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

Yuheng Bu Shaofeng Zou Venugopal V. Veeravalli
University of Illinois at Urbana-Champaign
Email: bu3@illinois.edu, szou3@illinois.edu, vvv@illinois.edu

***Abstract*—We study a universal outlying sequence detection problem, in which there are M sequences of samples out of which a small subset of outliers need to be detected. A sequence is considered as an outlier 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, the goal is to identify all the outliers without any knowledge about the underlying generating distributions. In prior work, this problem was studied as a universal hypothesis testing problem, and a generalized likelihood (GL) test was constructed and its asymptotic performance characterized. In this paper, we propose a different class of tests for this problem based on distribution clustering. Such tests are shown to be exponentially consistent and their time complexity is linear in the total number of sequences, in contrast with the GL test, which has time complexity that is exponential in the number of outliers. Furthermore, our tests based on clustering are applicable to more general scenarios. For example, when both the typical and outlier distributions form clusters, the clustering based test is exponentially consistent, but the GL test is not even applicable.**

I. Introduction

In this paper, we study a universal outlying sequence detection problem, where there are M sequences of samples out of which a small subset are outliers and need to be detected. Each sequence consists of 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 an outlier if its distribution is different from the typical distributions. We are interested in the universal setting of the problem, where nothing is known about the outlier and typical distributions, except that the outlier distributions are different from the typical distributions. The goal is to design a test, which does not depend on the typical and outlier distributions, to best discern all the outliers.

Outlying sequence detection finds possible applications in many domains (e.g., [1]–[3]). The problem of outlying sequence detection was studied as a universal outlier hypothesis testing problem for discrete observations in [4] and continuous observations in [5]. In [4], the exponential consistency of the generalized likelihood (GL) test under various universal settings was established. When there are a known number of identically distributed outliers, the GL test was shown to be asymptotically optimal as M goes to infinity.

However, the high time complexity, which is exponential in the number of outliers T, is a major drawback of the GL test when M and T are large. In this paper, we propose tests based on distribution clustering [6] for various scenarios. Such tests are shown to be exponentially consistent, with time complexity that is linear in M and independent of T.

The intuition for the clustering based test is that the typical distributions are usually closer to each other than the outlier distributions. Therefore, the typical distributions (and also possibly the outlier distributions) will naturally form a cluster. This implies that the outlying sequence detection problem is equivalent to clustering the distributions using KL divergence as the distance metric (see also [7]).

We note that our problem is different from the classical distribution clustering problem [6], [8]–[10]. In the distribution clustering problem, the goal is to find the cluster structure with the lowest cost (sum of distance functions of each point in the cluster to the center). However, for our problem, we are given a sequence of samples from each distribution rather than the actual underlying distribution itself. Therefore, we are interested in the statistical performance, i.e., the error probability, of our test. Previous studies on approximation algorithms for distribution clustering [8]–[10] only show that the cost corresponding to the cluster structure returned by the approximation is 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 a performance guarantee in a probabilistic sense.

Our contributions in this paper are as follows: (1) in all cases where the GL test is exponentially consistent, we construct tests based on clustering that are also exponentially consistent and have time complexity linear in M; (2) we show that the tests based on clustering are applicable to more general scenarios. For example, when both the typical and outlier distributions form clusters, the clustering based test is exponentially consistent, but the GL test is not even applicable.

II. Problem Model

Throughout the paper, all distributions are defined on the finite set $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ denotes the set of all probability mass functions (pmfs) on $\mathcal{Y}$.

We consider an outlying sequence detection problem, 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

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

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 outlying sequences, where $S \subset \{1, \ldots, M\}$ and $1 \leq |S| = T < \frac{M}{2}$. Each typical sequence j is distributed according to a typical distribution $\pi_j \in \mathcal{P}(\mathcal{Y}),\ j \notin S$. Each outlying sequence i is distributed according to an outlier distribution $\mu_i \in \mathcal{P}(\mathcal{Y}),\ i \in S$. Nothing is known about μ_i and π_j except that $\mu_i \neq \pi_j,\ \forall i \in S,\ j \notin S,\ S \subset \{1, \ldots, M\}$, and all of them have full support over a finite alphabet $\mathcal{Y}$.

We first study the case that all typical distributions are identical, i.e., $\pi_j = \pi,\ \forall j \notin S$.

