Linear-Complexity Exponentially-Consistent Tests for Universal Outlying Sequence Detection

Yuheng Bu, *Student Member, IEEE*, Shaofeng Zou, *Member, IEEE*, and Venugopal V. Veeravalli, *Fellow, IEEE*

***Abstract*—The problem of universal outlying sequence detection is studied, where the goal is to detect outlying sequences among M sequences of samples. A sequence is considered as outlying if the observations therein are generated by a distribution different from those generating the observations in the majority of the sequences. In the universal setting, we are interested in identifying all the outlying sequences without knowing the underlying generating distributions. We consider the outlying sequence detection problem in three different scenarios: first, known number of outlying sequences; second, unknown number of identical outlying sequences; and finally, typical and outlying distributions forming clusters. In this paper, a class of tests based on distribution clustering is proposed. These tests are shown to be exponentially consistent with *linear* time complexity in M. Numerical results demonstrate that our clustering-based tests achieve similar performance to existing tests, while being considerably more computationally efficient.**

***Index Terms*—Anomaly detection, clustering algorithm, exponential consistency, outlier detection, universal outlier hypothesis testing.**

I. Introduction

We study a universal outlying sequence detection problem, where the objective is to detect outlying sequences among M sequences of samples. Each sequence consists of n independent and identically distributed (i.i.d.) discrete observations. It is assumed that the observations in the majority of the sequences are distributed according to typical distributions. A sequence is considered as outlying if its distribution is different from the typical distributions. We are interested in the universal setting of the problem, where we do not know the probability mass functions (pmfs) of both the typical and outlying distributions, which are assumed to have full support over a finite alphabet. Moreover, we consider the following three scenarios for the outlying sequences detection problem in this paper: (1) known number of outlying sequences; (2) unknown number of identical outlying sequences; and (3) typical and outlying distributions forming clusters. The goal is to design universal tests, which do not depend on the typical and outlying distributions, to efficiently discern all the outlying sequences in these three different scenarios.

Outlying sequence detection finds possible applications in many domains [2]. For example, in cognitive wireless networks, channel measurements follow different distributions depending on whether the channel is busy or vacant. In order to utilize the vacant channels for improving spectral efficiency, such a network need to efficiently identify vacant channels out of a large number busy channels based on their corresponding signals. Other applications include anomaly detection in large data sets [3], security monitoring in sensor networks [4], and detecting an epidemic disease with aberrant genetic markers [5]. All of these applications require a reliable algorithm that can be implemented with low time complexity.

In the universal outlying sequence detection problem, we have no prior knowledge and no training data to learn these distributions before hand. Thus, the major challenges to solve this problem lie in: (1) building distribution-free consistent tests, and further guaranteeing their exponential consistency for distinct typical and outlying distributions; and (2) designing low-complexity tests that can be used in practical applications. To address theses challenges, we propose tests based on distribution clustering [6] for various scenarios. We show that our tests are exponentially consistent, with time complexity that is linear in the number of sequences M and independent of the number of outlying sequences T.

The basic idea behind our clustering-based tests is that if we observe a sequence of samples from each distribution, the empirical distributions of the sequences will converge to the true distributions as the number of samples goes to infinity. Moreover, the typical distributions (and also possibly the outlying distributions) usually form a cluster in the three scenarios considered in this paper. The typical distributions are thus closer to each other than to the outlying distributions. This suggests that the outlying sequence detection problem can be solved by clustering the empirical distributions using KL divergence as the distance metric (see also [7]).

We note that our clustering-based tests are closely related to the classical distribution clustering problem [6], [8]–[11], but there are essential differences. In the distribution clustering problem, the goal is to construct low-complexity algorithm to find the cluster structure of distributions with the lowest cost (sum of distance functions of each distribution in the cluster to

Manuscript received December 4, 2017; revised September 10, 2018, November 13, 2018, and January 3, 2019; accepted February 13, 2019. Date of publication February 25, 2019; date of current version March 15, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Romain Couillet. This work was supported in part by the National Science Foundation under Grants NSF 11-11342 and 1617789, through the University of Illinois at Urbana-Champaign. This paper was presented in part at the IEEE International Symposium on Information Theory, Aachen, Germany, June 2017 [1]. *(Corresponding author: Venugopal V. Veeravalli.)*

Y. Bu and V. V. Veeravalli are with the ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: bu3@illinois.edu; vvv@illinois.edu).

S. Zou is with the Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14228 USA (e-mail: szou3@buffalo.edu).

Digital Object Identifier 10.1109/TSP.2019.2901364

the center). Whereas in our problem, we are given samples from each distribution rather than the actual underlying distribution itself. Since we are considering a detection problem, we are more interested in the error probability of our test, rather than the minimal clustering cost. Previous studies on approximation algorithms for distribution clustering [8]–[10] only show that by carefully seeding the initialization step, the cost corresponding to the cluster structure returned by the approximation can be bounded within a $\log K$ factor of the minimal cost, where K is the number of clusters. And there are no results showing that the approximation algorithms will converge to the minimal cost. Therefore, their results cannot be directly applied to our problem to provide statistical performance guarantees.

The problem of outlying sequence detection when all the typical distributions are identical was also studied previously as a universal outlier hypothesis testing problem for discrete samples in [12] and for continuous samples in [13], [14]. In [12], the exponential consistency of the generalized likelihood (GL) test under various universal settings was established, where the GL test is based on computing the generalized likelihood function for each hypothesis by taking a maximum likelihood approach with respect to the unknown distributions. However, GL test cannot handle the scenario where the typical distributions are possibly different from each other, and so are the outlying distributions. Another major drawback of the GL test is its high time complexity, which is exponential in M and T (actually M^T or 2^M depending on assumptions).

Our contributions in this paper are summarized as follows. We construct clustering-based tests that are exponentially consistent and have time complexity that is linear in M for various scenarios. For the scenario where GL test is asymptotically optimal, we show that running more steps of the clustering-based test will not decrease the error exponent. We also show that the clustering-based tests are applicable to more general scenarios; for example, when both the typical and outlying distributions form clusters, the clustering-based test is exponentially consistent, but the GL test is not even applicable. We provide numerical results to demonstrate that the clustering-based tests can achieve an error exponent similar to that of the optimal test, but with time complexity that is linear in M. For all scenarios, our experiments indicate that running more steps of the clustering-based test results in a larger error exponent.

The rest of the paper is organized as follows. In Section II, we describe the problem model and three different test scenarios. In Section III, we introduce the GL test studied in [12], which motivates the connection between universal outlying sequence detection and distribution clustering. In Section IV, we reformulate the outlying sequence detection problem as a distribution clustering problem. In Section V, we propose linear-complexity tests based on the K-means clustering algorithm. In Section VI, we provide numerical results. Finally in Section VII, we conclude the paper.

II. Problem Model

Throughout the paper, all random variables are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All distributions are defined on the finite set $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ denotes the set of all probability mass functions on $\mathcal{Y}$.

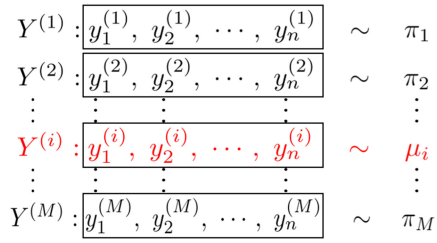


Fig. 1. Outlying sequence detection with data sequences generated by typical distributions denoted by π and outlying distributions denoted by μ.

Consider an outlying sequence detection problem (see Figure 1), where there are in total $M \geq 3$ data sequences denoted by $Y^{(i)}$ for $i = 1, \ldots, M$. Each data sequence $Y^{(i)}$ consists of n i.i.d. samples $Y_1^{(i)}, \ldots, Y_n^{(i)}$. The majority of the sequences are distributed according to typical distributions except for a subset S of outlying sequences, where $S \subset \{1, \ldots, M\}$ and $1 \leq |S| < \frac{M}{2}$. Each typical sequence j is distributed according to a typical distribution $\pi_j \in \mathcal{P}(\mathcal{Y})$, for $j \in S^C$. Each outlying sequence i is distributed according to an outlying distribution $\mu_i \in \mathcal{P}(\mathcal{Y})$, for $i \in S$. Nothing is known about the pmfs of μ_i and π_j except that $\forall i \in S$, $\forall j \in S^C$, $\forall S \subset \{1, \ldots, M\}$, $\mu_i \neq \pi_j$, and all of them have full support over $\mathcal{Y}$. Denote $\mathcal{S}$ as the set comprising all possible outlying subsets.

We use the notation $\boldsymbol{y}^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)})$, where $y_k^{(i)} \in \mathcal{Y}$ is the k-th observation of the i-th sequence. Let γ_i denote the empirical distribution of $\boldsymbol{y}^{(i)}$, and is defined as $\gamma_i(y) \triangleq \frac{1}{n}|\{k = 1, \ldots, n : y_k = y\}|$, for each $y \in \mathcal{Y}$.

A. Three Scenarios

In this paper, we focus on the following three scenarios.

1) Known Number of Outlying Sequences: We first study the scenario where all the typical distributions are identical, i.e., $\pi_j = \pi$, $\forall j \in S^C$, and the number of the outlying sequences is known at the outset, in Section III-A and Section V-A.

2) Unknown Number of Identical Outlying Sequences: We next study the scenario where all the typical distributions are identical ($\pi_j = \pi$, $\forall j \in S^C$), all the outlying distributions are also identical ($\mu_i = \mu$, $\forall i \in S$), and the number of the outlying sequences is unknown, in Section III-B and Section V-B. We note that without any further assumptions, when the number of the outlying sequences is unknown and the outlying sequences can be distinctly distributed, there does not exist a universally exponentially consistent test [12].

3) Typical and Outlying Distributions Forming Clusters: We then study a more general scenario where both the outlying distributions $\{\mu_i\}_{i \in S}$ and the typical distributions $\{\pi_j\}_{j \in S^C}$ form clusters in Section V-C. Moreover, the typical distributions and the outlying distributions are distinct. More concretely,

$$\max_{i,j \in S} D(\mu_i \| \mu_j) < \min_{i \in S, j \in S^C} \{D(\mu_i \| \pi_j), D(\pi_j \| \mu_i)\},$$
$$\max_{i,j \in S^C} D(\pi_i \| \pi_j) < \min_{i \in S, j \in S^C} \{D(\mu_i \| \pi_j), D(\pi_j \| \mu_i)\}. \quad (1)$$

This condition means that the divergence between any two distributions within the same cluster is less than the divergence between any two distributions from two different clusters.

