribution of all the observations (i.e., the likelihood function) is

$$p_i(\boldsymbol{y}^{Mn}) = p_i(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(M)}) = \prod_{t=1}^{n} \{\bar{f}(y_t^{(i)}) \prod_{j \neq i} f(y_t^{(j)})\}, \tag{10}$$

where

$$\boldsymbol{y}^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)}), i = 1, \ldots, M,$$

and $\bar{f}$ and f denote the probability mass functions (pmfs) of the outlier and typical streams, respectively.

For a universal detection rule $\delta : \mathcal{Y}^{Mn} \to \{1, \ldots, M\}$, which is not allowed to depend on f and $\bar{f}$, the maximal error probability, which will be a function of the test and $(\bar{f}, f)$, is

$$e(\delta, f, \bar{f}) \triangleq \max_{i=1,\ldots,M} \sum_{\boldsymbol{y}^{Mn}: \delta(\boldsymbol{y}^{Mn}) \neq i} p_i(\boldsymbol{y}^{Mn}), \tag{11}$$

with the corresponding error exponent, denoted by

$$\alpha(\delta, f, \bar{f}) \triangleq \lim_{n \to \infty} -\frac{1}{n} \log e(\delta, f, \bar{f}). \tag{12}$$

We consider the error exponent as n goes to infinity, while M, and hence the number of hypotheses, is kept fixed. Consequently, the error exponent in (12) also coincides with the one for the average probability of error.

A test is termed *universally consistent* if $e(\delta, f, \bar{f}) \to 0$ for any $(\bar{f}, f)$, $\bar{f} \neq f$ as $n \to \infty$. It is termed *universally exponentially consistent* if $\alpha(\delta, f, \bar{f}) > 0$.

UNIVERSAL TEST

For each $i = 1, \ldots, M$, denote the empirical distribution of $\boldsymbol{y}^{(i)}$ by γ_i. When f is known and $\bar{f}$ is unknown, we compute the likelihood for outlier hypothesis i by replacing $\bar{f}$ in (10) with its maximum likelihood (ML) estimate $\hat{\bar{f}}_i \triangleq \gamma_i$, as

$$L_i^{\text{typ}}(\boldsymbol{y}^{Mn}) = \prod_{t=1}^{n} \{\hat{\bar{f}}_i(y_t^{(i)}) \prod_{j \neq i} f(y_t^{(j)})\}. \tag{13}$$

Similarly, when neither $\bar{f}$ nor f is known, we compute the likelihood for outlier hypothesis i by replacing the $\bar{f}$ and f in (10) with their ML estimates $\hat{\bar{f}}_i \triangleq \gamma_i$, and $\hat{f}_i \triangleq \left(\sum_{j\neq i}\gamma_j\right)/(M-1)$, as

$$L_i^{\text{univ}}(y^{Mn}) = \prod_{t=1}^{n}\{\hat{\bar{f}}_i(y_t^{(i)})\prod_{j\neq i}\hat{f}_i(y_t^{(j)})\}. \tag{14}$$

Finally, we decide upon the coordinate with the largest likelihood to be the outlier. Using (13) and (14), our universal tests in the two cases can be described respectively as

$$\delta^{\text{typ}}(y^{Mn}) = \underset{i=1,\ldots,M}{\operatorname{argmax}}\, L_i^{\text{typ}}(y^{Mn}), \tag{15}$$

when only f is known, and

$$\delta^{\text{univ}}(y^{Mn}) = \underset{i=1,\ldots,M}{\operatorname{argmax}}\, L_i^{\text{univ}}(y^{Mn}), \tag{16}$$

when neither $\bar{f}$ nor f is known.

RESULTS

Our results will be stated in terms of a distance metric between a pair of pmfs $p, q \in \mathcal{P}(\mathcal{Y})$ called the Bhattacharyya distance, which is related to the Chernoff information (see, e.g., [60]), defined as

$$B(p,q) \triangleq -\log\Big(\sum_{y\in\mathcal{Y}} p(y)^{\frac{1}{2}} q(y)^{\frac{1}{2}}.\Big) \tag{17}$$

Our first theorem for models with one outlier characterizes the optimal exponent for the maximal error probability when both $\bar{f}$ and f are known, and when only f is known.