We then study the case where both the outlier distributions $\{\mu_i\}_{i \in S}$ and the typical distributions $\{\pi_j\}_{j \notin S}$ are distinct. Moreover, the typical distributions and the outlier distributions form two clusters. More specifically,

$$\begin{aligned} \max_{i,j \in S} D(\mu_i \| \mu_j) &< \min_{i \in S, j \notin S} \{D(\mu_i \| \pi_j), D(\pi_j \| \mu_i)\}, \\ \max_{i,j \notin S} D(\pi_i \| \pi_j) &< \min_{i \in S, j \notin S} \{D(\mu_i \| \pi_j), D(\pi_j \| \mu_i)\}. \end{aligned} \tag{1}$$

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

We use the notation $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)})$, where $y_k^{(i)} \in \mathcal{Y}$ denotes the k-th observation of the i-th sequence. Let $\mathcal{S}$ be the set comprising all possible outlier subsets. For the hypothesis corresponding to an outlier subset $S \in \mathcal{S}$, the joint distribution of all the observations is given by

$$\begin{aligned} p_S(y^{Mn}) &= L_S\Big(y^{Mn}, \{\mu_i\}_{i \in S}, \{\pi_j\}_{j \notin S}\Big) \\ &= \prod_{k=1}^{n} \Big\{ \prod_{i \in S} \mu_i(y_k^{(i)}) \prod_{j \notin S} \pi_j(y_k^{(j)}) \Big\}, \end{aligned} \tag{2}$$

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

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 $\{\mu_i\}_{i=1}^M$ and $\{\pi_j\}_{j=1}^M$.

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

$$e\Big(\delta, \{\mu_i\}_{i=1}^M, \{\pi_j\}_{j=1}^M\Big) \triangleq \max_{S \in \mathcal{S}} \sum_{y^{Mn} : \delta(y^{Mn}) \neq S} p_S(y^{Mn}),$$

and the corresponding error exponent is defined as

$$\alpha\Big(\delta, \{\mu_i\}_{i=1}^M, \{\pi_j\}_{j=1}^M\Big) \triangleq \lim_{n \to \infty} -\frac{1}{n} \log e\Big(\delta, \{\mu_i\}_{i=1}^M, \{\pi_j\}_{j=1}^M\Big).$$

A universal test δ is *universally exponentially consistent* if $\alpha\big(\delta, \{\mu_i\}_{i=1}^M, \{\pi_j\}_{j=1}^M\big) > 0$, for $\mu_i \neq \pi_j,\ i, j = 1, \ldots, M$.

III. Generalized Likelihood Test

In this section, we consider the case where the typical distributions are identical, i.e., $\pi_j = \pi,\ \forall j \notin S$. Let γ_i denote the empirical distribution of $\mathbf{y}^{(i)}$, and is defined as $\gamma_i(y) \triangleq \frac{1}{m}\big|\{k = 1, \ldots, m : y_k = y\}\big|$, for each $y \in \mathcal{Y}$. Let $D(p\|q)$ and $B(p,q)$ denote the *KL divergence* and *Bhattacharyya distance*, which are defined as $D(p\|q) \triangleq \sum_{y \in \mathcal{Y}} p(y) \log\big(p(y)/q(y)\big)$ and $B(p,q) \triangleq -\log(\sum_{y \in \mathcal{Y}} p(y)^{1/2} q(y)^{1/2})$, respectively. In the universal setting with π and $\{\mu_i\}_{i \in S}$ unknown, conditioned on the set of outliers 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

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

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

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

A. Known Number of Outliers

We first consider the case in which the number of outliers, denoted by $T \geq 1$, is known at the outset, i.e., $\mathcal{S} = \{S : S \subset \{1, \ldots, M\}, |S| = T\}$. We compute the generalized likelihood of y^{Mn} by replacing the $\mu_i,\ i \in S$ and π in (2) with their MLEs: $\hat{\mu}_i \triangleq \gamma_i$, and $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} \gamma_j}{M - T}$. Then the GL test in (4) is equivalent to

$$\delta_{\text{GL}}(y^{Mn}) = \arg\min_{S \subset \mathcal{S}} \sum_{j \notin S} D\Big(\gamma_j \| \frac{\sum_{j \notin S} \gamma_j}{M - T}\Big). \tag{5}$$

Proposition 1. *[4] When the number of outliers is known, the GL test in* (5) *is universally exponentially consistent. As $M \to \infty$, the achievable error exponent converges as*

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

When all outlier 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.*

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

B. Unknown Number of Identical Outliers

In this subsection, we consider the case where the number of outliers is unknown, i.e., $\mathcal{S} = \{S : S \subset \{1, \ldots, M\}, 1 \leq |S| < M/2\}$. Moreover, we assume that the outliers are identically distributed.