B. Error Exponent

Our goal is to build distribution-free tests to detect the outlying sequences. The test can be captured by a universal rule $\delta : \mathcal{Y}^{Mn} \to \mathcal{S}$, which must not depend on $\{\mu_i\}_{i\in S}$ and $\{\pi_j\}_{j\in S^C}$. We use $\mathbb{P}_S(\cdot)$ to denote the probability conditioned on the hypothesis that corresponds to the set of outlying sequences being $S \in \mathcal{S}$. Thus, the joint distribution of all the observations can be written as

$$\mathbb{P}_S(y^{Mn}) = L_S\left(y^{Mn}, \{\mu_i\}_{i\in S}, \{\pi_j\}_{j\in S^C}\right) = \prod_{k=1}^{n}\left\{\prod_{i\in S}\mu_i(y_k^{(i)})\prod_{j\in S^C}\pi_j(y_k^{(j)})\right\}, \quad (2)$$

where $L_S\left(y^{Mn}, \{\mu_i\}_{i\in S}, \{\pi_j\}_{j\in S^C}\right)$ denotes the likelihood.

The performance of a universal test is gauged by the maximal probability of error, which is defined as

$$e\big(\delta\big) \triangleq \max_{S\in\mathcal{S}} \sum_{y^{Mn}:\, \delta(y^{Mn})\neq S} \mathbb{P}_S(y^{Mn}),$$

and the corresponding error exponent is defined as

$$\alpha\big(\delta\big) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log e\big(\delta\big).$$

Definition 1: A universal test δ is said to be universally exponentially consistent if $\alpha\big(\delta\big) > 0$.

We use $D(\pi\|\mu)$ and $B(\pi,\mu)$ to denote the KL divergence and Bhattacharyya distance between distributions π and μ:

$$D(\pi\|\mu) \triangleq \sum_{y\in\mathcal{Y}}\pi(y)\log\left(\frac{\pi(y)}{\mu(y)}\right),$$

$$B(\pi,\mu) \triangleq -\log\left(\sum_{y\in\mathcal{Y}}\pi(y)^{\frac{1}{2}}\mu(y)^{\frac{1}{2}}\right).$$

III. Generalized Likelihood Test

In this section, we introduce the GL test for outlying sequence detection studied in [12], and summarize its consistency results (see [12] for details). Here, it is assumed that the typical distributions are identical, i.e., $\pi_j = \pi,\ \forall j \in S^C$.

In the universal setting with π and $\{\mu_i\}_{i\in S}$ unknown, conditioned on the outlying set being $S \in \mathcal{S}$, we compute the generalized likelihood of y^{Mn} by replacing π and $\{\mu_i\}_{i\in S}$ in (2) with their maximum likelihood estimates (MLEs) $\{\hat{\mu}_i\}_{i\in S}$, and $\hat{\pi}_S$, as

$$\hat{\mathbb{P}}_S^{\text{univ}}(y^{Mn}) = \hat{L}_S(y^{Mn}, \{\hat{\mu}_i\}_{i\in S}, \hat{\pi}_S). \quad (3)$$

The GL test [12] then selects the hypothesis under which the GL is maximized (ties are broken arbitrarily), i.e.,

$$\delta_{\text{GL}}(y^{Mn}) = \arg\max_{S\in\mathcal{S}} \hat{\mathbb{P}}_S^{\text{univ}}. \quad (4)$$

In the following subsections, we consider different scenarios, where the suitable set $\mathcal{S}$ and the MLE of $\{\hat{\mu}_i\}_{i\in S}$, and $\hat{\pi}_S$ may differ.

A. Known Number of Outlying Sequences

We first consider the scenario in which the number of outlying sequences, denoted by $T \geq 1$, is known at the outset, i.e., $\mathcal{S} = \{S : S \subset \{1,\ldots,M\}, |S| = T\}$. Moreover, the distributions of different outlying sequences $\mu_i, i \in S$, can be distinct from each other.

We compute the generalized likelihood of y^{Mn} by replacing the $\mu_i,\ i \in S$ and π in (2) with their MLEs:

$$\hat{\mu}_i = \gamma_i, \text{ and } \hat{\pi}_S = \frac{\sum_{j\in S^C}\gamma_j}{M-T}.$$

Then, as in [12], the GL test in (4) is equivalent to

$$\delta_{\text{GL}}(y^{Mn}) = \arg\min_{S\subset\mathcal{S}} \sum_{j\in S^C} D\left(\gamma_j \middle\| \frac{\sum_{j\in S^C}\gamma_j}{M-T}\right). \quad (5)$$

Proposition 1 ([12, Theorem 10]): Consider the scenario when the number of outlying sequences is known. If the typical distributions are identical, then the GL test in (5) is universally exponentially consistent. As $M \to \infty$, the achievable error exponent converges as

$$\lim_{M\to\infty}\alpha\big(\delta_{\text{GL}}\big) = \lim_{M\to\infty}\min_{i=1,\ldots,M} 2B(\mu_i,\pi).$$

When all the outlying sequences are identically distributed, i.e., $\mu_i = \mu \neq \pi, i = 1,\ldots,M$, the achievable error exponent of the GL test in (5) converges to the optimal one achievable when both μ and π are known, i.e., $2B(\mu,\pi)$.

Note that the number of hypotheses in the test (5) is $\binom{M}{T}$. An exhaustive search over all possible hypotheses has time complexity that is polynomial in M and exponential in T.

B. Unknown Number of Identical Outlying Sequences

In this subsection, we consider the scenario where the number of outlying sequences is unknown, i.e., $\mathcal{S} = \{S : S \subset \{1,\ldots,M\}, 1 \leq |S| < M/2\}$, and the hypotheses in $\mathcal{S}$ may have different numbers of outlying sequences. Moreover, it is assumed that the typical distributions are identical, and the outlying distributions are identical.

As shown in [12], by replacing the $\mu_i,\ i \in S$, and π in (2) with their MLEs:

$$\hat{\mu}_S = \hat{\mu}_i = \frac{\sum_{i\in S}\gamma_i}{|S|}, \text{ and } \hat{\pi}_S = \frac{\sum_{j\in S^C}\gamma_j}{M-|S|},$$

the GL test in (4) is equivalent to

$$\delta_{\text{GL}}(y^{Mn}) = \arg\min_{S\subset\mathcal{S}} \sum_{j\in S^C} D\left(\gamma_j \middle\| \frac{\sum_{j\in S^C}\gamma_j}{M-|S|}\right) + \sum_{i\in S} D\left(\gamma_i \middle\| \frac{\sum_{i\in S}\gamma_i}{|S|}\right). \quad (6)$$

Proposition 2 ([12, Theorem 11]): Consider the scenario when the number of the outlying sequences is unknown,

$1 \leq |S| < \frac{M}{2}$. If all the outlying sequences are identically distributed, and all the typical sequences are identically distributed, the GL test in (6) is universally exponentially consistent.

Note that the number of hypotheses in the GL test (6) is $\sum_{i=1}^{\lfloor M/2 \rfloor} \binom{M}{i}$, which is exponential in M. The time complexity of (6) is even larger than that of (5), without the knowledge of the number of outlying sequences.

Although the exponential consistency of the GL test under various universal settings was established in [12], the high time complexity, which is at least exponential in the number of outlying sequences T, limits its usage in applications.

In the following two sections, we reformulate the universal outlying sequence detection problem as a distribution clustering problem, and further propose clustering-based algorithms that are computationally efficient and exponentially consistent. In particular, we reduce the time complexity of our tests to $O(M)$, while still retaining a comparable error probability.

IV. Problem Reformulation as Distribution Clustering

The GL test can be interpreted as combinatorial clustering over the probability simplex with the KL divergence as the distance measure. More specifically, consider the problem of clustering distributions $p_1, p_2, \ldots$ into K clusters, using a set of cluster centers $c = \{c^{(1)}, \ldots, c^{(K)}\}$, and a cluster assignment $C = \{C^{(1)}, \ldots, C^{(K)}\}$. If we define the following cost function for distribution clustering

$$TC \triangleq \sum_{k=1}^{K} \sum_{i \in C^{(k)}} D(p_i \| c^{(k)}). \tag{7}$$

As shown in [6, Proposition 1], for a given cluster assignment $C = \{C^{(1)}, \ldots, C^{(K)}\}$, the cost is minimized when

$$c^{(k)} = \frac{\sum_{i \in C^{(k)}} p_i}{|C^{(k)}|},$$

which is the average of the distributions within the k-th cluster. Thus, for a given cluster assignment $C = \{C^{(1)}, \ldots, C^{(K)}\}$, we have

$$\min_{c^{(1)}, \ldots, c^{(K)}} TC = \sum_{k=1}^{K} \sum_{i \in C^{(k)}} D\left(p_i \middle\| \frac{\sum_{i \in C^{(k)}} p_i}{|C^{(k)}|}\right). \tag{8}$$

To connect the distribution clustering problem with the GL test, we first consider the scenario in Subsection III-B, in which the typical distributions are identical and the outlying distributions are also identical. In view of (8), the GL test in (6) can be interpreted as a distribution clustering algorithm for the empirical distributions γ_i, $1 \leq i \leq M$, with $K = 2$ clusters. The first term in (6) is the minimum cost in the typical cluster, and the second term is the minimum cost within the outlying cluster, for a given choice of S. The GL test then searches over all possible cluster assignments, and chooses the one with minimum cost.

We then consider the scenario in Subsection III-A, where the typical distributions are identically distributed, but outliers are not (i.e., outlying distributions may not form a cluster). We can utilize the knowledge of the number of outlying sequences, and it suffices to only cluster the empirical distributions of all the typical sequences, as shown in the GL test (5).

Algorithm 1: K-means Distribution Clustering Algorithm.

Input: M distributions $p_1, \ldots, p_M$, defined on $\mathcal{Y}$, number of clusters K.
Output: partition set $\{C^{(k)}\}_{k=1}^{K}$.
Initialization: $\{c^{(k)}\}_{k=1}^{K}$ (Will be specified in Algorithms 2 and 3.)
Method:
while not converge **do**
  {Assignment Step}
  Set $C^{(k)} \leftarrow \emptyset$, $1 \leq k \leq K$
  for $i = 1$ to M **do**
    $C^{(k^*)} \leftarrow C^{(k^*)} \cup \{p_i\}$
    where $k^* = \arg\min_k D(p_i \| c^{(k)})$
  end for
  {Re-estimation Step}
  for $k = 1$ to K **do**
    $c^{(k)} \leftarrow \frac{\sum_{i \in C^{(k)}} p_i}{|C^{(k)}|}$
  end for
end while
Return $\{C^{(k)}\}_{k=1}^{K}$

Thus, both the GL tests in (5) and (6) are equivalent to empirical distribution clustering on the probability simplex using KL divergence as the distance metric.

While the distribution clustering problem itself is known to be NP-hard [8], there are many existing approximation algorithms with low time complexity, e.g., the K-means algorithm [15]. Here, we introduce the K-means distribution clustering algorithm in Algorithm 1, as proposed in [6].

Proposition 3 ([6, Proposition 3]): The cost function in (8) of Algorithm 1 is monotonically decreasing with steps. Moreover, Algorithm 1 terminates in a finite number of steps at a partition that is locally optimal, i.e., the total cost cannot be decreased by either (a) the assignment step, or (b) changing the means of any existing clusters.

