UNIVERSAL OUTLYING SEQUENCE DETECTION FOR CONTINUOUS OBSERVATIONS

Yuheng Bu[⋆] *Shaofeng Zou* [†] *Yingbin Liang* [†] *Venugopal V. Veeravalli* [⋆]

[⋆] University of Illinois at Urbana-Champaign [†] Syracuse University

Email: bu3@illinois.edu, szou02@syr.edu, yliang06@syr.edu, vvv@illinois.edu

ABSTRACT

The following detection problem is studied, in which there are M sequences of samples out of which one outlier sequence needs to be detected. Each typical sequence contains n independent and identically distributed (i.i.d.) *continuous* observations from a known distribution π, and the outlier sequence contains n i.i.d. observations from an outlier distribution μ, which is distinct from π, but otherwise unknown. A universal test based on Kullback-Leibler (KL) divergence is built to approximate the maximum likelihood test, with known π and unknown μ. A KL divergence estimator based on data-dependent partitions is employed, and is shown to converge to its true value exponentially fast when the density ratio satisfies $0 < K_1 \leq \frac{d\mu}{d\pi} \leq K_2$, where K_1 and K_2 are positive constants. The performance of such a KL divergence estimator further implies that the outlier detection test is exponentially consistent. The detection performance of the KL divergence based test is compared with that of a recently introduced test for this problem based on the machine learning approach of maximum mean discrepancy (MMD). Regimes in which the KL divergence based test is better than the MMD based test are identified.

Index Terms— Kullback-Leibler divergence, maximum mean discrepancy, outlier hypothesis testing, universal exponential consistency

1. INTRODUCTION

In this paper, we study an outlying sequence detection problem, in which there are M sequences of samples out of which one outlier sequence needs to be detected. Each typical sequence consists of n independent and identically distributed (i.i.d.) *continuous* observations drawn from a *known* distribution π, whereas the outlier sequence consists of n i.i.d. samples drawn from a distribution μ, which is distinct from π, but otherwise *unknown*. The goal is to design a test to detect the outlier sequence.

The work of Y. Bu and V. V. Veeravalli was supported by the Air Force Office of Scientific Research (AFOSR) under the Grant FA9550-10-1-0458, and by the National Science Foundation under Grant NSF 11-11342, through the University of Illinois at Urbana-Champaign. The work of S. Zou and Y. Liang was supported by an NSF CAREER Award under Grant CCF-10-26565.

The study of such a model is useful in many applications [1]. For example, in cognitive wireless networks, signals follow different distributions depending on whether the channel is busy or vacant. The goal in such a network is to identify vacant channels out of busy channels based on their corresponding signals in order to utilize the vacant channels for improving spectral efficiency. Such a problem was studied in [2] and [3] under the assumption that both μ and π are known. Other applications include anomaly detection in large data sets [4,5], event detection and environment monitoring in sensor networks [6], understanding of visual search in humans and animals [7], and optimal search and target tracking [8].

The outlying sequence detection problem with *discrete* μ and π was studied in [9] in which both distributions were assumed to be unknown. A universal test based on generalized likelihood ratio test was proposed, and was shown to be exponentially consistent. The error exponent was further shown to be optimal as the number of sequences goes to infinity. The test utilizes empirical distributions to estimate μ and π, and is therefore applicable only for the case where μ and π are discrete.

In this paper, we study the case where distributions μ and π are *continuous* and μ is *unknown*. We construct a Kullback-Leibler (KL) divergence based test, and further show that this test is *exponentially consistent*.

Our exploration of the problem is inspired by the case in which both μ and π are known, and the maximum likelihood test is optimal. An interesting observation is that the test statistic of the optimal test converges to $D(\mu\|\pi)$ as the sample size goes to infinity if the sequence is the outlier. This motivates the use of a KL divergence estimator to approximate the test statistic for the case when μ is unknown. We apply a divergence estimator based on the idea of data-dependent partitions [10], which was shown to be consistent. Our first contribution here is to show that such an estimator converges *exponentially* fast to its true value when the density ratio satisfies the boundedness condition: $0 < K_1 \leq \frac{d\mu}{d\pi} \leq K_2$, where K_1 and K_2 are positive constants. We then design an outlying sequence detection test using such an estimator of the KL divergence, and further show that the resulting test is *exponentially* consistent.