By replacing the $\mu_i,\ i \in S$, and π in (2) with their MLEs $\hat{\mu}_S = \hat{\mu}_i \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$, and $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} \gamma_j}{M - |S|}$, the GL test in (4) is equivalent to

$$\begin{aligned} \delta_{\text{GL}}(y^{Mn}) = \arg\min_{S \subset \mathcal{S}} &\sum_{j \notin S} D\Big(\gamma_j \| \frac{\sum_{j \notin S} \gamma_j}{M - |S|}\Big) \\ &+ \sum_{i \in S} D\Big(\gamma_i \| \frac{\sum_{i \in S} \gamma_i}{|S|}\Big). \end{aligned} \tag{6}$$

Proposition 2. *[4] When the number of the outliers is unknown, $1 \leq |S| < \frac{M}{2}$, and all the outlier 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 complexity of the test in (6) is even larger than that of the test in (5).

We also note that when the number of the outliers is unknown and outliers can be distinctly distributed, there cannot exist a universally exponentially consistent test [4].

IV. Problem Reformulation as Distribution Clustering

In order to construct low time complexity algorithms without sacrificing much in performance, we reformulate the outlying sequence detection problem as a distribution clustering problem. Although the distribution clustering problem is known to be NP-hard [8], there exists many approximation approaches, e.g. K-means algorithm [11], with time complexity that is linear in M and linear in the number of clusters.

The typical distributions are closer to each other than to any of the outlier distributions, and the same also holds for the outlier distributions when the outliers form a cluster. The outliers can be identified by clustering the empirical distributions into two clusters, where the cluster with more members contains all typical sequences, and the other cluster contains outliers.

More specifically, if we define the following cost function for distribution clustering

$$TC = \sum_{k=1}^{K} \sum_{i \in C^k} D(\gamma_i \| c^k), \qquad (7)$$

where K is the number of clusters, $c = \{c^1, \ldots, c^K\}$ are the clustering centers, and disjoint partitions $C = \{C^1, \ldots, C^K\}$ denote the cluster assignment. As shown in [6, Proposition 1], for a given cluster assignment $\{C^k\}_{k=1}^K$, the total 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.

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: $\{c_k\}_{k=1}^K$ (Specified in Algorithm 2 and 3)
Method:
while not convergence **do**
  $\{$Assignment Step$\}$
  Set $C^k \leftarrow \emptyset$, $1 \le k \le K$
  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 $\{C^k\}_{k=1}^K$

Then the GL test in (6) can be interpreted as a distribution clustering algorithm with $K = 2$. The first term in (6) corresponds to the cost in the typical cluster and the second term is the cost within the outlier cluster. The GL test decides on the cluster assignment that minimizes the cost among all possible cluster assignments.

For the case where the typical distributions are identically distributed, but outliers are not, it suffices to only cluster the empirical distributions of all typical sequences as shown in the GL test (5).

Thus, both the GL test 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 complexity [11]. Here, we introduce the K-means distribution clustering algorithm in [6]. For Algorithm 1, [6] only shows that the cost function in (7) is monotonically decreasing and it terminates in a finite number of steps at a partition that is locally optimal.

V. Clustering Based Tests

In this section, we propose linear complexity tests based on the K-means clustering algorithm. We show in all cases where the GL test is exponentially consistent, the clustering based tests using KL divergence as the distance metric are also exponentially consistent, while only taking linear time in M. For the case that the typical and outlier 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 Outliers

We first consider the case where the number of outliers T is known and the typical distributions π are identical. Algorithm 1 cannot be directly applied here, because the outlier 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.