Remark 1: The proof of Proposition 3 in [6] follows by the fact that the number of distinct cluster assignments is finite, and the fact that Algorithm 1 monotonically decreases the cost function in (8). As shown in [16], for any Bregman divergence, the number of iterations of the K-means algorithm in the worst case can be upper bounded by $O(M^{K^2|\mathcal{Y}|})$, where K is the number of clusters.

For our problem, the number of clusters is 2. Then, Algorithm 1 has polynomial time complexity in M even in the worst case. In the following section, we will show that the exponential consistency can be established with a well-designed initialization in the first step, i.e., we do not need to wait for the algorithm to converge. As also will be shown in Section VI, our tests usually converge in a few steps.

V. Clustering-Based Tests

In this section, we propose linear-complexity tests based on the K-means clustering algorithm. We show in all three scenarios that the clustering-based tests using KL divergence as the distance metric are also exponentially consistent, while

Algorithm 2: Clustering with known Number of Outlying Sequences.

Input: $\gamma_1, \ldots, \gamma_M$, number of the outlying sequences T.
Output: A set of outlying sequences S.
Initialization:
$\gamma^{(0)}$: Choose one distribution from $\gamma_1, \ldots, \gamma_M$ uniformly and randomly
for $i = 1$ to M **do**
  Compute $D(\gamma_i \| \gamma^{(0)})$
end for
$\hat{\pi} \leftarrow \gamma^*$
where $D(\gamma^* \| \gamma^{(0)})$ is the $\lceil \frac{M}{2} \rceil$-th element among $D(\gamma_i \| \gamma^{(0)})$, $1 \le i \le M$ in an ascending order
Method:
While not converge **do**
  {Assignment Step}
  Set $S \leftarrow S^*$,
  where $S^* = \arg\max_{S' \in \mathcal{S}, |S'|=T} \sum_{i \in S'} D(\gamma_i \| \hat{\pi})$
  {Re-estimation Step}
  $\hat{\pi} \leftarrow \frac{\sum_{j \in S^C} \gamma_j}{M-T}$
end while
Return S

only taking linear time in M. For the scenario that the typical and outlying distributions form two clusters, we show that the clustering-based test is exponentially consistent, but the GL test is not even applicable.

A. Known Number of Outlying Sequences

We first consider the scenario where the number of outlying sequences T is known and the typical distributions are identical.

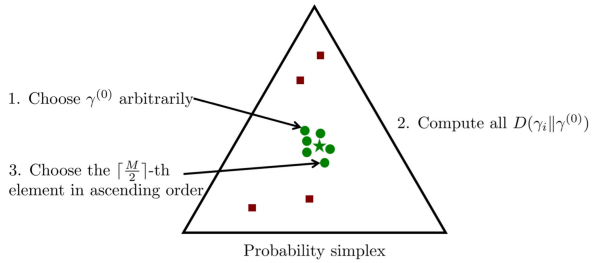
Note that Algorithm 1 cannot be directly applied here, because the outlying distributions may not form a cluster and Algorithm 1 does not employ the knowledge of T.

Motivated by the test in (5), we design Algorithm 2. The novelty of this algorithm lies in the construction of the first cluster center for the typical distribution and the iterative approach based on K-means to update it.

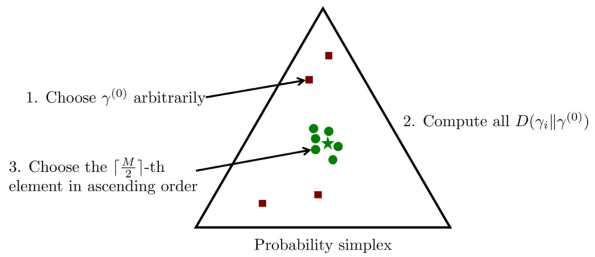
By the initialization in Algorithm 2, γ^* is generated from π with high probability. The intuition behind this is that: if $\gamma^{(0)}$ is generated from the typical distribution π as shown in Figure 2(a), then only $|S| < \frac{M}{2}$ empirical distributions which are generated from μ_i are far from $\gamma^{(0)}$; if $\gamma^{(0)}$ is generated from some μ_i as shown in Figure 2(b), then there are at least $M - |S| > \frac{M}{2}$ of $D(\gamma_i \| \gamma^{(0)})$ concentrating at $D(\pi \| \mu_i)$. Thus, the $\lceil \frac{M}{2} \rceil$-th element among all $D(\gamma_i \| \gamma^{(0)})$, $1 \le i \le M$, arranged in an ascending order, is close to $D(\pi \| \mu_i)$, and γ^* is generated from π with high probability.

Let δ_2 denote the test described in Algorithm 2, and $\delta_2^{(\ell)}$ denote the test that runs ℓ number of K-means iterations in Algorithm 2.

In the following theorem, we show that the test $\delta_2^{(1)}$ (only one iteration step) is universally exponentially consistent.



(a) $\gamma^{(0)}$ is generated from the typical distribution



(b) $\gamma^{(0)}$ is generated from the outlying distribution

Fig. 2. Diagrams of Algorithm 2, where $M = 10$, $|S| = 4$. The star denotes the (identical) typical distribution, and the circles and squares denote the empirical distributions of the typical and outlying sequences, respectively.

Theorem 1: Consider the scenario when the number of outlying sequences T is known. If the typical distributions are identical, the test $\delta_2^{(1)}$, which runs one K-means iteration in Algorithm 2 is universally exponentially consistent. The achievable error exponent of $\delta_2^{(1)}$ can be upper bounded by

$$\alpha\big(\delta_2^{(1)}\big) \le \lim_{M \to \infty} \min_{i=1,\ldots,M} 2B(\mu_i, \pi). \tag{9}$$

Furthermore, the time complexity of the test $\delta_2^{(1)}$ is $O(M)$.

Proof sketch: Errors made by $\delta_2^{(1)}$ in the initialization step can be decomposed into two scenarios. If $\gamma^{(0)}$ is generated from typical distribution π, an error occurs when γ^* is actually generated from an outlying distribution. The probability of this event can be upper bounded by the probability of the following event

$$E_1 = \{\exists\, i \in S,\ \exists\, j_1, j_2 \in S^C,\ D(\gamma_i \| \gamma_{j_1}) < D(\gamma_{j_2} \| \gamma_{j_1})\}.$$

If $\gamma^{(0)}$ is generated from an outlying distribution, the error probability can be upper bounded by the probability of the following event

$$E_2 = \{\exists\, i_1,\ i_2 \in S,\ \exists\, j_1, j_2 \in S^C,\ D(\gamma_{j_1} \| \gamma_{i_1}) < D(\gamma_{i_2} \| \gamma_{i_1}) < D(\gamma_{j_2} \| \gamma_{i_1})\}.$$

By Sanov's theorem [17], we can prove that the probabilities of both E_1 and E_2 decay exponentially fast.

The error probability in the assignment step can be upper bounded by the probability of the same event E_1, which decays exponentially fast by Sanov's theorem.

Both the initialization and the assignment steps in Algorithm 2 that select the $\lceil \frac{M}{2} \rceil$-th element and the T largest elements can be computed in linear time $O(M)$ using the Quickselect algorithm in [18].

The details of the proof can be found in Appendix B. ■

A comparison between Proposition 1 and Theorem 1 shows that $\delta_2^{(1)}$ has a smaller error exponent than that of the GL test in (5) as $M \to \infty$, but has a linear time complexity in M.

Although the exponential consistency can be established for the one step test $\delta_2^{(1)}$, it is of further interest to investigate whether the performance of Algorithm 2 improves with more iterations. In the following theorem, we show that the asymptotic performance of Algorithm 2 does not decrease with more iterations. We will also see in our numerical results in Section VI that running more iterations of K-means always results in better performance.

Theorem 2: For each $M \geq 3$, when the number of outlying sequences T is known, the test $\delta_2^{(\ell)}$ is universally exponentially consistent. As $M \to \infty$, the achievable error exponent of $\delta_2^{(\ell)}$ in Algorithm 2 can be lower bounded by

$$\lim_{M\to\infty} \alpha\big(\delta_2^{(\ell)}\big) \geq \lim_{M\to\infty} \alpha\big(\delta_2^{(1)}\big). \tag{10}$$

Furthermore, the time complexity of the test $\delta_2^{(\ell)}$ is $O(M\ell)$.

Proof: From Theorem 1 and Proposition 1, we know that both the one-step test $\delta_2^{(1)}$ and the GL test δ_{GL} are exponentially consistent. Then we have $\alpha\big(\delta_2^{(1)}\big) > 0$ and $\alpha\big(\delta_{\mathrm{GL}}\big) > 0$. Denote

$$A \triangleq \big\{y^{Mn} : \delta_2^{(1)}(y^{Mn}) \neq S\big\},$$
$$B \triangleq \big\{y^{Mn} : \delta_{\mathrm{GL}}(y^{Mn}) \neq S\big\}.$$

Then the set $A \cup B$ contains those y^{Mn} such that at least one of $\delta_2^{(1)}$ and δ_{GL} makes an error. Thus,

$$\lim_{n\to\infty} -\frac{1}{n} \log\Big(\mathbb{P}_S\big(A \cup B\big)\Big) = \min\big\{\alpha\big(\delta_2^{(1)}\big), \alpha\big(\delta_{\mathrm{GL}}\big)\big\}.$$

This shows that the probability that at least one of $\delta_2^{(1)}$ and δ_{GL} makes an error decays exponentially fast. Thus, the one-step test $\delta_2^{(1)}$ and the GL test δ_{GL} output the same correct S with high probability.

If $\delta_2^{(1)}$ and δ_{GL} achieve the same outcome, which means both $\delta_2^{(1)}$ and δ_{GL} have achieved the global minimum of the cost function (8), then running more iterations in Algorithm 2 will not change the outcome. Thus,

$$e\big(\delta_2^{(\ell)}\big) \leq \mathbb{P}_S\big(A \cup B\big). \tag{11}$$

The error exponent of running ℓ iterations will be lower bounded by

$$\begin{aligned}\alpha\big(\delta_2^{(\ell)}\big) &= \lim_{n\to\infty} -\frac{1}{n} \log e\big(\delta_2^{(\ell)}\big) \\ &\geq \lim_{n\to\infty} -\frac{1}{n} \log\Big(\mathbb{P}_S\big(A \cup B\big)\Big) \\ &= \min\big\{\alpha\big(\delta_2^{(1)}\big), \alpha\big(\delta_{\mathrm{GL}}\big)\big\}. \end{aligned} \tag{12}$$

As $M \to \infty$, we know that $\alpha(\delta_2^{(1)}) < \alpha(\delta_{\mathrm{GL}})$ from Theorem 1, then we can conclude that

$$\lim_{M\to\infty} \alpha(\delta_2^{(\ell)}) \geq \lim_{M\to\infty} \min(\alpha(\delta_{\mathrm{GL}}), \alpha(\delta_2^{(1)})) \geq \lim_{M\to\infty} \alpha(\delta_2^{(1)}).$$

Algorithm 3: Clustering with Unknown Number of Outlying Sequences.

