Unsupervised Nonparametric Anomaly Detection: A Kernel Method

Shaofeng Zou[1] Yingbin Liang[1] H. Vincent Poor[2] Xinghua Shi[3]

***Abstract*— An anomaly detection problem is investigated, in which s out of n sequences are anomalous and need to be detected. Each sequence consists of m independent and identically distributed (i.i.d.) samples drawn either from a normal distribution p or from an anomalous distribution q that is distinct from p. Neither p nor q is unknown a priori. Two scenarios respectively with s known and unknown are studied. Distribution-free tests are constructed based on the metric of the maximum mean discrepancy (MMD). It is shown that if the value of s is known, as n goes to infinity, the number m of samples in each sequence should be of order $\mathcal{O}(\log n)$ or larger to guarantee that the constructed test is exponentially consistent. Conversely, under the special case that $s = 1$, it is shown that if $m \leq \mathcal{O}(\log n)$, then there exists $p \neq q$ for which no test guarantees consistent test. These two bounds are within a constant factor, which means the tests are near-optimal. It is also shown that if the value of s is unknown, the number of samples m in each sequence should be of the order strictly greater than $\mathcal{O}(\log n)$ to guarantee consistent tests. The computational complexity of all tests are shown to be polynomial. Numerical results are provided to demonstrate the performance of the tests. Comparisons with other approaches on both synthetic data sets and real data sets show that MMD-based tests outperform or perform as well as other approaches.**

I. Introduction

In this paper, we study an anomalous sequences detection problem (see Figure **??**). In this problem, there are s anomalous sequences out of n sequences to be detected. There are m independent and identically distributed (i.i.d.) samples drawn from a distribution p in each normal sequence. Whereas there are m i.i.d. samples drawn from a distribution q in each anomalous sequence. The distributions p and q are assumed to be distinct and unknown a priori. We aim to build distribution-free tests to detect the s anomalous sequences generated by q out of all sequences.

We note that the anomaly detection studied here is different from the usual anomaly or outlier detection problem in machine learning [**?**], [**?**]. In the usual anomaly or outlier detection problem, the anomaly is in the sense of one or multiple realizations of a certain process which are quite different from normal realizations, whereas in our problem, the anomaly is in the sense of s sequences of observations from an anomalous distribution q. The parametric case of such a problem has been well studied in [**?**]. In [**?**], the distribution p and q are known a priori which could be exploited to design the tests. However the nonparametric model in which p and q are unknown and arbitrary is less studied. Recently, Li, Nitinarat and Veeravalli proposed a nonparametric divergence-based generalized likelihood tests in [**?**], and characterized the error decay exponents in the asymptotic regime. In [**?**], the non-asymptotic regime (finite samples in each sequence) for such a problem is studied. However, the tests in [**?**], [**?**] are limited to the case that distributions p and q are discrete since their tests utilized empirical probability mass function of p and q from samples.

In this paper, we study the nonparametric model that distributions p and q are arbitrary, unknown and can be continuous. Compared to previous nonparametric studies [**?**], [**?**] which focus on discrete distributions, there are a few challenges to solve the problem for arbitrary continuous distributions: (1) it is difficult to accurately estimate continuous distributions from samples for further anomaly detection; (2) it is difficult to design tests for continuous distributions with low computational complexity; (3) distribution-free consistent tests (and further exponentially consistent tests) are challenging to build for arbitrary unknown distributions.

Such a problem has a great potential for practical applications. For example, for a group of people with one certain genetic disease, there might be a few genes related to this genetic disease whose expression levels follow different distributions comparing to those genes not related to this disease. A major issue is to identify those related genes out of a large number of genes based on there expression level. Another potential application is in the cognitive wireless networks, signals follow different distributions depending on whether the channels are busy or not. The major task is to identify the vacant channels such that users can transmit over those vacant channels to improve the spectral efficiency. There are also many other applications, e.g., detecting virus infected computers from other virus free computers, detecting slightly modified images from other untouched images.