THEOREM 4

When $\bar{f}$ and f are both known, the optimal exponent for the maximal error probability is equal to [59]

$$2B(\bar{f}, f). \tag{18}$$

Furthermore, the error exponent in (18) is achievable by a test that uses only the knowledge of f. In particular, such a test is our proposed test in (15).

Consequently, in the completely universal setting, when nothing is known about $\bar{f}$ and f except that $\bar{f}\neq f$, and both $\bar{f}$ and f have full supports, it holds that for any universal test δ,

$$\alpha(\delta, f, \bar{f}) \leq 2B(\bar{f}, f). \tag{19}$$

Given the second assertion in Theorem 4, it might be tempting to think that it would be possible to design a test to achieve the optimal error exponent of $2B(\bar{f}, f)$ universally when neither $\bar{f}$ nor f is known. A counterexample given in [59] shows that this is not possible. This motivates us to seek instead a test that yields just a positive (no matter how small) error exponent $\alpha(\delta, f, \bar{f}) > 0$ for every $\bar{f}$ and f, $\bar{f}\neq f$, i.e., a test that achieves universally exponential consistency. Without knowing either $\bar{f}$ or f, it is not clear at the outset that even this lesser objective can be met. One of the main contributions in [59] is to show that the proposed universal test in (16) is indeed universally exponentially consistent for every fixed M.

THEOREM 5

For every pair $\bar{f}\neq f$

$$\alpha(\delta^{\text{univ}}, f, \bar{f}) = \min_{q_1,\ldots,q_M} D(q_1\|\bar{f}) + \ldots + D(q_M\|f),$$

where the minimum is over the set of $(q_1, \ldots, q_M)$ such that

$$\sum_{j\neq 1} D\left(q_j \middle\| \frac{\sum_{k\neq 1} q_k}{M-1}\right) \geq \sum_{j\neq 2} D\left(q_j \middle\| \frac{\sum_{k\neq 2} q_k}{M-1}\right). \tag{20}$$

It can be shown that the solution $\alpha(\delta^{\text{univ}}, f, \bar{f}) > 0$ [59].

Note that for any fixed $M \geq 3$ and $\theta > 0$, regardless of which coordinate is the outlier, it holds that the random empirical distributions $(\gamma_1, \ldots, \gamma_M)$ satisfy

$$\lim_{n\to\infty} \mathbb{P}_i\Big\{\Big\|\frac{1}{M}\sum_{j=1}^{M}\gamma_j - \Big(\frac{1}{M}\bar{f} + \frac{M-1}{M}f\Big)\Big\|_1 > \theta\Big\} = 0, \tag{21}$$

where $\|\cdot\|_1$ denotes the 1-norm of the argument distribution. Since $(1/M)\bar{f} + (M-1/M)f \to f$ as $M\to\infty$, heuristically speaking, a consistent estimate of the typical distribution can readily be obtained asymptotically in M at the outset from the entire observation set before deciding upon which coordinate is the outlier. This observation and the second assertion of Theorem 4 motivate a study of the asymptotic performance (achievable error exponent) of δ^{univ} when $M\to\infty$ (after having taken the limit as n goes to infinity).

THEOREM 6

For each $M \geq 3$

$$\alpha(\delta^{\text{univ}}, f, \bar{f}) \geq \min_{\substack{q\in\mathcal{P}(\mathcal{Y}) \\ D(q\|f)\leq\frac{1}{M-1}(2B(\bar{f},f)+C_f)}} 2B(\bar{f}, q), \tag{22}$$

where $C_f \triangleq -\log(\min_{y\in\mathcal{Y}} f(y)) < \infty$ by the fact that f has a full support [59].