Input: M empirical distributions $\gamma_1, \ldots, \gamma_M$ defined on finite alphabet $\mathcal{Y}$.
Output: A set of outlying sequences S.
Initialization:
Choose one distribution $\gamma^{(0)}$ arbitrarily,
$c^{(1)} \leftarrow \arg\max_{\gamma_i} D(\gamma_i \| \gamma^{(0)})$
$c^{(2)} \leftarrow \gamma^{(0)}$
Method: Same as in Algorithm 1 with $K = 2$
Return the smaller of the sets $C^{(1)}$ and $C^{(2)}$

As for the time complexity, since each iteration has time complexity $O(M)$, $\delta_2^{(\ell)}$ which runs ℓ iterations has time complexity $O(M\ell)$. ■

Remark 2: There are other works applying different distance metrics in clustering algorithm, for example, using Rényi divergence [19] or maximum mean discrepancy (MMD) [14]. However, the choice of KL divergence as the distance metric is crucial in our theoretical analysis, and the reason is two-fold:

1) For a given cluster, the cost function with KL divergence is minimized when the center is the average of all distributions in this cluster [6]. This property holds for all Bregman divergences and ensures that the clustering algorithm will terminate in finite steps. We have noticed that using MMD (not a Bregman divergence) will result in oscillating between two different clustering partitions.
2) As shown in [12], when there are a known number of identically distributed outlying sequences, the GL test was shown to be asymptotically optimal as M goes to infinity. It is based on this asymptotic optimality of using KL divergence that we prove that the error exponent of our clustering-based test is not decreasing as running more iterations.

B. Unknown Number of Identical Outlying Sequences

In this section, we consider the scenario where the number of outlying sequences is unknown. Moreover, the typical distributions are identical, and the outlying distributions are identical.

Since there is no prior information on the number of outlying sequences, we apply Algorithm 1 directly. Motivated by the test in (6), we design the following initialization in Algorithm 3 to set the cluster centers in Algorithm 1.

With high probability, $c^{(1)}$ and $c^{(2)}$ chosen by the initialization step in Algorithm 3 are generated by different distributions.

Let δ_3 denote the test described in Algorithm 3, and $\delta_3^{(\ell)}$ denote the test that runs ℓ iterations in Algorithm 3. The following theorem shows that the clustering-based test $\delta_3^{(\ell)}$, is universally exponentially consistent, and has time complexity that is linear in M.

Theorem 3: Consider the scenario when the number of the outlying sequences is unknown, and $1 \leq |S| < \frac{M}{2}$. If all the outlying sequences are identically distributed, and all the typical sequences are identically distributed, the test $\delta_3^{(\ell)}$, which runs ℓ

steps of Algorithm 3, is exponentially consistent, and has time complexity $O(M\ell)$.

Proof sketch: The exponential consistency of $\delta_3^{(\ell)}$ can be established using similar techniques to those in Theorem 1 and Theorem 2. The major difference between the proof of Theorem 1 and Theorem 3 is that there are two cluster centers in the initialization step and assignment step in Algorithm 3. The details can be found in Appendix C. ■

C. Typical and Outlying Distributions Forming Clusters

In this subsection, we consider the scenario that both the outlying distributions $\{\mu_i\}_{i\in S}$ and the typical distributions $\{\pi_j\}_{j\in S^c}$ are not identically distributed. Moreover, the typical distributions and the outlying distributions form clusters as defined in (1), which means that the divergence within the same cluster is always less than the divergence between different clusters.

The following theorem shows that under the condition (1), the one step test $\delta_3^{(1)}$ proposed in Algorithm 3 is universally exponentially consistent, and has time complexity that is linear in M.

Theorem 4: For each $M \geq 3$, when both the outlying distributions $\{\mu_i\}_{i\in S}$ and the typical distributions $\{\pi_j\}_{j\in S^c}$ form clusters, i.e. condition (1) holds, the test $\delta_3^{(1)}$, which runs one step of Algorithm 3, is universally exponentially consistent, and has time complexity $O(M)$.

Proof sketch: The exponential consistency of $\delta_3^{(1)}$ can be established using techniques similar to those in Theorem 3. The details can be found in Appendix D. ■

The GL approach of replacing the true distribution in (2) by their MLEs leads to identical likelihood estimates under each hypothesis. Thus, the GL approach is not applicable here. One could apply the test in (6) to this problem, but the following example shows that the test in (6) is not universally exponentially consistent, even if condition (1) holds.

Example 1: As shown in [12], the error exponent of the GL test in (6) is established by showing the following optimization problem has a positive value

$$\min_{q_1,q_2,\ldots,q_M\in C_{(S,S')}} \sum_{i\in S} D(q_i\|\mu_i) + \sum_{j\in S^C} D(q_j\|\pi_j), \qquad (13)$$

where

$$C_{(S,S')} = \Bigg\{(q_1,\ldots,q_M): \sum_{i\in S} D\left(q_i \Bigg\| \frac{\sum_{k\in S} q_k}{|S|}\right) + \sum_{j\notin S} D\left(q_j \Bigg\| \frac{\sum_{k\notin S} q_k}{M-|S|}\right) \geq \sum_{i\in S'} D\left(q_i \Bigg\| \frac{\sum_{k\in S'} q_k}{|S'|}\right) + \sum_{j\notin S'} D\left(q_j \Bigg\| \frac{\sum_{k\notin S'} q_k}{M-|S'|}\right)\Bigg\}.$$

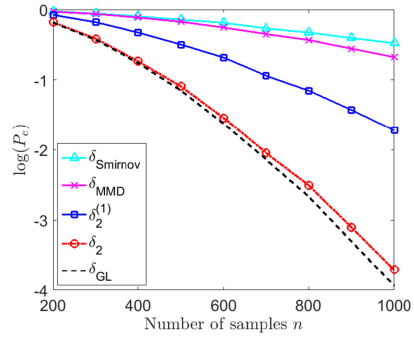


Fig. 3. Comparison of tests $\delta_2^{(1)}$, δ_2, δ_{GL}, MMD-based test and FR-Smirnov test with known number of distinct outlying distributions.

We consider the scenario where $M = 1000$, $S = \{1, 2\}$, the typical and outlying distributions are specified as follows:

$$\mu_1 = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right), \mu_2 = \left(\frac{1}{5}, \frac{7}{15}, \frac{1}{3}\right),$$
$$\pi_3 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \pi_4 = \cdots = \pi_{1000} = \left(\frac{247}{500}, \frac{32}{125}, \frac{1}{4}\right).$$

It can be verified the clustering condition (1) holds for this example. However, if we let $q_1 = \mu_1$, $q_2 = \mu_2$, $q_3 = \pi_3$, $q_4 = \cdots = q_{1000} = \pi_4$, $S = \{1, 2\}$ and $S' = \{1, 2, 3\}$, then

$$\sum_{i\in S} D\left(q_i \Big\| \frac{\sum_{k\in S} q_k}{|S|}\right) + \sum_{j\notin S} D\left(q_j \Big\| \frac{\sum_{k\notin S} q_k}{M-|S|}\right) \geq \sum_{i\in S'} D\left(q_i \Big\| \frac{\sum_{k\in S'} q_k}{|S'|}\right) + \sum_{j\notin S'} D\left(q_j \Big\| \frac{\sum_{k\notin S'} q_k}{M-|S'|}\right) \qquad (14)$$

also holds, i.e., $(q_1, q_2, \ldots, q_M) \in C_{S,S'}$, which means the error exponent in (13) is equal to zero. Thus, the test in (6) is not universally exponentially consistent for the scenario where both typical and outlying distributions form clusters.

VI. Numerical Results

In this section, we compare the performance of the proposed clustering-based tests δ_2, δ_3 (run until convergence) and the one step tests $\delta_2^{(1)}$, $\delta_3^{(1)}$ with the GL test δ_{GL}, and other baseline tests including the MMD-based test [14] and the FR-Smirnov test [20].

For the scenario with identical typical distribution, we set π to be the uniform distribution with alphabet size 10, and generate outlying distributions randomly.

We first simulate the scenario where the outlying distributions are distinct and T is known. We choose $M = 20$, $T = 3$. In Figure 3, we plot the logarithm of the error probability as a function of n for δ_{GL}, δ_2, $\delta_2^{(1)}$, the MMD-based test and the FR-Smirnov test, averaged over 5000 Monte Carlo simulation runs. As we can see from Figure 3, all the compared tests are exponentially consistent, and the clustering-based tests outperform the MMD-based test and the FR-Smirnov test. Moreover, δ_2 outperforms $\delta_2^{(1)}$ significantly, as suggested by Theorem 2. A comparison of δ_2 and δ_{GL} shows that they are close in

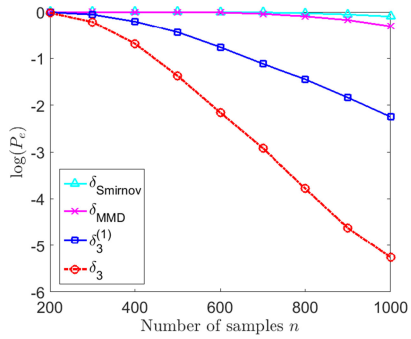


Fig. 4. Comparison of tests δ_3, $\delta_3^{(1)}$, MMD-based test and FR-Smirnov test with unknown number of identical outlying distributions.

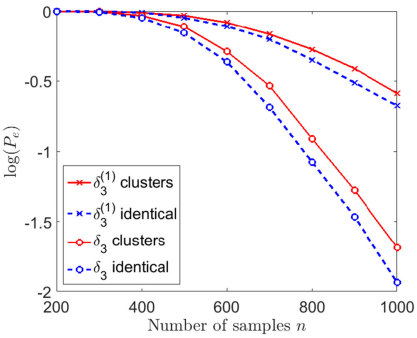


Fig. 5. Comparison of tests δ_3, $\delta_3^{(1)}$ when typical distributions and outlying distributions form clusters, respectively.

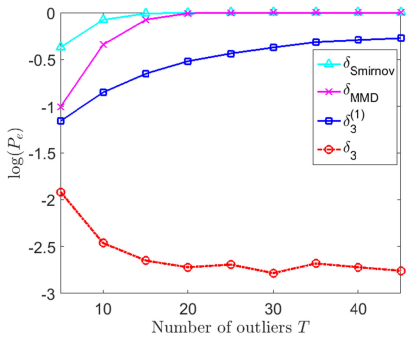


Fig. 6. Comparison of tests δ_3, $\delta_3^{(1)}$, MMD-based test and FR-Smirnov test with different unknown number of identical outlying sequences.

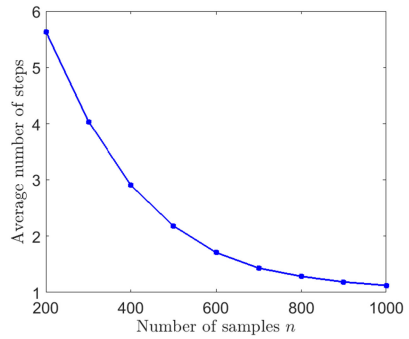


Fig. 7. Average number of steps for convergence of test δ_3 versus number of samples n.

performance, but δ_2 is about 50 times faster than δ_{GL} in terms of computation time.