Algorithm 2 Clustering with known number of outliers

Input: $\gamma_1, \ldots, \gamma_M$, number of the outliers T.
Output: A set of outliers S.
Initialization:
Choose one distribution $\gamma^{(0)}$ from $\gamma_1, \ldots, \gamma_M$ arbitrarily
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 all $D(\gamma_i \| \gamma^{(0)})$
Method:
while not convergence **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$\}$
  $\hat{\pi} \leftarrow \frac{\sum_{j \notin S} \gamma_j}{M - T}$
end while
Return S

The novelty of this algorithm lies in the construction of the first clustering center for the typical distribution and an iterative approach based on K-means to update it.

Using 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 typical distribution π, 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, 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 is close to $D(\pi \| \mu_i)$, and γ^* is generated from π with high probability.

Let δ_{c2} denote the test described in Algorithm 2, and $\delta_{c2}^{(\ell)}$ denote the test that runs ℓ number of K-means iterations in Algorithm 2.

In the following theorem, we show that the test $\delta_{c2}^{(1)}$ with only one iteration step is universally exponentially consistent.

Theorem 1. *For each $M \geq 3$, when the number of outliers T is known, the test $\delta_{c2}^{(1)}$, which runs one K-means iteration in Algorithm 2 is universally exponentially consistent. The achievable error exponent of $\delta_{c2}^{(1)}$ can be upper bounded by*

$$\alpha\Big(\delta_{c2}^{(1)}, \{\mu_i\}_{i=1}^M, \pi\Big) < \lim_{M\to\infty} \min_{i=1,\dots,M} 2B(\mu_i, \pi). \quad (8)$$

Furthermore, the time complexity of the test $\delta_{c2}^{(1)}$ is $O(M)$.

Proof sketch: Errors made by $\delta_{c2}^{(1)}$ in the initialization step can be decomposed into two scenarios. If $\gamma^{(0)}$ is generated from typical distribution π, an error occurs when $\hat{\pi}$ is actually generated from an outlier distribution. The probability of this event can be upper bounded by the probability of the following event $E_1 = \{D(\gamma_i \| \gamma_{j_1}) < D(\gamma_{j_2} \| \gamma_{j_1}), \exists i \in S, \exists\ j_1, j_2 \notin S\}$. If $\gamma^{(0)}$ is generated from an outlier distribution, the error probability can be upper bounded by the probability of the following event $E_2 = \{D(\gamma_{j_1} \| \gamma_{i_1}) < D(\gamma_{i_2} \| \gamma_{i_1}) < D(\gamma_{j_2} \| \gamma_{i_1}), \exists i_1, i_2 \in S, \exists\ j_1, j_2 \notin S\}$. By Sanov's theorem, we can prove that the probability of 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.

Moreover, the assignment step in Algorithm 2 can be solved in linear time $O(M)$ [12], which is independent of T.

The details of the proof can be found in [13]. ■

Comparison of Proposition 1 and Theorem 1 shows that $\delta_{c2}^{(1)}$ has a smaller error exponent than that of the GL test in (5) as $M \to \infty$, but has a linear time complexity.

In the following theorem, we further show that the performance of Algorithm 2 will improve with more iterations.

Theorem 2. *For each $M \geq 3$, when the number of outliers T is known, the test $\delta_{c2}^{(\ell)}$ is universally exponentially consistent. As $M \to \infty$, the achievable error exponent of $\delta_{c2}^{(\ell)}$ in Algorithm 2 can be lower bounded by*

$$\lim_{M\to\infty} \alpha\Big(\delta_{c2}^{(\ell)}, \{\mu_i\}_{i=1}^M, \pi\Big) \geq \alpha\Big(\delta_{c2}^{(1)}, \{\mu_i\}_{i=1}^M, \pi\Big). \quad (9)$$

Furthermore, the time complexity of the test $\delta_{c2}^{(\ell)}$ is $O(M\ell)$.