The rest of the paper is organized as follows. In Section 2, we describe the problem formulation. In Section 3, we

present the KL divergence based test and establish its exponential consistency. In Section 4, we review the maximum mean discrepancy (MMD) based test. In Section 5, we provide a numerical comparison of our KL divergence based test and the MMD based test. We omit detailed proofs in this paper due to space limitations. The full proofs can be found in [11].

2. PROBLEM MODEL

Throughout the paper, random variables are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All logarithms are with respect to the natural base.

We study an outlier detection problem, in which there are in total M data sequences denoted by $Y^{(i)}$ for $1 \leq i \leq M$. Each data sequence $Y^{(i)}$ consists of n i.i.d. samples $Y_1^{(i)}, \ldots, Y_n^{(i)}$ drawn from either a typical distribution π or an outlier distribution μ, where π and μ are *continuous*, i.e., defined on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, and $\mu \neq \pi$. We use the notation $\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_n^{(i)})$, where $y_k^{(i)} \in \mathbb{R}$ denotes the k-th observation of the i-th sequence. We assume that there is exactly one outlier among M sequences. If the i-th sequence is the outlier, the joint distribution of all observations is given by

$$p_i(y^{Mn}) = p_i(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}) = \prod_{k=1}^{n} \left\{ \mu(y_k^{(i)}) \prod_{j \neq i} \pi(y_k^{(j)}) \right\}.$$

We are interested in the scenario in which the outlier distributions μ is unknown a priori, but the typical distribution π is known exactly. This is reasonable because in many practical scenarios, systems typically start without outliers and it is not difficult to collect sufficient information about π.

Our goal is to build a distribution-free test to detect the outlier sequence generated by μ. The the test can be captured by a universal rule $\delta : \pi \times \mathbb{R}^{Mn} \to 1, \ldots, M$, which must not depend on μ.

The maximum error probability, which is a function of the detector and (μ, π), is defined as

$$e(\delta, \pi, \mu) \triangleq \max_{i=1,\ldots,M} \int_{y^{Mn}: \delta(\pi, y^{Mn}) \neq i} p_i(y^{Mn}) dy^{Mn},$$

and the corresponding error exponent is defined as

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

A test is said to be *universally consistent* if

$$\lim_{n \to \infty} e(\delta, \pi, \mu) = 0,$$

for any $\mu \neq \pi$. It is said to be *universally exponentially consistent* if

$$\lim_{n \to \infty} \alpha(\delta, \pi, \mu) > 0,$$

for any $\mu \neq \pi$.

3. KL DIVERGENCE BASED TEST

We first introduce the optimal test when both μ and π are known, which is the maximum likelihood test. We then construct a KL divergence estimator, and prove its exponential consistency. Next, we employ the KL divergence estimator to approximate the test statistics of the optimal test for the outlying sequence detection problem, and construct the KL divergence based test.

3.1. Optimal test with π and μ known

If both μ and π are known, the optimal test for the outlying sequence detection problem is the maximum likelihood test:

$$\delta_{\mathrm{ML}}(y^{Mn}, \pi, \mu) = \arg\max_{1 \leq i \leq M} \; p_i(y^{Mn}). \tag{1}$$

By normalizing $p_i(y^{Mn})$ with $\pi(y^{Mn})$, (1) is equivalent to:

$$\begin{aligned} \delta_{\mathrm{ML}}(y^{Mn}, \pi, \mu) &= \arg\max_{1 \leq i \leq M} \; \frac{p_i(y^{Mn})}{\pi(y^{Mn})} \\ &= \arg\max_{1 \leq i \leq M} \left\{ \frac{1}{n} \sum_{k=1}^{n} \log \frac{\mu(y_k^{(i)})}{\pi(y_k^{(i)})} \right\} \\ &= \arg\max_{1 \leq i \leq M} \; L_i. \end{aligned}$$

where

$$L_i \triangleq \frac{1}{n} \sum_{k=1}^{n} \log \frac{\mu(y_k^{(i)})}{\pi(y_k^{(i)})}. \tag{2}$$

The following theorem characterizes the error exponent of test (1).