We then simulate the scenario with unknown number of identical outlying distributions. We set $M = 100$, $T = 10$. Figure 4 shows that δ_3 outperforms $\delta_3^{(1)}$, the MMD-based test, and the FR-Smirnov test. Note that the MMD-based test and the FR-Smirnov test cannot deal with the scenario where T is unknown. Here these tests are implemented with the knowledge of T. We note that the GL test is not computationally feasible here, since the number of hypotheses one needs to search over is exponential in M.

For the scenario where both the typical and outlying distributions form clusters, we set the alphabet size to be 10. And we choose the uniform distribution as the center of the typical cluster. We generate the typical distributions by adding some Gaussian noise to the cluster center, and then normalizing them. The cluster of the outlying distributions are generated in the same way, but with a randomly chosen cluster center. We set $M = 100$, $T = 10$. The dotted lines in Figure 5 correspond to the scenario where the typical distributions are identical and the outlying distributions are identical, and equal to the corresponding cluster center. The solid lines correspond to the scenario where both the typical and outlying distributions are generated by the approach mentioned above, which form clusters. Figure 5 shows that the tests $\delta_3^{(1)}$ and δ_3 are exponentially consistent, and that δ_3 outperforms $\delta_3^{(1)}$ for both scenarios.

Figures 3, 4 and 5 demonstrate the exponential consistency of the proposed test by plotting the logarithm of the error probability as a function of n. To illustrate how the number of outlying sequences influences the test performance, we plot the logarithm of the error probability as a function of T in Figure 6. We use the same setting as in Figure 4, with the knowledge of T being used when implementing the MMD-based and FR-Smirnov tests. We set $M = 100$, $n = 400$, and let T range from 5 to 45. Figure 6 shows that δ_3 outperforms all the other tests. In addition, the error probability of δ_3 decreases as T increases, while the error probabilities of the other compared tests increase. The improved performance of δ_3 is due to the fact that the K-means clustering algorithm performs better when the clusters become more balanced, which happens as T increases.

We further study the number of iterations that δ_3 takes to converge. Figure 7 plots the average number of steps of test δ_3 versus the number of samples, using the same setting as in Figure 4. It is seen that the more the samples collected, the fewer the iterations needed. Moreover, when the number of samples n goes to infinity, δ_3 converges in just one step, which explains the exponential consistency of the test $\delta_3^{(1)}$.

Moreover, we compare how the time complexity of the clustering-based test δ_3, $\delta_3^{(1)}$, the MMD-based test, and the FR-Smirnov test varies as a function of the number of sequences M. Note that as seen in Theorems 1, 2 and 3, the time complexity of these tests is independent of the number of outlying sequences T. Thus, we simulate the scenario where the number of outlying sequences T changes with the number of sequences

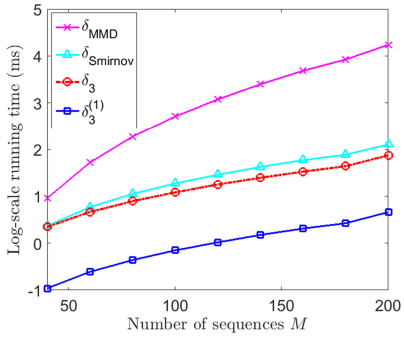


Fig. 8. Comparison of the log-scale average running time versus number of sequences M.

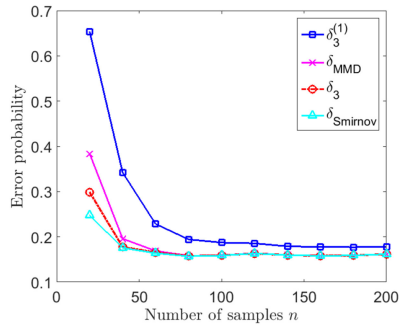


Fig. 9. Comparison of tests δ_3, $\delta_3^{(1)}$ with other tests on the climate type dataset.

M, and plot the log-scale average running time in Figure 8. The experiments are simulated using a 3.6 GHz i7 CPU. We use the same setting as in Figure 4. The knowledge of T is used when implementing the MMD-based test and the FR-Smirnov test. Here, we set $T = M/5$, where M ranges from 40 to 200, and $n = 400$. Figure 8 shows that δ_3, $\delta_3^{(1)}$ and the FR-Smirnov test have time complexity that is linear in M. However, the time complexity of the MMD-based test is $O(M^2)$ [14].

The constant gap between the curve of δ_3 and $\delta_3^{(1)}$ in Figure 8 shows that the number of iterations that δ_3 takes to converge does not change much as M increases. Thus, both Figure 7 and Figure 8 show that in general our clustering-based test δ_3 converges in very few steps.

Finally, we compare the performance of these tests on a real climate type dataset. The climate data taken by different weather stations are obtained from the National Center for Atmospheric Research data archive [21]. We label the climate type of each station using the Koppen-Geiger climate classification method [22]. The sequences are constructed by quantizing the precipitation records of each month at different weather stations into 20 different levels, i.e., the precipitation records across months form a sequence consisting of discrete observations for each station. We randomly choose 16 stations in southeast China and southeast North America (191 stations in total) to construct the typical sequences, and randomly choose $T = 4$ stations in north Africa and central Australia (13 stations in total) to construct the outlying sequences. We randomly choose n months from 1987 until 2012, and let n vary. We plot the probability of error for δ_3, $\delta_3^{(1)}$, the MMD-based test and the FR-Smirnov test as a function of n in Figure 9. Again, T is known in the implementation of the MMD-based test and the FR-Smirnov test. It can be seen that the clustering-based test δ_{c3} achieves similar performance as the MMD-based test and the FR-Smirnov test, despite lacking the knowledge of T.

In many practical applications, e.g., natural language processing, the alphabet size of the underlying distributions can be very large. To estimate the KL divergence efficiently, we can utilize the minimax optimal KL divergence estimator proposed in [23] and random projection technique [24] to reduce the computational complexity of K-means clustering algorithm, which might be a path for future work.

VII. Conclusion

In this paper, we have investigated the universal outlying sequence detection problem. We have constructed clustering-based tests that are exponentially consistent and have time complexity that is linear in M for various scenarios. For the scenario where GL test is asymptotically optimal, we have shown that running more steps of the clustering-based test does not decrease the error exponent. We have further shown that the clustering-based tests are applicable to more general scenarios. For example, when both the typical and outlying distributions form clusters, the clustering-based test is exponentially consistent, but the GL test is not even applicable. We have provided numerical results to demonstrate that our clustering-based test can achieve a similar error exponent as the GL test.

Our study provides a new way to quantify the performance of distribution clustering algorithms, via the lens of exponential consistency. We believe that this approach can be applied to universal outlying sequence detection with continuous distributions and also to other nonparametric problems.

Appendix A
Useful Lemmas

Lemma 1 ([12, Lemma 1]): Let $Y^{(1)}, \ldots, Y^{(J)}$ be mutually independent random vectors with each $Y^{(j)}$, $j = 1, \ldots, J$, being n i.i.d. samples of a random variable distributed according to $p_j \in \mathcal{P}(\mathcal{Y})$. Let A_n be the set of all J tuples $(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(J)}) \in \mathcal{Y}^{Jn}$ whose empirical distributions $(\gamma_1, \ldots, \gamma_J)$ lie in a closed set $E \in \mathcal{P}(\mathcal{Y})^J$. Then,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left\{\left(Y^{(1)}, \ldots, Y^{(J)}\right) \in A_n\right\} = \min_{(q_1,\ldots,q_J)\in E} \sum_{j=1}^{J} D(q_j \| p_j). \tag{15}$$

Lemma 2 ([12, Lemma 2]): For any two pmfs $p_1, p_2 \in \mathcal{P}(\mathcal{Y})$ with full supports, it holds that

$$2B(p_1, p_2) = \min_{q\in\mathcal{P}(\mathcal{Y})} \Big(D(q\|p_1) + D(q\|p_2)\Big). \tag{16}$$

In particular, the minimum on the right side is achieved by

$$q^* = \frac{p_1^{1/2}(y) p_2^{1/2}(y)}{\sum_{y\in\mathcal{Y}} p_1^{1/2}(y) p_2^{1/2}(y)},\ y \in \mathcal{Y}. \tag{17}$$

APPENDIX B
PROOF OF THEOREM 1

Due to the structure of the test we know that errors may occur at the following two steps:

1) *Initialization Step:* The constructed cluster center for typical sequences $\hat{\pi}$ is actually generated from an outlying distribution.
2) *Assignment Step:* Given that the cluster center $\hat{\pi}$ is actually generated from typical distribution π, the empirical distribution of an outlying sequence is closer to $\hat{\pi}$.

We use E to denote the event that errors occur in the initialization step. It is difficult to write the explicit form of the event E. However, we can find upper bounds for the probability of E for the following two scenarios.

If $\gamma^{(0)}$ is generated from the typical distribution π, an error occurs when $\hat{\pi}$ is actually generated from an outlying distribution, then $D(\gamma_i\|\gamma^{(0)}) \le D(\gamma_j\|\gamma^{(0)})$ must hold for some $i \in S, j \in S^C$. Due to the arbitrariness of $\gamma^{(0)}$, the probability of this error event can be upper bounded by the probability of the following event:

$$E_1 \triangleq \big\{\exists\, i \in S,\ \exists\, j_1, j_2 \in S^C,\ D(\gamma_i\|\gamma_{j_1}) \le D(\gamma_{j_2}\|\gamma_{j_1})\big\}. \tag{18}$$

If $\gamma^{(0)}$ is generated from an outlying distribution, the error probability can be upper bounded by the probability of the following event:

$$E_2 \triangleq \big\{\exists\, i_1, i_2 \in S,\ \exists\, j_1, j_2 \in S^C,\ D(\gamma_{j_1}\|\gamma_{i_1}) < D(\gamma_{i_2}\|\gamma_{i_1}) < D(\gamma_{j_2}\|\gamma_{i_1})\big\}. \tag{19}$$

Thus, $\mathbb{P}_S(E) \le \mathbb{P}_S(E_1) + \mathbb{P}_S(E_2)$.

We then use F to denote the event that errors occur at the assignment step, then

$$F \triangleq E^C \bigcap \big\{\exists\, i \in S,\ \exists\, j \in S^C,\ D(\gamma_i\|\hat{\pi}) \le D(\gamma_j\|\hat{\pi})\big\}. \tag{20}$$

Note that $F \subset E_1$, then the probability of error event F can be upper bounded by that of the event E_1.