There are a number of statistical approaches that can be applied to solving this problem, e.g. FR-Smirnov test [**?**], t-test, FR-Wolf test [**?**] and Hall test [**?**]. The FR-Smirnov test is to first estimate the probability distribution functions based on data samples, and then measure the difference of the estimated distributions for anomaly detection. For the continuous distributions, it requires a large number of samples to accurately estimate the probability distribution functions. Furthermore, the propagation of the error of estimating the distributions to anomaly detection is unpredictable. For the t-test, FR-wolf test and Hall test, although they do not

*The work of S. Zou and Y. Liang was supported by a NSF CAREER Award under Grant CCF-10-26565 and by the NSF under Grant CNS-11-16932. The work of H. V. Poor was supported in part by the NSF under Grants DMS-1118605 and ECCS-1343210.

[1] Shaofeng Zou and Yingbin Liang are with the department of Electrical Engineering and Computer Science, Syracuse University, USA {szou02,yliang06}@syr.edu

[2] H. V. Poor is with the Department of Electrical Engineering, Princeton University, USA poor@princeton.edu

[3] Xinghua Shi is with department of Bioinformatics and Genomics, University of North Carolina at Charlotte, xshi3@uncc.edul

need to estimate the probability distribution function as intermediate steps, t-test only works well for distributions with mean or variance difference, FR-wolf test does not perform well for some arbitrary distributions, and Hall test has a large computational complexity. Recently, the kernel-based approaches such as kernel density ratio (KDR) test [**?**] and kernel Fisher discriminant analysis (KFDA) test [**?**] have been developed to estimate certain distance metrics between two distributions. For the sake of comparison, we develop/implement tests based on the approaches mentioned above, although those approaches were not designed to solve our anomaly detection problem previously. We demonstrate that our tests outperform or equal the tests under various testing cases.

In this paper, our approach utilizes an emerging kernel-based approach which is based on mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [**?**], [**?**]. The idea is to map probability distributions into a RKHS such that the images capture the information of all moments of the distributions. Since this mapping is injective [**?**], [**?**], [**?**], [**?**], distinguishing the two probabilities can be carried out by distinguishing the corresponding mean embeddings in the RKHS. The mean embedding of distributions can be compared easily by evaluating their distances in the RKHS. This distance metric is defined as the *maximum mean discrepancy (MMD)* [**?**], [**?**]. There are a few advantages of this MMD metric: (1) it is computational efficient to estimate MMD from samples; (2) MMD-based approach do not need to estimate probability density functions as intermediate steps, hence can avoid error propagation. Our anomaly detection tests are constructed utilizing the metric MMD.

In this paper, we mainly focus on the asymptotic regime but in a different way compared to [**?**] in which n is fixed and m goes to infinity. In this paper, the total number n of data sequences goes to infinity. This is a natural assumption in the era of big data. It becomes more challenging as the number n of sequences increases (and possibly the number s of anomalous sequences also increases). It then requires that the number m of samples in each sequence increase to balance out the error from the increase of n and s. We want to know that how m should scale with (n, s) to guarantee consistent tests as n goes to infinity. In this paper, we focus on the unsupervised case in which no information about p and q is available, which extends the semi-supervised nonparametric study in [**?**], in which a reference sequence from p is available.

In summary, our main contributions in this paper are as follows. (1) We construct computational efficient distribution-free MMD-based tests for two scenarios (the number s of anomalous sequences known and unknown). (2) We characterize how the number m of samples in each sequence should scale with (n, s) to guarantee consistency of the tests. We show that m can be much smaller than n (i.e., in the order of $\mathcal{O}(\log n)$ if s is known, in the order slightly greater than $\mathcal{O}(\log n)$ if s is unknown). Therefore, lack of the knowledge of s results in an order level increase of m for consistency of the tests. (3) For the special case $s = 1$, we derive the necessary condition on m, i.e., for $m \leq \mathcal{O}(\log n)$, there exists p and q that no test is consistent. Hence, the lower and upper bounds on m are within a constant factor, which means our test are near-optimal under the case $s = 1$. (4) We provide extensive numerical results. We provide numerical results on synthetic data to demonstrate our theoretical assertions. We further compare our MMD-based approach with nonparametric generalized likelihood tests in [**?**] on discrete distributions. Finally, we develop/implement traditional statistical approaches together with our tests to a real data set to show the consistency of our tests and comparison with other approaches. We note that in this paper, we omit the proofs. The details can be found in [**?**].