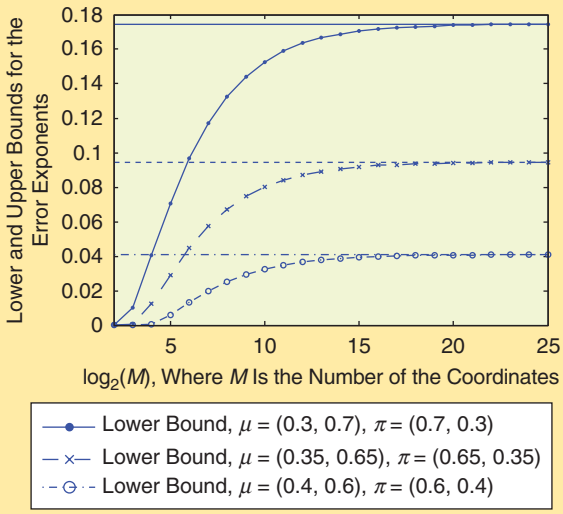
The lower bound on the error exponent in (22) is nondecreasing in $M \geq 3$. Furthermore, as $M\to\infty$, this lower bound converges to the optimal error exponent $2B(\bar{f}, f)$; hence, our test is asymptotically optimal:

$$\lim_{M\to\infty} \alpha(\delta^{\text{univ}}, f, \bar{f}) = 2B(\bar{f}, f), \tag{23}$$

which from Theorem 4 is equal to the optimal error exponent when both $\bar{f}$ and f are known.

Example 1

We now provide some numerical results for an example with $\mathcal{Y} = \{0, 1\}$. Specifically, the three plots in Figure 2 are for three pairs of outlier and typical distributions being $\bar{f} = (p(0) = 0.3, p(1) = 0.7)$, $f = (0.7, 0.3)$; $\bar{f} = (0.35, 0.65)$, $f = (0.65, 0.35)$; and $\bar{f} = (0.4, 0.6)$, $f = (0.6, 0.4)$, respectively. Each horizontal line corresponds to $2B(\bar{f}, f)$, and each curved line corresponds to the lower bound in (22) for the error exponent achievable by δ^{univ}. As



[FIG2] An illustration of the asymptotic optimality of δ^{univ}.

shown in these plots, the lower bounds converge to $2B(\bar{f}, f)$ as $M \to \infty$, i.e., δ^{univ} is asymptotically optimal for all three pairs $\bar{f}, f$.

MODELS WITH AT MOST ONE OUTLIER

A natural question that arises at this point is what would happen if it is also possible that no outlier is present? To answer this question, we now consider models that append an additional null hypothesis with no outlier to the set of possible hypotheses. In particular, under the null hypothesis, the likelihood function is given by

$$p_0(y^{Mn}) = \prod_{t=1}^{n} \prod_{i=1}^{M} f(y_t^{(i)}).$$

A universal test $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$ will now also accommodate an additional decision for the null hypothesis. Correspondingly, the maximal error probability is now computed with the inclusion of the null hypothesis according to

$$e(\delta, f, \bar{f}) \triangleq \max_{i=0,1,\ldots,M} \sum_{y^{Mn}: \delta(y^{Mn}) \neq i} p_i(y^{Mn}).$$

With just one additional null hypothesis, contrary to the previous models with one outlier, it becomes impossible to achieve universal exponential consistency even with the knowledge of the typical distribution. This pessimistic result reaffirms that our previous finding that universal exponential consistency is attained for the models with one outlier is indeed quite surprising.

PROPOSITION 1

For the setting with the additional null hypothesis, there cannot exist a universally exponentially consistent test even when the typical distribution is known [59].

In typical applications such as environment monitoring and fraud detection, the consequence of a missed detection of the outlier can be much more catastrophic than that of a false positive. In addition, Proposition 1 tells us that there cannot exist a universal test that yields exponential decays for both the conditional probability of false positive (under the null hypothesis) and the conditional probabilities of missed detection (under all nonnull hypotheses). Consequently, it is natural to look for a universal test fulfilling a lesser objective: attaining universal exponential consistency for conditional error probabilities under only all the nonnull hypotheses, while seeking only universal consistency for the conditional error probability under the null hypothesis. We now show that such a test can be obtained by slightly modifying our earlier test. Furthermore, in addition to achieving universal consistency under the null hypothesis, this new test achieves the same exponent as in (20) in Theorem 5 for the conditional error probabilities under all nonnull hypotheses.