Proof sketch: It is shown in Theorem 1 and Proposition 1, both the test $\delta_{c2}^{(1)}$ and the GL test δ_{GL} are exponentially consistent, and the error exponent of δ_{GL} is larger than that of $\delta_{c2}^{(1)}$, when $M \to \infty$. Since $\mathbb{P}(\delta_{c2}^{(1)} \neq S) \leq \mathbb{P}(\delta_{c2}^{(1)} \neq S \text{ or } \delta_{\mathrm{GL}} \neq S) \leq \mathbb{P}(\delta_{c2}^{(1)} \neq S) + \mathbb{P}(\delta_{\mathrm{GL}} \neq S)$, we can conclude that $\mathbb{P}(\delta_{c2}^{(1)} \neq S \text{ or } \delta_{\mathrm{GL}} \neq S)$ also decays exponentially fast with the same error exponent of $\delta_{c2}^{(1)}$, which means that δ_{GL} and $\delta_{c2}^{(1)}$ both output the same correct S with high probability. Given that $\delta_{c2}^{(1)}$ and δ_{GL} have same outcomes, i.e., the initialization achieves the global optimum of the cost function, then running ℓ steps will not change the output of $\delta_{c2}^{(1)}$. Thus, $\delta_{c2}^{(\ell)}$ at least achieves the error exponent of $\delta_{c2}^{(1)}$. The details of the proof can be found in [13]. ■

B. Unknown Number of Identical Outliers

In this section, we consider the case where the number of outliers is unknown. Moreover, the typical distributions π are identical and the outlier distributions μ are identical. Motivated by the test in (6), we design the following initialization algorithm to set the clustering centers in Algorithm 1.

Algorithm 3 Clustering with unknown number of outliers

Input: M empirical distributions $\gamma_1, \dots, \gamma_M$ defined on finite alphabet $\mathcal{Y}$.
Output: A set of outliers 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 C^1 and C^2

It can be seen that, c_1 and c_2 chosen by the initialization step in Algorithm 3 are generated by π and μ, with high probability.

Let δ_{c3} denote the test described in Algorithm 3, and $\delta_{c3}^{(\ell)}$ denote the test that runs ℓ iterations in Algorithm 2, where $\ell = 1, 2, \dots$ is the number of iterations. The following theorem shows that the clustering based algorithm $\delta_{c3}^{(\ell)}$, is universally exponentially consistent, and has time complexity linear in M.

Theorem 3. *For each $M \geq 3$, when the number of outliers is unknown, the test $\delta_{c3}^{(\ell)}$, which runs ℓ steps of Algorithm 3, is exponentially consistent, and has time complexity $O(M\ell)$.*

Proof sketch: The exponential consistency of $\delta_{c3}^{(\ell)}$ 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 clustering centers in the initialization step and assignment step in Algorithm 3. The details of the proof can be found in [13]. ■

C. Typical and Outlier Distributions Forming Clusters

In this subsection, we consider the case that both the outlier distributions $\{\mu_i\}_{i \in S}$ and the typical distributions $\{\pi_j\}_{j \notin S}$ are distinct. Moreover, the typical distributions and the outlier 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 $\delta_{c3}^{(1)}$ proposed in Algorithm 3 is universally exponentially consistent, and has time complexity linear in M.

Theorem 4. *For each $M \geq 3$, when both the outlier distributions $\{\mu_i\}_{i\in S}$ and the typical distributions $\{\pi_j\}_{j\notin S}$ form clusters, i.e. condition* (1) *holds, the test $\delta_{c3}^{(1)}$, which runs one step of Algorithm 3, is universally exponentially consistent, and has time complexity $O(M)$.*

Proof sketch: The exponential consistency of $\delta_{c3}^{(1)}$ can be established using similar techniques as shown in Theorem 3. The details of the proof can be found in [13]. ■

The GL approach of replacing the true distribution in (2) by their MLEs leads to identical likelihood estimates for each hypothesis. Thus, the GL approach is not applicable here. One could apply the test in (6) to this problem, but it can be shown (see [13]) that the test in (6) is not universally exponentially consistent, even if condition (1) holds.

VI. Numerical Results

In this section, we compare the performance of the GL test δ_{GL}, the clustering based tests δ_{c2}, δ_{c3} and the one step tests $\delta_{c2}^{(1)}$, $\delta_{c3}^{(1)}$. We consider the case with identical typical distribution. We set π to be the uniform distribution with alphabet size 10, and generate outlier distributions randomly.