II. Problem Statement and Preliminaries on MMD

A. Problem Statement

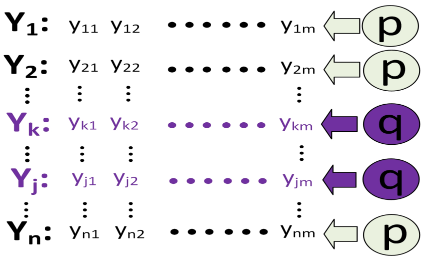

Fig. 1. An anomaly detection model with data sequences from distribution p and anomalous distribution q.

In this paper, we study an anomaly detection problem (see Figure **??**). There are n data sequences denoted by Y_k for $1 \leq k \leq n$. There are m i.i.d. samples $y_{k1}, \ldots, y_{km}$ drawn from either a distribution p or an anomalous distribution q in each data sequence Y_k. We assume that p, q are arbitrary unknown a priori and $p \neq q$. We use the notation $Y_k := (y_{k1}, \ldots, y_{km})$. We aim to build distribution-free tests to detect data sequences generated by the anomalous distribution q.

We assume that s out of n data sequences are anomalous, i.e., are generated by the anomalous distribution q. We study both cases with the value of s known and unknown a priori, respectively. We are interested in the asymptotical regime, in which the number n of data sequences goes to infinity. We assume that the number s of anomalous sequences satisfies $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha < \frac{1}{2}$. Symmetrically, the case that $\frac{1}{2} < \alpha \leq 1$ is included. The test for unknown s and corresponding analysis are also applicable to the case that $s = 0$, i.e., the null hypothesis in which there is no anomalous sequence. We will comment on this when the corresponding results are presented. In this paper, $f(n) = \mathcal{O}(g(n))$ denotes $f(n)/g(n)$ converges to a constant as $n \to \infty$.

In contrast to [**?**], we focus on the unsupervised case, in which no reference sequence from p is available.

The performance criteria of the tests are defined as follows. We use $\mathcal{I}$ to denote the index set of all anomalous data sequences.

***Definition** 1:* A sequence of tests are said to be consistent if

$$\lim_{n\to\infty} P_e = \lim_{n\to\infty} \max_{|\mathcal{I}|=s} P\{\hat{\mathcal{I}}^n \neq \mathcal{I}|\mathcal{I}\} = 0. \tag{1}$$

We note the limit is taken with respect to n instead of m. However, as n increases, it becomes more difficult to detect the anomalous sequences with asymptotically small probability of error. Therefore, m should also increase with n to provide more information about p and q, and to balance out the increase of probability of error brought by the increase of n. For convenience, we take the limit with respect to n, but it is equivalent to taking the limit with respect to m.

Furthermore, for a consistent test, we are interested in whether the error probability decays exponentially fast with respect to the number m of samples.

***Definition** 2:* A sequence of tests are said to be exponentially consistent if

$$\liminf_{m\to\infty} -\frac{1}{m}\log P_e > 0. \tag{2}$$

In this paper, we aim to build distribution-free tests, characterize how m should scale with (n, s) to guarantee consistency (or exponentially consistency), and study the necessary condition on m for a consistent test for such a problem.

B. Introduction of MMD

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

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

$\mu_p(\cdot)$ is referred to as the *mean embedding* of p into $\mathcal{H}$.

It is desirable that such a mapping is *injective*, such that the problem of distinguishing two distributions can be solved by distinguishing the mean embedding of the two distributions. It has been shown in [**?**], [**?**], [**?**], [**?**] that for many RKHSs such as those associated with Gaussian and Laplace kernels, the mean embedding is injective. And in [**?**], they introduced the metric of MMD which is the distance between the mean embeddings μ_p and μ_q of p and q:

$$\mathrm{MMD}[p,q] := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{3}$$

By the reproducing property of kernel, it is clear that

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

where x and x' have independent but the same distribution p, and y and y' have independent but the same distribution q. Naturally, an unbiased estimator of $\mathrm{MMD}^2[p,q]$ based on l_1 samples of X and l_2 samples of Y is given as follows,

$$\mathrm{MMD}_u^2[X,Y] = \frac{1}{l_1(l_1-1)}\sum_{i=1}^{l_1}\sum_{j\neq i}^{l_1} k(x_i,x_j) + \frac{1}{l_2(l_2-1)}\sum_{i=1}^{l_2}\sum_{j\neq i}^{l_2} k(y_i,y_j) - \frac{2}{l_1 l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2} k(x_i,y_j). \tag{4}$$