Theorem 1. *[9, Theorem 1] Consider the outlying sequence detection problem with both μ and π known. The error exponent for the maximum likelihood test* (1) *is given by*

$$\alpha(\delta_{\mathrm{ML}}, \pi, \mu) = 2B(\pi, \mu),$$

where $B(\pi, \mu)$ is the Bhattacharyya distance between μ and π which is defined as

$$B(\pi, \mu) \triangleq -\log \left(\int \mu(y)^{\frac{1}{2}} \pi(y)^{\frac{1}{2}} dy \right).$$

Consider L_i defined in (2). If $\mathbf{y}^{(i)}$ is generated from μ, L_i is an empirical estimate of the KL divergence between μ and π, then $L_i \to D(\mu||\pi)$ almost surely as $n \to \infty$, by the law of large numbers. Here,

$$D(\mu||\pi) \triangleq \int d\mu \log \frac{d\mu}{d\pi}$$

is the KL divergence between μ and π. Similarly, if $\mathbf{y}^{(i)}$ is generated from π, $L_i \to -D(\pi||\mu)$ almost surely as $n \to \infty$. These observations motivate us to construct a generalized likelihood test based on an estimator of the KL divergence between μ and π, if μ is unknown.

3.2. KL divergence estimator

We propose to use a KL divergence estimator based on data-dependent partitions [10].

Assume that the distribution p is *unknown* and the distribution q is known, and both p and q are continuous. A sequence of i.i.d. samples $Y \in \mathbb{R}^n$ is generated from p. We wish to estimate the KL divergence between p and q. We denote the order statistics of Y by $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$ where $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. We further partition the real line into empirically equiprobable segments as follows:

$$\{I_t^n\}_{t=1,\ldots,T_n} = \{(-\infty, Y_{(\ell_n)}], (Y_{(\ell_n)}, Y_{(2\ell_n)}], \ldots, (Y_{(\ell_n(T_n-1))}, \infty)\},$$

where $\ell_n \in \mathbb{N} \leq n$ is the number of points in each interval except possibly the last one, and $T_n = \lfloor n/\ell_n \rfloor$ is the number of intervals. A divergence estimator between the sequence $Y \in \mathbb{R}^n$ and the distribution π was proposed in [10], which is given by

$$\hat{D}_n(Y||q) = \sum_{t=1}^{T_n-1} \frac{\ell_n}{n} \log \frac{\ell_n/n}{q(I_t^n)} + \frac{\epsilon_n}{n} \log \frac{\epsilon_n/n}{q(I_{T_n}^n)}, \quad (3)$$

where $\epsilon_n = (n - \ell_n(T_n - 1))$ is the number of points in the last segment.

The consistency of such an estimator was shown in [10]. Here, we further characterize the convergence rate by introducing the following boundedness condition on the density ratio between p and q, i.e.,

$$0 < K_1 \leq \frac{\mathrm{d}p}{\mathrm{d}q} \leq K_2, \quad (4)$$

where K_1 and K_2 are positive constants. In practice, such a boundedness condition is often satisfied, for example, for truncated Gaussian distributions.

The following theorem characterizes a lower bound on the convergence rate of estimator (3).

Theorem 2. *If the density ratio between p and q satisfies* (4)*, and estimator* (3) *is applied with $T_n, \ell_n \to \infty$, as $n \to \infty$, then for $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} -\frac{1}{n} \log \left(\mathbb{P}\left\{ \left| \hat{D}_n(Y||q) - D(p||q) \right| > \epsilon \right\} \right) \geq \frac{1}{32} \frac{K_1^2}{K_2^2} \epsilon^2.$$

Proof. See [11]. □

Remark 1. *The convergence rate of estimator* (3) *in Theorem 2 is equivalent to*

$$\left| \hat{D}_n(Y||q) - D(p||q) \right| = \mathcal{O}_p(n^{-1/2}),^1$$

where $\mathcal{O}_p$ denotes "bounded in probability " [12].

[1] $X_n = \mathcal{O}_p(a_n)$: $\forall \epsilon > 0, \exists M > 0, P(|\frac{X_n}{a_n}| > M) < \epsilon, \forall n$.

3.3. Test and performance

In this subsection, we utilize the estimator based on data-dependent partitions to construct our test.

It is clear that if $Y^{(i)}$ is the outlier, then $\hat{D}_n(Y^{(i)}||\pi)$ is a good estimator of $D(\mu||\pi)$, which is a positive constant. On the other hand, if $Y^{(i)}$ is a typical sequence, $\hat{D}_n(Y^{(i)}||\pi)$ should be close to $D(\pi||\pi) = 0$. Based on this understanding and the convergence guarantee in Theorem 2, we use $\hat{D}_n(Y^{(i)}||\pi)$ in place of L_i in (2), and construct the following test for the outlying sequence detection problem:

$$\delta_{\mathrm{KL}}(y^{Mn}) = \arg \max_{1\leq i\leq M} \hat{D}_n(Y^{(i)}||\pi). \quad (5)$$

The following theorem provides a lower bound on the error exponent of δ_{KL}, which further implies that δ_{KL} is universally exponentially consistent.