We first simulate the case with distinct outliers where T is known. We choose $M = 20$, $T = 3$. In Fig. 1, we plot $\log P_e$ as a function of n for δ_{GL}, δ_{c2} and $\delta_{c2}^{(1)}$ after 5000 times of Monte Carlo simulations. As we can see from Fig. 1, $\delta_{c2}^{(1)}$ and δ_{c2} are both exponentially consistent, and δ_{c2} outperforms $\delta_{c2}^{(1)}$ as shown in Theorem 2. Comparison between δ_{c2} and δ_{GL} shows that without sacrificing much in performance, δ_{c2} is about 50 times faster than δ_{GL}.

We then simulate the case with unknown number of identical outliers. We set $M = 100$, $T = 10$. Fig. 2 shows that δ_{c3} outperforms $\delta_{c3}^{(1)}$. We note that running the clustering based tests for 5000 times takes 5 minutes on a 3.6 GHz i7 CPU. However, the GL test is not feasible here, since the number of hypotheses needs to search over is exponential in M.

We have also simulated the case where both the typical and outlier distributions form clusters. These results, which are not reported here due to space limitations, show that the test $\delta_{c3}^{(1)}$ is also exponentially consistent.

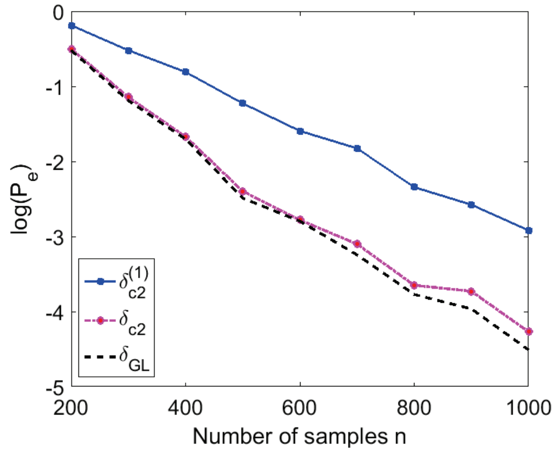

Fig. 1. Comparison of test $\delta_{c2}^{(1)}$, δ_{c2}, δ_{GL}, known number of distinct outliers

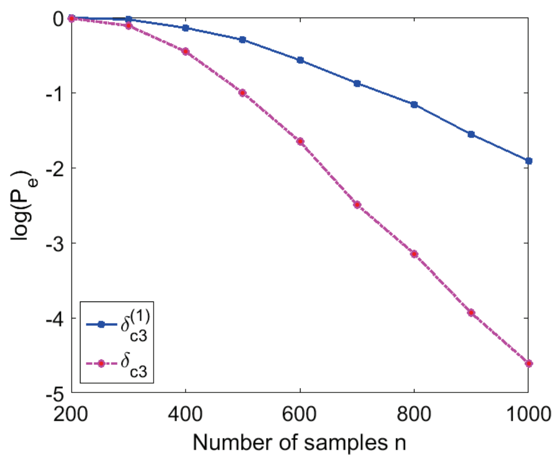

Fig. 2. Comparison of test δ_{c3}, $\delta_{c3}^{(1)}$, unknown number of identical outliers

References

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

[2] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical science*, pp. 235–249, 2002.

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

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

[5] Y. Bu, S. Zou, Y. Liang, and V. V Veeravalli, "Universal outlying sequence detection for continuous observations," in *Proc. IEEE In. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4254–4258.

[6] A. Banerjee, S. Merugu, I. S Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of machine learning research*, vol. 6, no. Oct, 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 In. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6180–6184.

[8] K. Chaudhuri and A. McGregor, "Finding metric structure in information theoretic clustering.," in *COLT*. Citeseer, 2008, vol. 8, p. 10.

[9] M. R Ackermann, J. Blömer, and C. Sohler, "Clustering for metric and nonmetric distance measures," *ACM Transactions on Algorithms (TALG)*, vol. 6, no. 4, pp. 59, 2010.

[10] R. Nock, P. Luosto, and J. Kivinen, "Mixed bregman clustering with approximation guarantees," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 154–169.

[11] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[12] M. Blum, R. W Floyd, V. Pratt, R. L Rivest, and R. E Tarjan, "Time bounds for selection," *Journal of computer and system sciences*, vol. 7, no. 4, pp. 448–461, 1973.

[13] Y. Bu, S. Zou, and V. V. Veeravalli, "Linear complexity exponentially consistent tests for outlying sequence detection," *available at https://arxiv.org/abs/1701.06084*, 2017.