III. Main Results

In this section, we start with the case that s is known. We use a a simple case with $s=1$ to introduce the idea of our tests, and then study the more general case for arbitrary s, in which $\frac{s}{n}\to\alpha$ as $n\to\infty$, where $0\le\alpha\le\frac{1}{2}$. And then we further study the case when s is unknown.

When s is known, we first study the case with $s=1$. For each sequence Y_k, we use $\overline{Y}_k$ to denote the $(n-1)m$ dimensional sequence that stacks all other sequences together, as given by

$$\overline{Y}_k = \{Y_1,\dots,Y_{k-1},Y_{k+1},\dots,Y_n\}.$$

We then compute $\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]$ for $1\le k\le n$. It is clear that if Y_k is an anomalous sequence, then $\overline{Y}_k$ is fully composed of sequences from p. Hence, $\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]$ is a good estimator of $\mathrm{MMD}^2[p,q]$, which is a positive constant. Nonetheless, if Y_k is a sequence from p, $\overline{Y}_k$ is composed of $n-2$ sequences generated by p and only one sequence generated by q. As n increases, the impact of the anomalous sequence on $\overline{Y}_k$ is negligible, and $\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]$ should be close to zero. Based on the above understanding, we construct the following test when $s=1$. The sequence k^* is the index of the anomalous data sequence if

$$k^* = \arg\max_{1\le k\le n} \mathrm{MMD}_u^2[Y_k,\overline{Y}_k]. \tag{5}$$

The following theorem characterizes the condition under which the above test is consistent.

***Theorem** 1:* Consider the anomaly detection model with one anomalous sequence. Suppose the test (**??**) applies a bounded kernel with $0\le k(x,y)\le K$ for any (x,y). Then, the test (**??**) is consistent if

$$m \ge \frac{16K^2(1+\eta)}{\mathrm{MMD}^4[p,q]}\log n, \tag{6}$$

where η is any positive constant. Furthermore, under the above condition, the test (**??**) is also exponentially consistent.

We now consider the more general case in which $\frac{s}{n}\to\alpha$ as $n\to\infty$, where $0\le\alpha<\frac{1}{2}$, and the value of s is known. By symmetry, the case with $\alpha>\frac{1}{2}$ is included by swapping the role of p and q. Our test is a natural generalization of the test (**??**) except now the test picks the sequences with the largest s values of $\mathrm{MMD}_u^2[Y_k,\overline{Y}_k]$, which is given by

$$\begin{aligned}\hat{\mathcal{I}} = \{k : &\ \mathrm{MMD}_u^2[Y_k,\overline{Y}_k] \text{ is among the } s \text{ largest}\\ &\text{values of } \mathrm{MMD}_u^2[Y_i,\overline{Y}_i] \text{ for } i=1,\dots,n\}.\end{aligned} \tag{7}$$

The following theorem characterizes the condition under which the above test is consistent.

***Theorem** 2:* Consider the anomaly detection model with s anomalous sequences, where $\frac{s}{n} \to \alpha$ as $n \to \infty$ and $0 \leq \alpha < \frac{1}{2}$. Assume the value of s is known. Further assume that the test (**??**) applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test (**??**) is consistent if

$$m \geq \frac{16K^2(1+\eta)}{(1-2\alpha)^2 \text{MMD}^4[p,q]} \log(s(n-s)), \qquad (8)$$

where η is any positive constant. Furthermore, under the above condition, the test (**??**) is also exponentially consistent. The computational complexity of the test (**??**) is $\mathcal{O}(n^3m^2)$.

Since $s \leq \mathcal{O}(n)$, we have $\log s(n-s) \sim \mathcal{O}(\log n)$. Hence, Theorem **??** and Theorem **??** imply that, our MMD based tests (**??**) and (**??**) require only $\mathcal{O}(\log n)$ samples in each data sequence in order to guarantee consistency of the tests.