In particular, we modify our previous test in (16) to allow for the possibility of deciding for the null hypothesis as:

$$\delta^{\text{null}}(y^{Mn}) = \begin{cases} \arg\max\limits_{i=1,\ldots,M} L_i^{\text{univ}}(y^{Mn}) & \text{if } \max\limits_{j \neq k} \dfrac{L_j^{\text{univ}}(y^{Mn})}{L_k^{\text{univ}}(y^{Mn})} > \lambda_n \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where $\lambda_n = O(n)$.

THEOREM 7

For every pair of distributions $\bar{f}, f, \bar{f} \neq f$, the test in (24) yields a positive exponent for the conditional probability of error under every nonnull hypothesis $i = 1, \ldots, M$, and a vanishing conditional probability of error under the null hypothesis [59]. In particular, the achievable error exponent under every nonnull hypothesis is the same as that given in (20).

Furthermore, as $M \to \infty$, the test in (24) is asymptotically optimal under each of the nonnull hypotheses, i.e.,

$$\lim_{M \to \infty} \lim_{n \to \infty} -\frac{1}{n} \log(\mathbb{P}_i\{\delta \neq i\}) = 2B(\bar{f}, f), \tag{25}$$

while also yielding that

$$\lim_{n \to \infty} \mathbb{P}_0\{\delta \neq 0\} = 0.$$

EXTENSION TO MULTIPLE OUTLIERS

The aforementioned results on universal outlying sequence detection can be extended to the setting with more than one outlier [59]:

- For the setting with a known number of distinctly distributed outliers, we can construct a universally exponentially consistent test using the generalized likelihood principle as in the section "Universal Test." A key difference from the single outlier case is that the error exponent when both the outlier and typical distributions are known can be larger than that when only the typical distribution is known.
- For the setting with a known number of identically distributed outliers, the error exponent when both the outlier and typical distributions are known is equal to that when only the typical distribution is known, which is equal to $2B(\bar{f}, f)$ (the same as for the case of a single outlier). Furthermore, the universally exponentially consistent test when both the outlier

and typical distributions are unknown is asymptotically optimum as $M \to \infty$ (with the number of outliers fixed) in that its error exponent is also equal to $2B(\bar{f}, f)$.

- For the setting with an unknown number of identically distributed outliers, we construct a test based on modified application of the generalized likelihood principle to achieve a positive error exponent under each nonnull hypothesis, and also consistency under the null hypothesis universally.
- When the outliers can be distinctly distributed (with their total number being unknown), it can be shown that a universally exponentially consistent test cannot exist, even when the typical distribution is known and the null hypothesis is excluded.

CONCLUDING REMARKS

In this article, we have discussed the problem of identifying outlying sequences from a large pool of sequences that is populated by typical sequences. By crafting the problem as a natural dichotomous hypothesis testing problem, we have discussed three general classes of strategies for outlying sequence detection based on different detection objectives and available information about the statistics of the outliers. In this class, we have discussed sequential data-adaptive approaches in which there is no prespecified order for making measurements from the sequences, and the sampling decisions are made dynamically at each time and based on the information accumulated up to that time. Depending on whether one is interested in identifying all outlying sequences, a fraction of them, or only one of them, the data-adaptive sampling strategies exhibit different structures. An important insight one gains from these approaches is that if the objective is not identifying all outliers, incorporating an exploration stage, which uses rough observations to reduce the dimension of the data set with more condensed proportion of outliers, translates into substantial reduction in the cost of sensing. Motivated by this insight, in the second class of approaches we have discussed the notion of group sampling, in which the sequences are split into groups and in the exploration stage the sequences are not measured individually, but instead, rough observations in the form of combined measurements from sequence groups are made. Finally, in the third class, we have investigated the effects of uncertainties about the statistics of the outliers and have discussed a universal approach for identifying outliers, in which besides the premise that the outliers follow a distribution distinct from that governing the typical data streams, no knowledge of their statistics is assumed. Our generalized likelihood approach was based on using the empirical distributions of the data streams. A recent study [61] adopts an alternative kernel-based test, which applies the metric of maximum mean discrepancy that measures the distance between embeddings of distributions into a reproducing kernel Hilbert space. We further note that in our discussion of universal outlying sequence detection, we have restricted attention to the fixed sample size setting in which every sequence is sampled at every time step. Extending the study of universal outlying sequence detection to the sequential and adaptive sampling settings is a challenging open area of research that is worthy of pursuit.