Theorem 3. *If the density ratio between μ and π satisfies* (4)*, then δ_{KL} defined in* (5) *is exponentially consistent, and the error exponent is lower bounded as follows,*

$$\alpha(\delta_{\mathrm{KL}}, \pi, \mu) \geq \frac{1}{32} \left(\frac{K_1}{K_1 + K_2} \right)^2 D^2(\mu||\pi). \quad (6)$$

Proof. See [11]. □

4. MMD-BASED TEST

In this section, we introduce the MMD based test, which we previously studied in [13]. We will compare δ_{KL} to the MMD based test.

4.1. Introduction to MMD

In this subsection, we briefly introduce the idea of mean embedding of distributions into RKHS [14] and the metric of MMD. Suppose $\mathcal{P}$ is a set of probability distributions, and suppose $\mathcal{H}$ is the RKHS with an associated kernel $k(\cdot, \cdot)$. We define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped to an element in $\mathcal{H}$ as follows

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot, x)] = \int k(\cdot, x) dp(x).$$

Here, $\mu_p(\cdot)$ is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$. Due to the reproducing property of $\mathcal{H}$, it is clear that $\mathbb{E}_p[f] = \langle \mu_p, f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

In order to distinguish between two distributions p and q, Gretton *et al.* [15] introduced the following quantity of maximum mean discrepancy (MMD) based on the mean embeddings μ_p and μ_q of p and q in RKHS:

$$\mathrm{MMD}[p, q] := \|\mu_p - \mu_q\|_{\mathcal{H}}.$$

It can be shown that

$$\mathrm{MMD}[p, q] = \sup_{\|f\|_{\mathcal{H}} \leq 1} \mathbb{E}_p[f] - \mathbb{E}_q[f].$$

Due to the reproducing property of kernel, we have

$$\mathrm{MMD}^2[p,q] = \mathbb{E}[k(X,X')] - 2\mathbb{E}[k(X,Y)] + \mathbb{E}[k(Y,Y')],$$

where X and X' are independent but have the same distribution p, and Y and Y' are independent but have the same distribution q. An unbiased estimator of $\mathrm{MMD}^2[p,q]$ based on q and n samples of $X = \{x_1, x_2, \ldots, x_n\}$ generated from p is given as follows,

$$\mathrm{MMD}_u^2[X,q] = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(x_i, x_j) + \mathbb{E}[k(Y,Y')] - \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}[k(x_i, Y)],$$

where Y and Y' are independent but have the same distribution q.

4.2. Test and performance

For each sequence $Y^{(i)}$, we compute $\mathrm{MMD}_u^2[Y^{(i)}, \pi]$ for $1 \leq i \leq M$. It is clear that if $Y^{(i)}$ is the outlier, $\mathrm{MMD}_u^2[Y^{(i)}, \pi]$ is a good estimator of $\mathrm{MMD}^2[\mu, \pi]$, which is a positive constant. On the other hand, if $Y^{(i)}$ is a typical sequence, $\mathrm{MMD}_u^2[Y^{(i)}, \pi]$ should be a good estimator of $\mathrm{MMD}^2[\pi, \pi]$, which is zero. Based on the above understanding, we construct the following test:

$$\delta_{\mathrm{MMD}} = \arg\max_{1 \leq i \leq M} \mathrm{MMD}_u^2[Y^{(i)}, \pi]. \tag{7}$$

The following theorem provides a lower bound on the error exponent of δ_{MMD}, and further demonstrates that the test δ_{MMD} is universally exponentially consistent.

Theorem 4. *Consider the universal outlying sequence detection problem. Suppose δ_{MMD} defined in* (7) *applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then, the error exponent is lower bounded as follows,*

$$\alpha(\delta_{MMD}, \mu, \pi) \geq \frac{\mathrm{MMD}^4[\mu, \pi]}{9K^2}. \tag{8}$$

Proof. See [11]. □

5. NUMERICAL RESULTS AND DISCUSSION

In this section, we compare the performance of δ_{KL} and δ_{MMD}. We set the number of sequences $M = 5$. We choose the typical distribution $\pi = \mathcal{N}(0,1)$, and choose the outlier distribution $\mu = \mathcal{N}(0,0.2), \mathcal{N}(0,1.2), \mathcal{N}(0,1.8), \mathcal{N}(0,2.0)$, respectively. In Fig. 1, Fig. 2, Fig. 3 and Fig. 4, we plot the logarithm $\log P_e$ of the probability of error as a function of the sample size n.