***Remark** 1:* For the case with $\frac{s}{n} \to 0$, as $n \to \infty$, we can build a test that has reduced computational complexity. For each Y_k, instead of using $n-1$ sequences to build $\overline{Y}_k$ as in the test (**??**), we take any l sequences out of the remaining $n-1$ sequences to build a sequence $\widetilde{Y}_k$, such that $\frac{l}{n} \to 0$ and $\frac{s}{l} \to 0$ as $n \to \infty$. Such an l exists for any s and n satisfying $\frac{s}{n} \to 0$ (e.g., $l = \sqrt{sn}$). It can be shown that using $\widetilde{Y}_k$ to replace $\overline{Y}_k$ in the test (**??**) still leads to consistent detection under the same condition given in Theorem **??**. The computational complexity of such a test becomes $\mathcal{O}(nl^2m^2)$, which is substantially smaller than $\mathcal{O}(n^3m^2)$ of the test (**??**), considering that l is less than n in the order sense.

We further consider the case, in which the value of s is unknown a priori. For this case, the previous test (**??**) is not applicable due the the lack of knowledge of s. We observe that for large value of m, $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ should be close to 0 if Y_k is drawn from p, and should be close to $\text{MMD}^2[p,q]$ if Y_k is drawn from q. Based on this understanding, we build the following test:

$$\widehat{\mathcal{I}} = \{k : \text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\} \qquad (9)$$

where $\delta_n \to 0$ and $\frac{s^2}{n^2\delta_n} \to 0$ as $n \to \infty$. We note that the above requirements on δ_n implies that the test (**??**) is applicable only when $\frac{s}{n} \to 0$ as $n \to \infty$. This includes two cases: (1) s is fixed and (2) $s \to \infty$ and $\frac{s}{n} \to 0$ as $n \to \infty$. Furthermore, the scaling behavior of s as n increases needs to be known in order to pick δ_n for the test. This is reasonable to assume because mostly in practice the scale of anomalous data points can be estimated based on domain knowledge.

The following theorem characterizes the condition under which the test (**??**) is consistent.

***Theorem** 3:* Consider the anomaly detection model with s anomalous sequences, where $\lim_{n\to\infty} \frac{s}{n} = 0$. We assume that the value of s is unknown a priori. Further assume that the test (**??**) adopts a threshold δ_n such that $\delta_n \to 0$ and $\frac{s^2}{n^2\delta_n} \to 0$, as $n \to \infty$, and the test applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test (**??**) is consistent if

$$m \geq 16(1+\eta)K^2 \max\left\{ \frac{\log(\max\{1,s\})}{(\text{MMD}^2[p,q]-\delta_n)^2}, \frac{\log(n-s)}{(\delta_n - \mathbb{E}\big[\text{MMD}_u^2[Y,\overline{Y}]\big])^2} \right\}, \qquad (10)$$

where η is any positive constant, and $E[\text{MMD}_u^2[Y,\overline{Y}]]$ is a constant, where Y is a sequence generated by p and $\overline{Y}$ is a stack of $(n-1)$ sequences with s sequences generated by q and the remaining sequences generated by p. The computational complexity of the test (**??**) is $\mathcal{O}(n^3m^2)$.

We note that Theorem **??** is also applicable to the case with $s = 0$, i.e., the null hypothesis when there is no anomalous sequence. We further note that the test (**??**) is not exponentially consistent. If there is no null hypothesis (i.e., $s > 0$ and unknown), an exponentially consistent test can be built as follows. For each subset $\mathcal{S}$ of $\{1,\dots,n\}$ we compute the average $\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \text{MMD}_u^2[Y_k, \overline{Y}_k]$, and the test finds the set of indices corresponding to the largest average value. However, m need to scale linearly with n for the test to be consistent, and the computational complexity is exponential with n, which is not desirable.

Theorem **??** implies that the threshold on m to guarantee consistent detection has an order strictly larger than $\mathcal{O}(\log n)$, because $\frac{s}{n} \to 0$ and $\delta_n \to 0$ as $n \to \infty$. This is the price paid due to not knowing s.