ACKNOWLEDGMENTS

The work of Ali Tajer was supported by the National Science Foundation (NSF) under grant ECCS-1343326. The work of Venugopal V. Veeravalli was supported by the Air Force Office of Scientic Research under grant FA9550-10-1-0458 through the University of Illinois at Urbana-Champaign, by the U.S. Defense Threat Reduction Agency through subcontract 147755 at the University of Illinois from prime award HDTRA1-10-1-0086, and by the NSF under grant NSF CCF 11-11342. The work of H. Vincent Poor was supported by the NSF under grants DMS-1118605 and ECCS-1343210.

AUTHORS

Ali Tajer (tajer@ecse.rpi.edu) received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology in 2002 and 2004, respectively, and the M.A. and Ph.D. degrees from Columbia University in 2010 in statistics and electrical engineering, respectively. He was with Princeton University from 2010 to 2012 as a postdoctoral research associate. He is currently an assistant professor of electrical, computer, and systems engineering at Rensselaer Polytechnic Institute. His research interests include estimation and detection theory, network information theory, wireless communications, and smart grids.

Venugopal V. Veeravalli (vvv@illinois.edu) received the B. Tech. degree (Silver Medal Honors) from the Indian Institute of Technology, Bombay, in 1985, the M.S. degree from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1987, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, in 1992, all in electrical engineering. He joined the University of Illinois at Urbana-Champaign in 2000, where he is currently a professor in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the Information Trust Institute. His research interests include detection and estimation theory, information theory, sensor networks, and wireless communication.

H. Vincent Poor (poor@princeton.edu) is the dean of engineering and applied science at Princeton University, where he is also the Michael Henry Strater University Professor of Electrical Engineering. His interests include the areas of statistical signal processing, stochastic analysis, and information theory, with applications in wireless networks and related fields. Among his publications is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge, 2014). He is an IEEE Fellow and a member of the National Academy of Engineering, the National Academy of Sciences, and the Royal Society. His recent recognitions include the 2011 IEEE Signal Processing Society Award and the 2014 URSI Booker Gold Medal.

REFERENCES

[1] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. Hoboken, NJ: Wiley, 2003.

[2] B. Pierce, "Criterion for the rejection of doubtful observations," *Astron. J.*, vol. 2, no. 21, pp. 161–163, July 1852.

[3] W. Chauvenet, *A Manual of Spherical and Practical Astronomy*, 5th ed. New York: Dover, 1960, vol. II.

[4] R. B. Dean and W. J. Dixon, "Simplified statistics for small numbers of observations," *Anal. Chem.*, vol. 24, no. 4, pp. 636–638, Apr. 1951.

[5] V. Barnett and T. Lewis, *Outliers in Statistical Data*. New York: Wiley, 1978.

[6] M. Gupta, J. Gao, C. Aggarwal, and J. Han, *Outlier Detection for Temporal Data*. San Rafael, CA: Morgan and Claypool, 2014.

[7] P.-N. Tand, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Reading, MA: Addison-Wesley, 2005, ch. 2, pp. 19–96.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley Interscience, 2000.

[9] D. Dasgupta and N. Majumdar, "Anomaly detection in multidimensional data using negative selection algorithm," in *Proc. IEEE Conf. Evolutionary Computation*, Hawaii, 2002, pp. 1039–1044.

[10] D. Endler, "Intrusion detection applying machine learning to solaris audit data," in *Proc. 14th Annu. Computer Security Applications Conf.*, 1998, pp. 268–279.