The error probability of the test $\delta_2^{(1)}$ can be bounded by

$$e\big(\delta_2^{(1)}\big) = \mathbb{P}_S\big(E \cup F\big) \le \mathbb{P}_S(E) + \mathbb{P}_S(F). \tag{21}$$

The right hand side of (21) can be further bounded by

$$\begin{aligned}
&\mathbb{P}_S(E) + \mathbb{P}_S(F)\\
&\le \mathbb{P}_S(E_1) + \mathbb{P}_S(E_2) + \mathbb{P}_S(E_1)\\
&\le 2\mathbb{P}_S\left(\bigcup_{\substack{j_1,j_2\in S^C\\ i\in S}} \{D(\gamma_i\|\gamma_{j_1}) \le D(\gamma_{j_2}\|\gamma_{j_1})\}\right)\\
&\quad + \mathbb{P}_S\left(\bigcup_{\substack{j_1,j_2\in S^C\\ i_1,i_2\in S}} \{D(\gamma_{j_1}\|\gamma_{i_1}) < D(\gamma_{i_2}\|\gamma_{i_1}) < D(\gamma_{j_2}\|\gamma_{i_1})\}\right)\\
&\overset{(a)}{\le} (M-T)^2T^2\Big(2\max_{i\in S}\mathbb{P}_S\big(D(\gamma_i\|\gamma_{j_1}) \le D(\gamma_{j_2}\|\gamma_{j_1})\big)\\
&\quad + \max_{i_1,i_2\in S}\mathbb{P}_S\big(D(\gamma_{j_1}\|\gamma_{i_1}) < D(\gamma_{i_2}\|\gamma_{i_1}) < D(\gamma_{j_2}\|\gamma_{i_1})\big)\Big),
\end{aligned} \tag{22}$$

where the union bound (a) holds for all $j_1, j_2 \in S^C$, since all typical sequences are generated from the same distribution π.

From Lemma 1, we know the exponent can be computed as

$$\begin{aligned}
\alpha_1 &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{i\in S}\mathbb{P}_S\big(D(\gamma_i\|\gamma_{j_1}) \le D(\gamma_{j_2}\|\gamma_{j_1})\big)\\
&= \min_{\substack{q_1,q_2,q_3\in C_1\\ i\in S}} D(q_1\|\mu_i) + D(q_2\|\pi) + D(q_3\|\pi),
\end{aligned} \tag{23}$$

where $C_1 \triangleq \{(q_1,q_2,q_3) : D(q_1\|q_2) \le D(q_3\|q_2)\}$, and

$$\begin{aligned}
\alpha_2 &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{i_1,i_2\in S}\mathbb{P}_S\big(D(\gamma_{j_1}\|\gamma_{i_1})\\
&\qquad\qquad < D(\gamma_{i_2}\|\gamma_{i_1}) < D(\gamma_{j_2}\|\gamma_{i_1})\big)\\
&= \min_{\substack{q_1,q_2,q_3,q_4\in C_2\\ i_1,i_2\in S}} \big((D(q_1\|\pi) + D(q_2\|\pi)\\
&\qquad\qquad + D(q_3\|\mu_{i_1}) + D(q_4\|\mu_{i_2})\big),
\end{aligned} \tag{24}$$

$$C_2 \triangleq \big\{(q_1,q_2,q_3,q_4) : D(q_1\|q_3) < D(q_4\|q_3) < D(q_2\|q_3)\big\}.$$

It can be verified that the objective function in (23) can only be zero for the case $q_1 = \mu_i$, $q_2 = q_3 = \pi$, which are not in the constraint set C_1. The objective function in (24) can only be zero when $q_1 = q_2 = \pi$, $q_3 = \mu_{i_1}, q_4 = \mu_{i_2}$, which cannot meet the constraint in set C_2 either. Thus, we can conclude that $\alpha_1,\ \alpha_2 > 0$. From the fact that $\lim_{n\to\infty}\frac{\log M(M-T)}{n} = 0$, we get that

$$\alpha\big(\delta_2^{(1)}\big) = \lim_{n\to\infty} -\frac{1}{n}\log e\big(\delta_2^{(1)}\big) \ge \min\{\alpha_1, \alpha_2\}. \tag{25}$$

This result shows that the one step test $\delta_2^{(1)}$ is universally exponentially consistent.

We next derive a lower bound for $e(\delta_2^{(1)})$ by considering the error exponents for specific error events. To simplify the notation, we assume that $S = \{M-T+1, \ldots, M\}$, i.e., that $Y^{(M-T+1)}, \ldots, Y^{(M)}$ are the outlying sequences, and that

$$D(\mu_{M-T+1}\|\pi) \le D(\mu_{M-T+2}\|\pi) \le \cdots \le D(\mu_M\|\pi). \tag{26}$$

Since $T < M/2$, it holds that $Y^{(\lceil\frac{M}{2}\rceil)}$ is a typical sequence. We then bound $e(\delta_2^{(1)})$ for the following two cases, depending on whether $Y^{(\lceil\frac{M}{2}\rceil)+1}$ is typical or not.

Case I: $T < \lfloor\frac{M}{2}\rfloor$*:* It can be verified that $\lceil\frac{M}{2}\rceil + 1 \le M - T$, and hence $Y^{(\lceil\frac{M}{2}\rceil+1)}$ and $Y^{(M-T)}$ are two distinct typical sequences. We consider the following event,

$$A \triangleq \big\{\gamma^{(0)} = \gamma_1, D(\gamma_2\|\gamma_1) \le D(\gamma_3\|\gamma_1) \le \cdots \le D(\gamma_M\|\gamma_1)\big\}.$$

Since $Y^{(\lceil\frac{M}{2}\rceil)}$ is a typical sequence, the constructed cluster center in the initialization step $\hat{\pi} = \gamma_{\lceil\frac{M}{2}\rceil}$ is generated from the typical distribution, conditioned on the event A. This means that $A \subseteq E^C$.

Note that $Y^{(\lceil \frac{M}{2} \rceil)}$ and $Y^{(M-T)}$ are distinct. We further consider the following error event in the assignment step,

$$B \triangleq \left\{ D(\gamma_{M-T+1} \| \gamma_{\lceil \frac{M}{2} \rceil}) \leq D(\gamma_{M-T} \| \gamma_{\lceil \frac{M}{2} \rceil}) \right\}. \quad (27)$$

Since $Y^{(M-T+1)}$ is an outlying sequence, and $Y^{(M-T)}$ is a typical one, we have that $A \cap B \subset F$. Thus,

$$e\big(\delta_2^{(1)}\big) \geq \mathbb{P}_S(F) \geq \mathbb{P}_S(A \cap B). \quad (28)$$

Case II: $T = \lfloor \frac{M}{2} \rfloor$ It can be verified that $M - T = \lceil \frac{M}{2} \rceil$, and hence $Y^{(\lceil \frac{M}{2} \rceil + 1)}$ is an outlying sequence. Consider the event

$$\begin{aligned} A' \triangleq \big\{ & \gamma^{(0)} = \gamma_1, D(\gamma_2 \| \gamma_1) \leq \cdots \leq D(\gamma_{M-T-1} \| \gamma_1) \\ & \leq D(\gamma_{M-T+1} \| \gamma_1) \leq D(\gamma_{\lceil \frac{M}{2} \rceil} \| \gamma_1) \\ & \leq D(\gamma_{M-T+2} \| \gamma_1) \leq \cdots \leq D(\gamma_M \| \gamma_1) \big\}, \end{aligned} \quad (29)$$

which means that the $\hat{\pi}$ chosen in the initialization step is γ_{M-T+1}. Since $Y^{(M-T+1)}$ is an outlying sequence, we have that $A' \subset E$, and $e\big(\delta_2^{(1)}\big) \geq \mathbb{P}_S(E) \geq \mathbb{P}_S(A')$.

We note that by the algorithm, the event of choosing $\gamma^{(0)} = \gamma_1$ has probability $\frac{1}{M}$, and is independent of the observations in all the sequences. By Lemma 1, the error exponents of $\mathbb{P}_S(A \cap B)$ and $\mathbb{P}_S(A')$ can be computed as follows,

$$\begin{aligned} \alpha_3 & \triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S(A \cap B) \\ & = \min_{q_1, \ldots, q_M \in C_3} \sum_{j=1}^{M-T} D(q_j \| \pi) + \sum_{i=M-T+1}^{M} D(q_i \| \mu_i), \end{aligned} \quad (30)$$

$$\begin{aligned} C_3 \triangleq \big\{ & (q_1, \ldots, q_M) \colon D(q_{M-T+1} \| q_{\lceil \frac{M}{2} \rceil}) \leq D(q_{M-T} \| q_{\lceil \frac{M}{2} \rceil}), \\ & D(q_2 \| q_1) \leq D(q_3 \| q_1) \leq \cdots \leq D(q_M \| q_1) \big\}, \end{aligned} \quad (31)$$

$$\begin{aligned} \alpha_4 & \triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S(A') \\ & = \min_{q_1, \ldots, q_M \in C_4} \sum_{j=1}^{M-T} D(q_j \| \pi) + \sum_{i=M-T+1}^{M} D(q_i \| \mu_i), \end{aligned} \quad (32)$$

$$\begin{aligned} C_4 \triangleq \big\{ & (q_1, \ldots, q_M) : \ D(q_2 \| q_1) \leq \cdots \leq D(q_{M-T-1} \| q_1) \\ & \leq D(q_{M-T+1} \| q_1) \leq D(q_{M-T} \| q_1) \\ & \leq D(q_{M-T+2} \| q_1) \leq \cdots \leq D(q_M \| q_1) \big\}. \end{aligned} \quad (33)$$

If we add the constraint that $q_{M-T+1} = q_{M-T}$, i.e., $C_3' = C_3 \cap \{q_{M-T+1} = q_{M-T} = q\}$ and $C_4' = C_4 \cap \{q_{M-T+1} = q_{M-T} = q\}$, then $C_3' \subset C_3$ and $C_4' \subset C_4$, and

$$\begin{aligned} \alpha_3 & \leq \min_{q_1, \ldots, q_M \in C_3'} \sum_{j=1}^{M-T} D(q_j \| \pi) + \sum_{i=M-T+1}^{M} D(q_i \| \mu_i) \\ & = \min_{q \in \mathcal{P}(\mathcal{Y})} D(q \| \pi) + D(q \| \mu_{M-T+1}), \end{aligned} \quad (34)$$

$$\begin{aligned} \alpha_4 & \leq \min_{q_1, \ldots, q_M \in C_4'} \sum_{j=1}^{M-T} D(q_j \| \pi) + \sum_{i=M-T+1}^{M} D(q_i \| \mu_i) \\ & = \min_{q \in \mathcal{P}(\mathcal{Y})} D(q \| \pi) + D(q \| \mu_{M-T+1}), \end{aligned} \quad (35)$$

where the last steps of both inequalities follow by setting $q_1 = \cdots = q_{M-T-1} = \pi$, and $q_i = \mu_i$, for $i = M-T+2, \ldots, M$. It can be verified that these distributions satisfy the constraints in C_3' and C_4'. From Lemma 2, it follows that the minima are both equal to the Bhattacharyya distance between the distributions μ_{M-T+1} and π. Therefore,

$$\max\{\alpha_3, \alpha_4\} \leq 2B(\mu_{M-T+1}, \pi) = \min_{i \in S} 2B(\mu_i, \pi), \quad (36)$$

where the last step follows from (26). Thus, as $M \to \infty$,

$$\alpha\big(\delta_2^{(1)}\big) \leq \max\{\alpha_3, \alpha_4\} \leq \lim_{M \to \infty} \min_{i \in S} 2B(\mu_i, \pi). \quad (37)$$

As for the time complexity, it is obvious that the initialization step in Algorithm 2 can be executed within $O(M)$ time. The assignment step in Algorithm 2, which finds the largest T elements from size M array, can be solved in linear time $O(M)$ using the Quickselect algorithm proposed in [18]. Thus the overall time complexity is $O(M)$ and independent of T.