Next, we show the necessary condition for a consistent test under the special case $s = 1$.

***Theorem** 4:* Consider the anomaly detection problem with one anomalous sequence, if

$$m \leq \mathcal{O}(\log n), \qquad (11)$$

there exists $p \neq q$ for which there is no consistent test.

Notice that under the case $s = 1$, we have the sufficient condition for a consistent test from our MMD-based test which is $m \geq \mathcal{O}(\log n)$. If we compare the sufficient condition with the necessary condition, this two bounds are within a constant factor. In this sense, our MMD-based test are near-optimal.

***Corollary** 1:* Consider the anomaly detection problem with one anomalous sequence, the MMD-based test (**??**) is near-optimal.

IV. Numerical Results

In this section, we provide numerical results to demonstrate our theoretical assertions, and compare our MMD-based tests with a number of tests based on other approaches on both synthetic and real data sets.

A. Demonstration of Theorems

We choose the distribution p to be Gaussian with mean zero and variance one, and choose the anomalous distribution q to be Laplace distribution with mean one and variance one. We use the Gaussian kernel $k(x,x') = \exp(-\frac{|x-x'|^2}{2\sigma^2})$ with $\sigma = 1$. We choose $s = 1$. We run the test for cases with the numbers of sequences being $n = 40, 100$, respectively. In Figure **??**, we plot the probability of error as a function of

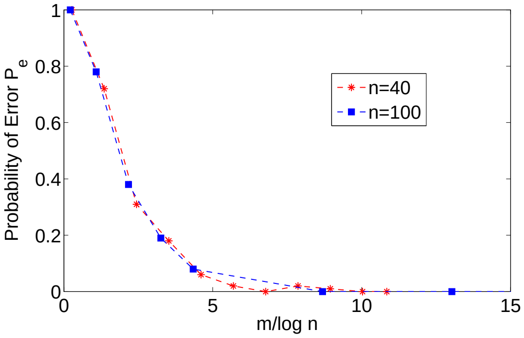

Fig. 2. Performance of the test with $s = 1$.

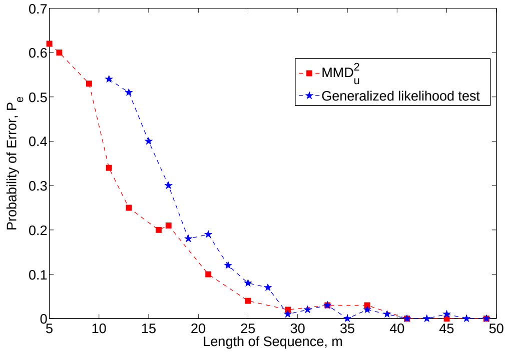

Fig. 3. Comparison of the MMD-based test with divergence-based generalized likelihood test.

$\frac{m}{\log n}$. It can be seen that, as m increases, the probability of error converges to zero. In particular, both curves drop to zero almost at the same threshold, which agrees with Theorem **??**.

B. Comparison with Other Tests

In this subsection, we compare our MMD-based tests with tests based on other nonparametric approaches. We first compare our test with the divergence-based generalized likelihood approach developed in [**?**]. Since the test in [**?**] is applicable only when the distributions p and q are discrete and have finite alphabets, we set the distributions p and q to be binary with p having probability 0.3 to take "0" (and probability 0.7 to take "1"), and q having probability 0.7 to take "0" (and probability 0.3 to take "1"). We let $s = 1$ and assume that s is known. We let $n = 50$.

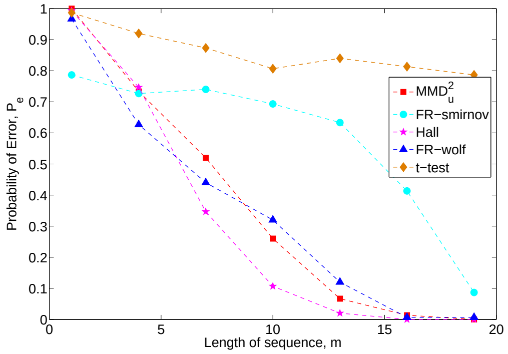

Fig. 4. Comparison of the MMD-based test with four other tests on a real data set.