[11] A. K. Gosh, J. Wanken, and F. Charron, "Detecting anomalous and unknown intrusions against programs," in *Proc. 14th Annu. Computer Security Applications Conf.*, 1998, pp. 259–267.

[12] A. Ghosh, A. Schwartzbard, and M. Schatz, "A study in using neural networks for anomaly and misuse detection," in *Proc. 8th Conf. USENIX Security Symp.*, 1999, pp. 12–23.

[13] D.-K. Kang, D. Fuller, and V. Honavar, "Learning classifiers for misuse detection using a bag of system calls representation," in *Proc. 3rd IEEE Int. Conf. Intelligence and Security Informatics*, 2005, pp. 511–516.

[14] D. Dasgupta and F. Nino, "A comparison of negative and positive selection algorithms in novel pattern detection," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, Nashville, TN, 2000, pp. 125–130.

[15] T. Lane and C. Brodley, "Sequence matching and learning in anomaly detection for computer security," in *Proc. AAAI Workshop: AI Approaches to Fraud Detection and Risk Management*, 1997, pp. 43–49.

[16] S. Budalakoti, A. Srivastava, R. Akella, and E. Turkov, "Anomaly detection in large sets of high-dimensional symbol sequences," NASA Ames Research Center, Tech. Rep. TM-2006-214553, 2006.

[17] S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Trans. Syst., Man, Cybern. C*, vol. 39, no. 1, pp. 101–113, Jan. 2009.

[18] V. Chandola, V. Mithal, and V. Kumar, "A comparative evaluation of anomaly detection techniques for sequence data," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 743–748.

[19] K. Sequeira and M. Zaki, "ADMIT: anomaly-based data mining for intrusions," in *Proc. 8th ACM Int. Conf. Knowledge Discovery and Data Mining*, 2002, pp. 386–395.

[20] C. Marceau, "Characterizing the behavior of a program using multiple-length N-grams," in *Proc. Workshop New Security Paradigms*, 2000, pp. 101–110.

[21] C. C. Michael and A. Ghosh, "Two state-based approaches to program-based anomaly detection," in *Proc. 16th Annu. Computer Security Applications Conf.*, 2000, pp. 21–30.

[22] E. Eskin, W. Lee, and S. Stolfo, "Modeling system calls for intrusion detection with dynamic window sizes," in *Proc. DARPA Information Survivability Conf. Expo. II*, 2001, pp. 165–175.

[23] G. Florez-Larrahondo, S. M. Bridges, and R. Vaughn, "Efficient modeling of discrete events for anomaly detection using hidden Markov models," in *Proc. 8th Int. Conf. Information Security*, 2005, pp. 506–514.

[24] B. Gao, H.-Y. Ma, and Y.-H. Yang, "HMMs (Hidden Markov Models) based on anomaly intrusion detection method," in *Proc. Int. Conf. Machine Learning and Cybernetics*, 2002, pp. 381–385.

[25] X. Li and J. Han, "Mining approximate top-K subspace anomalies in multi-dimensional time-series data," in *Proc. 33rd Int. Conf. Very Large Databases*, 2007, pp. 447–458.

[26] N. D. Le, R. D. Martin, and A. E. Raftery, "Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models," *J. Am. Stat. Assoc.*, vol. 91, no. 436, pp. 1504–1515, 1996.

[27] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4215–4229, July 2012.

[28] D. M. Hawkins, *Identification of Outliers*. London, U.K.: Chapman & Hall, 1980.

[29] R. J. Bolten and D. J. Hand, "Unsupervised profiling methods for fraud detection," in *Proc. Credit Scoring and Credit Control Conf.*, vol. VII, Edinburgh, Scotland, 2001, pp. 3–5.

[30] K. I. Penny and I. T. Jolliffe, "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *Statistician*, vol. 50, no. 3, pp. 295–308, 2001.

[31] C. Brownlees and G. Gallo, "Financial econometric analysis at ultra-high frequency: Data handling concerns," *Computat. Stat. Data Anal.*, vol. 51, no. 4, pp. 2232–2245, Dec. 2006.