Appendix C
Proof of Theorem 3

The exponential consistency of $\delta_3^{(\ell)}$ can be established using techniques similar to those in Theorem 1 and Theorem 2. The major difference between the proof of Theorem 1 and Theorem 3 is that there are two cluster centers in the initialization step and assignment step in Algorithm 3.

We first establish the exponential consistency of the one-step test $\delta_3^{(1)}$. Due to the structure of the test we know that errors may occur at two different steps:

1) *Initialization Step:* The constructed cluster center for typical sequences $\hat{\pi}$ and outlying sequences $\hat{\mu}$ are actually generated from the same distribution.
2) *Assignment Step:* The empirical distribution of an outlying sequence is closer to the cluster center of the typical sequence $\hat{\pi}$, and vice versa.

We use E to denote the event that errors occur in the initialization step. The error event E can be decomposed into two parts, since $\gamma^{(0)}$ is chosen arbitrarily and can be generated from π or μ:

$$E \triangleq E_1 \cup E_2, \quad (38)$$

where

$$\begin{aligned} E_1 \triangleq \Big\{ & \max_{j \in S^C} D(\gamma_j \| \gamma^{(0)}) > \max_{i \in S} D(\gamma_i \| \gamma^{(0)}), \\ & \gamma^{(0)} \text{ is generated from } \pi \Big\}, \end{aligned} \quad (39)$$

$$\begin{aligned} E_2 \triangleq \Big\{ & \max_{i \in S} D(\gamma_i \| \gamma^{(0)}) > \max_{j \in S^C} D(\gamma_j \| \gamma^{(0)}), \\ & \gamma^{(0)} \text{ is generated from } \mu \Big\}. \end{aligned} \quad (40)$$

Denote

$$A_i \triangleq \left\{ \exists \, j_2 \in S^C, \ \max_{j_1 \in S^C} D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_i \| \gamma_{j_2}) \right\}, \quad (41)$$

for all $i \in S$, and

$$B_j \triangleq \left\{ \exists\, i_2 \in S,\ \max_{i_i \in S} D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_j \| \gamma_{i_2}) \right\}, \quad (42)$$

for all $j \in S^C$. Since $\gamma^{(0)}$ is chosen arbitrarily, we have

$$E = \left(\bigcap_{i \in S} A_i \right) \bigcup \left(\bigcap_{j \in S^C} B_j \right). \quad (43)$$

We use F to denote the event that errors occur at the assignment step, given that the clustering center c^1 and c^2 chosen by Algorithm 3 are coming from different distributions. We further denote the cluster center which is actually generated from the typical (outlying) distribution by $\hat{\pi}$ ($\hat{\mu}$). Then F can be written as

$$F \triangleq F_1 \cup F_2, \quad (44)$$

$$F_1 \triangleq E^C \bigcap \left\{ \exists\, j \in S^C,\ D(\gamma_j \| \hat{\pi}) > D(\gamma_j \| \hat{\mu}) \right\}, \quad (45)$$

$$F_2 \triangleq E^C \bigcap \left\{ \exists\, i \in S,\ D(\gamma_i \| \hat{\mu}) > D(\gamma_i \| \hat{\pi}) \right\}. \quad (46)$$

Thus, we can upper bound the error probability of the one-step test $\delta_3^{(1)}$ by

$$e\big(\delta_3^{(1)}\big) = \mathbb{P}_S\big(E \cup F\big) \le \mathbb{P}_S(E) + \mathbb{P}_S(F). \quad (47)$$

The first term on the right hand side can be bounded as,

$$\begin{aligned}
\mathbb{P}_S(E) &\le \mathbb{P}_S\left(\bigcap_{i \in S} A_i \right) + \mathbb{P}_S\left(\bigcap_{j \in S^C} B_j \right) \\
&\le \mathbb{P}_S(A_i) + \mathbb{P}_S(B_j) \\
&\overset{(a)}{\le} (M - |S|)\mathbb{P}_S\Big(\max_{j_1 \in S^C} D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_i \| \gamma_{j_2}) \Big) \\
&\quad + |S|\mathbb{P}_S\Big(\max_{i_i \in S} D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_j \| \gamma_{i_2}) \Big) \\
&\overset{(b)}{\le} (M - |S|)^2 \mathbb{P}_S\Big(D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_i \| \gamma_{j_2}) \Big) \\
&\quad + |S|^2 \mathbb{P}_S\Big(D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_j \| \gamma_{i_2}) \Big), \quad (48)
\end{aligned}$$

where the union bound (a) and (b) holds for all $j, j_1, j_2 \in S^C$ and $i, i_1, i_2 \in S$, since all typical distributions are identical and all outlying distributions are identical.

From Lemma 1, we obtain

$$\begin{aligned}
\alpha_5 &\triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S\Big(D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_i \| \gamma_{j_2}) \Big) \\
&= \min_{q_1, q_2, q_3 \in C_5} D(q_1 \| \pi) + D(q_2 \| \pi) + D(q_3 \| \mu), \quad (49)
\end{aligned}$$

where $C_5 \triangleq \big\{ (q_1, q_2, q_3) : D(q_1 \| q_2) > D(q_3 \| q_2) \big\}$, and

$$\begin{aligned}
\alpha_6 &\triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S\Big(D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_j \| \gamma_{i_2}) \Big) \\
&= \min_{q_1, q_2, q_3 \in C_6} D(q_1 \| \mu) + D(q_2 \| \mu) + D(q_3 \| \pi), \quad (50)
\end{aligned}$$

where $C_6 \triangleq \big\{ (q_1, q_2, q_3) : D(q_1 \| q_2) > D(q_3 \| q_2) \big\}$.

We then upper bound $\mathbb{P}_S(F)$ by using the Union Bound [25] as follows:

$$\begin{aligned}
\mathbb{P}_S(F) &\le \mathbb{P}_S(F_1) + \mathbb{P}_S(F_2) \\
&\le \mathbb{P}_S\left(\bigcup_{j \in S^C} \big\{ D(\gamma_j \| \hat{\pi}) > D(\gamma_j \| \hat{\mu}) \big\} \right) \\
&\quad + \mathbb{P}_S\left(\bigcup_{i \in S} \big\{ D(\gamma_i \| \hat{\mu}) > D(\gamma_i \| \hat{\pi}) \big\} \right) \\
&\le |S|(M - |S|)^2 \mathbb{P}_S\Big(D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_{j_1} \| \gamma_i) \Big) \\
&\quad + |S|^2 (M - |S|) \mathbb{P}_S\Big(D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_{i_1} \| \gamma_j) \Big), \quad (51)
\end{aligned}$$

where $j, j_1, j_2 \in S^C$ and $i, i_1, i_2 \in S$. From Lemma 1, we obtain

$$\begin{aligned}
\alpha_7 &\triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S\Big(D(\gamma_{j_1} \| \gamma_{j_2}) > D(\gamma_{j_1} \| \gamma_i) \Big) \\
&= \min_{q_1, q_2, q_3 \in C_7} D(q_1 \| \pi) + D(q_2 \| \pi) + D(q_3 \| \mu), \quad (52)
\end{aligned}$$

where $C_7 \triangleq \{(q_1, q_2, q_3) : D(q_1 \| q_2) > D(q_1 \| q_3)\}$, and

$$\begin{aligned}
\alpha_8 &\triangleq \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}_S\Big(D(\gamma_{i_1} \| \gamma_{i_2}) > D(\gamma_{i_1} \| \gamma_j) \Big) \\
&= \min_{q_1, q_2, q_3 \in C_8} D(q_1 \| \mu) + D(q_2 \| \mu) + D(q_3 \| \pi), \quad (53)
\end{aligned}$$

where $C_8 \triangleq \big\{ (q_1, q_2, q_3) : D(q_1 \| q_2) > D(q_1 \| q_3) \big\}$.

Due to the fact that the objective functions in (49) and (52) can only be zero for the case $q_1 = q_2 = \pi$, $q_3 = \mu$, which is not in the constraint sets C_5 and C_7, respectively. The objective functions in (50) and (53) can only be zero when $q_1 = q_2 = \mu$, $q_3 = \pi$, which cannot meet the constraints in sets C_6 and C_8 either. Thus, we conclude that $\alpha_5,\ \alpha_6,\ \alpha_7,\ \alpha_8 > 0$.

From the fact that $\lim_{n \to \infty} \frac{\log M(M - |S|)}{n} = 0$, it then follows that

$$\alpha\big(\delta_3^{(1)}\big) \ge \min\big\{ \alpha_5,\ \alpha_6,\ \alpha_7,\ \alpha_8 \big\}. \quad (54)$$

From the above argument and Proposition 2, both the one-step test $\delta_3^{(1)}$ and the GL test δ_{GL} are exponentially consistent. Thus, based on the same technique used in the proof of Theorem 2, we establish the exponential consistency of the test $\delta_3^{(\ell)}$ proposed in Algorithm 3, for any $\ell \ge 1$.

Finally, since each iteration has the time complexity $O(M)$, $\delta_3^{(\ell)}$ which runs ℓ iterations has time complexity $O(M\ell)$.

Appendix D
Proof of Theorem 4

The exponential consistency of $\delta_3^{(1)}$ for the scenario where the typical and outlying distributions form clusters can be established using the same techniques as in Theorem 3. The major difference between the proofs of Theorem 3 and Theorem 4 is that here both the typical distributions and the outlying distributions are distinct.