It can be seen that for both tests as the number of samples increases, the probability of error converges to zero as the sample size increases. Furthermore, $\log P_e$ decreases with n linearly, which demonstrates the exponential consistency of both δ_{KL} and δ_{MMD}. Note that the pair of distributions in Fig. 2 are the closest to each other, which results in a larger probability of error than in the other three cases.

By comparing the four figures, it can be seen that as the variance of μ deviates from the variance of π, δ_{KL} outperforms δ_{MMD}. The numerical results and theoretical lower bounds on error exponents give us some intuitions to identify regimes in which one test outperforms the other. As shown above, when the distribution μ and π become more different from each other, δ_{KL} outperforms δ_{MMD}. The reason is that for any pair of distributions, MMD is bounded between $[0, 2K]$, while the KL divergence is not bounded. As the distributions become more different from each other, the KL divergence increases, and the KL divergence based test thus has a larger error exponent than the MMD based test.

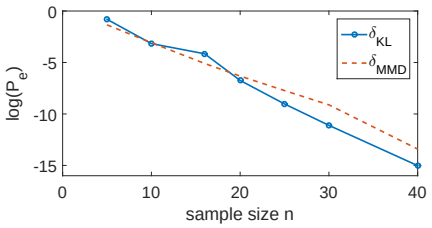

Fig. 1. Comparison of the performance between KL divergence and MMD based tests with $\pi = \mathcal{N}(0,1)$ and $\mu = \mathcal{N}(0,0.2)$

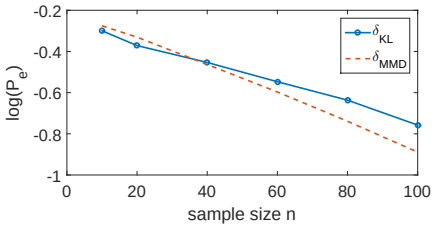

Fig. 2. Comparison of the performance between KL divergence and MMD based tests with $\pi = \mathcal{N}(0,1)$ and $\mu = \mathcal{N}(0,1.2)$

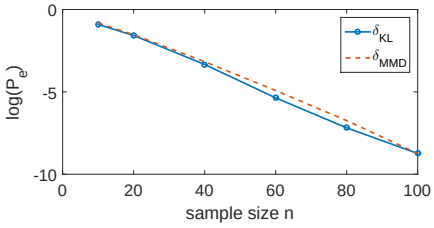

Fig. 3. Comparison of the performance between KL divergence and MMD based tests with $\pi = \mathcal{N}(0,1)$ and $\mu = \mathcal{N}(0,1.8)$

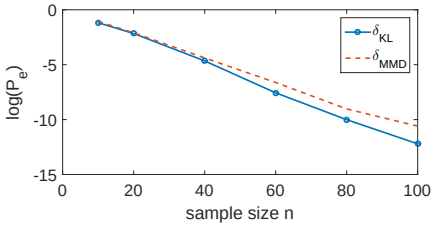

Fig. 4. Comparison of the performance between KL divergence and MMD based tests with $\pi = \mathcal{N}(0,1)$ and $\mu = \mathcal{N}(0,2)$

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

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

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

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, July 2009.

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

[7] N. K. Vaidhiyan, S. P. Arun, and R. Sundaresan, "Active sequential hypothesis testing with application to a visual search problem," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2012, pp. 2201–2205.

[8] L. D. Stone, "Theory of optimal search," Topics in Operations Research Series, INFORMS, 2004.

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

[10] Q. Wang, S. R Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inform. Theory*, vol. 51, no. 9, pp. 3064–3074, 2005.

[11] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Universal outlying sequence detection for continuous observations," *http://arxiv.org/abs/1509.07040*.

[12] X. Nguyen, M. J Wainwright, M. Jordan, et al., "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[13] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Unsupervised nonparametric anomaly detection: A kernel method," in *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014, pp. 836–841.

[14] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, 2010.

[15] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.