[32] R. M. Alvarez, S. D. Hyde, and T. E. Hall, Eds., *Election Fraud: Detecting and Deterring Electoral Manipulation*, ser. Brookings Series on Election Administration and Reform. Washington, DC: Brookings Institution, 2008.

[33] A. Tajer, R. Castro, and X. Wang, "Adaptive sensing of congested spectrum bands," *IEEE Trans. Inform. Theory*, vol. 58, no. 9, pp. 6110–6125, Sept. 2012.

[34] A. Lazarevic, A. Ozgur, L. Ertoz, J. Srivastava, and V. Kumar, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proc. 3rd SIAM Int. Conf. Data Mining*, San Francisco, CA, May 2003, pp. 25–36.

[35] M. D. Goldberg, Y. Qu, L. M. McMillin, W. Wolf, L. Zhou, and M. Divakarla, "AIRS near-real-time products and algorithms in support of operational numerical weather prediction," *IEEE Trans. Geosci. Remote Sensing*, vol. 41, no. 2, pp. 379–389, Feb. 2003.

[36] S. Wanga, W. A. Woodward, H. L. Grayb, S. Wiecheckib, and S. R. Sain, "A new test for outlier detection from a multivariate mixture distribution," *J. Computat. Graph. Stat.*, vol. 6, no. 3, pp. 285–299, 1997.

[37] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15.1–15.58, July 2009.

[38] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Ann. Math. Stat.*, vol. 19, no. 3, pp. 326–339, 1948.

[39] M. V. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problemy Peredachi Informastii*, vol. 10, no. 3, pp. 51–61, 1974.

[40] A. Korostelev, "On minimax rates of convergence in image models under sequential design," *Stat. Probability Lett.*, vol. 43, no. 4, pp. 369–375, July 1999.

[41] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.

[42] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6222–6235, Sept. 2011.

[43] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Trans. Inform. Theory*, vol. 59, no. 7, pp. 4462–4481, 2013.

[44] A. Tajer and H. V. Poor, "Adaptive sampling for sparse recovery," in *Proc. 4th Workshop on Information Theoretic Methods in Science and Engineering*, Helsinki, Finland, Aug. 2011.

[45] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, 2013.

[46] A. Tajer and H. V. Poor, "Hypothesis testing for partial sparse recovery," in *Proc. 50th Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2012, pp. 901–908.

[47] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, UK: Cambridge Univ. Press, 2009.

[48] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis, "Quickest search over multiple sequences," *IEEE Trans. Inform. Theory*, vol. 57, no. 8, pp. 5375–5386, Aug. 2011.

[49] A. Tajer and H. V. Poor, "Quick search for rare events through adaptive group sampling," in *Proc. 47th Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2013, pp. 757–761.

[50] J. Geng, W. Xu, and L. Lai, "Quickest search over multiple sequences with mixed observations," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, July 2013, pp. 2582–2586.

[51] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer, 1994.

[52] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?," *IEEE Trans. Inform. Theory*, vol. 38, no. 5, pp. 1597–1602, Sept. 1992.

[53] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Stat.*, vol. 36, no. 2, pp. 369–401, Apr. 1965.

[54] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, no. 1–2, pp. 250–254, July 1911.

[55] O. Shiyevitz, "On Rényi measures and hypothesis testing," in *Proc. IEEE Int. Symp. Information Theory*, July 31–Aug. 5, 2011, pp. 894–898.

[56] J. Unnikrishnan, "On optimal two sample homogeneity tests for finite alphabets," in *Proc. IEEE Int. Symp. Information Theory*, July 1–6, 2012, pp. 2027–2031.

[57] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inform. Theory*, vol. 34, no. 2, pp. 278–286, Mar. 1988.

[58] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[59] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," in *Proc. IEEE Int. Symp. Information Theory*, Istanbul, Turkey, July 2013, pp. 2666–2670.

[60] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 2006.

[61] S. Zou, Y. Liang, H. V. Poor, and X. Sh, "Kernel-based nonparametric anomaly detection," in *Proc. IEEE 15th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Toronto, Canada, June 2014.

[SP]