Using the same events defined in Appendix C, the error probability of the one-step test $\delta_3^{(1)}$ can be upper bounded by

$$e\big(\delta_3^{(1)}\big) = \mathbb{P}_S\big(E \cup F\big) \leq \mathbb{P}_S(E) + \mathbb{P}_S(F). \tag{55}$$

The first term on the right hand side can be bounded as,

$$\begin{aligned}
&\mathbb{P}_S(E)\\
&\leq \mathbb{P}_S\left(\bigcap_{i\in S} A_i\right) + \mathbb{P}_S\left(\bigcap_{j\in S^C} B_j\right)\\
&\leq \mathbb{P}_S(A_i) + \mathbb{P}_S(B_j)\\
&\leq (M-|S|)^2 \max_{\substack{j_1,j_2\in S^C\\ i\in S}} \mathbb{P}_S\left(D(\gamma_{j_1}\|\gamma_{j_2}) > D(\gamma_i\|\gamma_{j_2})\right)\\
&\quad + |S|^2 \max_{\substack{i_1,i_2\in S\\ j\in S^C}} \mathbb{P}_S\Big(D(\gamma_{i_1}\|\gamma_{i_2}) > D(\gamma_j\|\gamma_{i_2})\Big).
\end{aligned} \tag{56}$$

From Lemma 1, we obtain

$$\begin{aligned}
\alpha_9 &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{\substack{j_1,j_2\in S^C\\ i\in S}} \mathbb{P}_S\Big(D(\gamma_{j_1}\|\gamma_{j_2}) > D(\gamma_i\|\gamma_{j_2})\Big)\\
&= \min_{\substack{j_1,j_2\in S^C\\ i\in S}} \min_{q_1,q_2,q_3\in C_9} D(q_1\|\pi_{j_1}) + D(q_2\|\pi_{j_2})\\
&\qquad + D(q_3\|\mu_i),
\end{aligned} \tag{57}$$

where $C_9 \triangleq \big\{(q_1,q_2,q_3) : D(q_1\|q_2) > D(q_3\|q_2)\big\}$, and

$$\begin{aligned}
\alpha_{10} &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{\substack{i_1,i_2\in S\\ j\in S^C}} \mathbb{P}_S\Big(D(\gamma_{i_1}\|\gamma_{i_2}) > D(\gamma_j\|\gamma_{i_2})\Big)\\
&= \min_{\substack{i_1,i_2\in S\\ j\in S^C}} \min_{q_1,q_2,q_3\in C_{10}} D(q_1\|\mu_{i_1}) + D(q_2\|\mu_{i_2})\\
&\qquad + D(q_3\|\pi_j),
\end{aligned} \tag{58}$$

where $C_{10} \triangleq \big\{(q_1,q_2,q_3) : D(q_1\|q_2) > D(q_3\|q_2)\big\}$.

We then upper bound $\mathbb{P}_S(F)$ by using the Union Bound [25] as follows,

$$\begin{aligned}
&\mathbb{P}_S(F)\\
&\leq \mathbb{P}_S(F_1) + \mathbb{P}_S(F_2)\\
&\leq \mathbb{P}_S\left(\bigcup_{j\in S^C}\big\{D(\gamma_j\|\hat{\pi}) > D(\gamma_j\|\hat{\mu})\big\}\right)\\
&\quad + \mathbb{P}_S\left(\bigcup_{i\in S}\big\{D(\gamma_i\|\hat{\mu}) > D(\gamma_i\|\hat{\pi})\big\}\right)\\
&\leq |S|(M-|S|)^2 \max_{\substack{j_1,j_2\in S^C\\ i\in S}} \mathbb{P}_S\Big(D(\gamma_{j_1}\|\gamma_{j_2}) > D(\gamma_{j_1}\|\gamma_i)\Big)\\
&\quad + |S|^2(M-|S|) \max_{\substack{i_1,i_2\in S\\ j\in S^C}} \mathbb{P}_S\Big(D(\gamma_{i_1}\|\gamma_{i_2}) > D(\gamma_{i_1}\|\gamma_j)\Big).
\end{aligned} \tag{59}$$

From Lemma 1, we obtain

$$\begin{aligned}
\alpha_{11} &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{\substack{j_1,j_2\in S^C\\ i\in S}} \mathbb{P}_S\Big(D(\gamma_{j_1}\|\gamma_{j_2}) > D(\gamma_{j_1}\|\gamma_i)\Big)\\
&= \min_{\substack{j_1,j_2\in S^C\\ i\in S}} \min_{q_1,q_2,q_3\in C_{11}} D(q_1\|\pi_{j_1}) + D(q_2\|\pi_{j_2})\\
&\qquad + D(q_3\|\mu_i),
\end{aligned} \tag{60}$$

where $C_{11} \triangleq \big\{(q_1,q_2,q_3) : D(q_1\|q_2) > D(q_1\|q_3)\big\}$, and

$$\begin{aligned}
\alpha_{12} &\triangleq \lim_{n\to\infty} -\frac{1}{n}\log \max_{\substack{i_1,i_2\in S\\ j\in S^C}} \mathbb{P}_S\Big(D(\gamma_{i_1}\|\gamma_{i_2}) > D(\gamma_{i_1}\|\gamma_j)\Big)\\
&= \min_{\substack{i_1,i_2\in S\\ j\in S^C}} \min_{q_1,q_2,q_3\in C_{12}} D(q_1\|\mu_{i_1}) + D(q_2\|\mu_{i_2}) + D(q_3\|\pi_j),
\end{aligned} \tag{61}$$

where $C_{12} \triangleq \big\{(q_1,q_2,q_3) : D(q_1\|q_2) > D(q_1\|q_3)\big\}$.

Note that the objective functions in (57) and (60) can only be zero for the case $q_1 = \pi_{j_1}, q_2 = \pi_{j_2}, q_3 = \mu_i$, which is not in the constraint sets C_9 and C_{11}, due to our clustering assumption (1). The objective functions in (58) and (61) can only be zero when $q_1 = \mu_{i_1}, q_2 = \mu_{i_2}, q_3 = \pi_j$, which cannot meet the constraints in sets C_{10} and C_{12} either. Thus, we conclude that $\alpha_9,\ \alpha_{10},\ \alpha_{11},\ \alpha_{12} > 0$.

From the fact that $\lim_{n\to\infty} \frac{\log M(M-|S|)}{n} = 0$, it follows

$$\alpha\big(\delta_3^{(1)}\big) \geq \min\big\{\alpha_9,\ \alpha_{10},\ \alpha_{11},\ \alpha_{12}\big\}. \tag{62}$$

REFERENCES

[1] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2017, pp. 988–992.

[2] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 44–56, Sep. 2014.

[3] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Sci.*, vol. 17, pp. 235–249, 2002.

[4] J. Chamberland and V. V. Veeravalli, "Wireless sensors in distributed detection applications," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 16–25, May 2007.

[5] S. Vucetic, D. Pokrajac, H. Xie, and Z. Obradovic, "Detection of underrepresented biological sequences using class-conditional distribution models," in *Proc. SIAM Int. Conf. Data Mining*, 2003, pp. 279–283.

[6] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[7] Y. Li and V. V. Veeravalli, "Outlying sequence detection in large datasets: Comparison of universal hypothesis testing and clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 6180–6184.

[8] K. Chaudhuri and A. McGregor, "Finding metric structure in information theoretic clustering," in *Proc. Annu. Conf. Learn. Theory*, 2008, vol. 8, pp. 391–402.

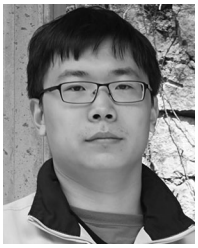
[9] M. R Ackermann, J. Blömer, and C. Sohler, "Clustering for metric and nonmetric distance measures," *ACM Trans. Algorithms*, vol. 6, no. 4, pp. 59:1–59:26, 2010.

[10] R. Nock, P. Luosto, and J. Kivinen, "Mixed Bregman clustering with approximation guarantees," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2008, pp. 154–169.

[11] L. Xiong, B. Póczos, and J. Schneider, "Group anomaly detection using flexible genre models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1071–1079.

[12] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, Jul. 2014.

[13] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Universal outlying sequence detection for continuous observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4254–4258.
[14] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5785–5797, Nov. 2017.
[15] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[16] J. Blömer, C. Lammersen, M. Schmidt, and C. Sohler, "Theoretical analysis of the K-means algorithm—A survey," in *Algorithm Engineering*. Berlin, Germany: Springer, 2016, pp. 81–116.
[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.
[18] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, "Time bounds for selection," *J. Comput. Syst. Sci.*, vol. 7, no. 4, pp. 448–461, 1973.
[19] B. Póczos, L. Xiong, and J. Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," in *Proc. Conf. Uncertainty Artif. Intell.*, 2011, pp. 599–608.
[20] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, pp. 697–717, 1979.
[21] Climate Prediction Center, "CPC global summary of day/month observations, 1979-continuing," Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO, 1987. [Online]. Available: http://rda.ucar.edu/datasets/ds512.0/
[22] M. C. Peel, B. Finlayson, and T. A. McMahon, "Updated world map of the Köppen–Geiger climate classification," *Hydrol. Earth Syst. Sci. Discuss.*, vol. 4, no. 2, pp. 439–473, 2007.
[23] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of KL divergence: Optimal minimax rate," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2648–2674, Apr. 2018.
[24] C. Boutsidis, A. Zouzias, and P. Drineas, "Random projections for k-means clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 298–306.
[25] R. Durrett, *Probability: Theory and Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2010.



Yuheng Bu (S'16) received the B.S. (Hons.) degree in electrical engineering from Tsinghua University, Beijing, China, in 2014, and the M.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2016. He is currently working toward the Ph.D. degree with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. His research interests include statistical signal processing, machine learning, and information theory.

Shaofeng Zou (S'14–M'16) received the B.E. (Hons.) degree from Shanghai Jiao Tong University, Shanghai, China, in 2011, and the Ph.D. degree in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 2016. From 2016 to 2018, he was a Postdoctoral Research Associate with the Coordinated Science Lab, University of Illinois at Urbana-Champaign. He joined the Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA, in 2018, where he is currently an Assistant Professor. His research interests include statistical signal processing, machine learning, and information theory.

Venugopal V. Veeravalli (M'92–SM'98–F'06) received the B.Tech. (Silver Medal Hons.) degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1985, the M.S. degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1987, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992.

He joined the University of Illinois at Urbana-Champaign in 2000, where he is currently the Henry Magnuski Professor with the Department of Electrical and Computer Engineering, and where he is also affiliated with the Department of Statistics, the Coordinated Science Laboratory, and the Information Trust Institute. From 2003 to 2005, he was a Program Director for communications research at the U.S. National Science Foundation, Arlington, VA, USA. He has previously held academic positions at Harvard University, Rice University, and Cornell University, and has been on sabbatical at MIT, IISc Bangalore, and Qualcomm, Inc. His research interests include statistical signal processing, machine learning, detection and estimation theory, information theory, and stochastic control, with applications to sensor networks, cyberphysical systems, and wireless communications. A recent emphasis of his research has been on signal processing and machine learning for data science applications.

Prof. Veeravalli was a Distinguished Lecturer for the IEEE Signal Processing Society from 2010 to 2011. He has been on the Board of Governors of the IEEE Information Theory Society. He has been an Associate Editor for Detection and Estimation for the IEEE TRANSACTIONS ON INFORMATION THEORY and for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was the recipient (for research and teaching) of the IEEE Browder J. Thompson Best Paper Award, the National Science Foundation CAREER Award, and the Presidential Early Career Award for Scientists and Engineers, and the Wald Prize in Sequential Analysis.