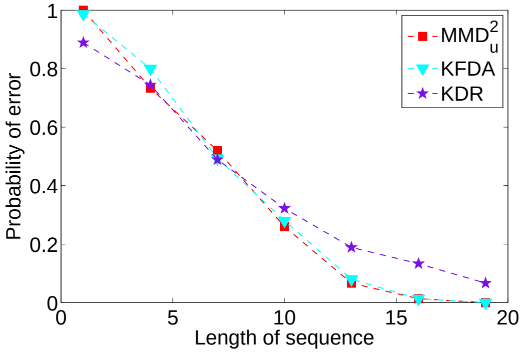

Fig. 5. Comparison of the MMD-based test with two other kernel-based tests on a real data set.

In Figure **??**, we plot the probability of error as a function m. It can be seen that the MMD-based test has a slightly better performance than the divergence-based generalized likelihood test in both cases. We note that it has been shown in [**?**] that the divergence-based test has optimal convergence rate in the limiting case when n is infinite, which suggests that such a test should also perform well for the case with finite n. Thus, the comparison demonstrates that the MMD-based test can provide comparable or even better performance than the well-performed divergence-based test.

C. Application to Real Data Set

In this subsection, we study how the MMD-based test performs on a real data set. We choose the collection of daily maximum temperature of Syracuse (New York, USA) in July from 1993 to 2012 as the normal data sequences, and the collection of daily maximum temperature of Makapulapai (Hawaii, USA) in May from 1993 to 2012 as anomalous sequences. Here, each data sequence contains daily maximum temperatures of a certain day across twenty years from 1993 to 2012. In our experiment, the data set contains 32 sequences in total, including one temperature sequence of Hawaii and 31 sequences of Syracuse. The probability of error is averaged over all cases with each using one sequence of Hawaii as the anomalous sequence. Although it seems easy to detect the sequence of Hawaii out of the sequences of Syracuse, the temperatures we compare for the two places are in May for Hawaii and July for Syracuse, during which the two places have approximately the same mean in temperature. In this way, it may not be easy to detect the anomalous sequence (in fact, some tests do not perform well as shown in Figure **??**).

We apply the MMD-based test and compare its performance with t-test, FR-Wolf test, FR-Smirnov test, and Hall test. For the MMD-based test, we use the Gaussian kernel with $\sigma = 1$. In Figure **??**, we plot the probability of error as a function of m for all tests. It can be seen that the MMD-based test, Hall test, and FR-wolf test have the best performances, and all of the three tests are consistent with the probability of error converging to zero as m increases. Furthermore, comparing to Hall and FR-wolf tests, the MMD-based test has the lowest computational complexity.

We also compare the performance of MMD-based test with the kernel-based tests KFDA and KDR. We use Gaussian kernel with $\sigma = 1$. In Figure **??**, we plot the probability of error as a function of m. It can be seen that all tests are consistent with the probability of error converging to zero as m increases, and the MMD-based test has the best performance among the three tests.

V. Conclusion

In this paper, we have studied a nonparametric anomaly detection problem, in which s anomalous sequences need to be detected out of n sequences. Each data sequence contains m i.i.d. samples drawn from distribution p (normal sequences) or q (anomalous sequences). We have build MMD-based distribution-free tests and characterized how m should scale with n to guarantee consistent (or exponentially consistent) tests for both the case with known s and the case with unknown s. If s is known, we have shown that m should scale with the order of $\mathcal{O}(\log n)$. If s is unknown, we have shown that m should scale with the order of slightly greater than $\mathcal{O}(\log n)$. Conversely, we have also shown that under the special case with only one anomalous sequence, if $m \leq \mathcal{O}(\log n)$, there exists p and q that there is no consistent test. The lower and upper bounds on m are within a constant factor, which shows the near-optimality of our MMD-based tests. Furthermore, we have demonstrated our theoretic results by numerical results on synthetic data and have demonstrated the performance of our tests by comparing it to other appealing tests on a real data set. Our study of this problem demonstrates a useful application of the mean embedding of distributions and MMD, and we believe that such an approach can be applied to solving various other nonparametric problems.