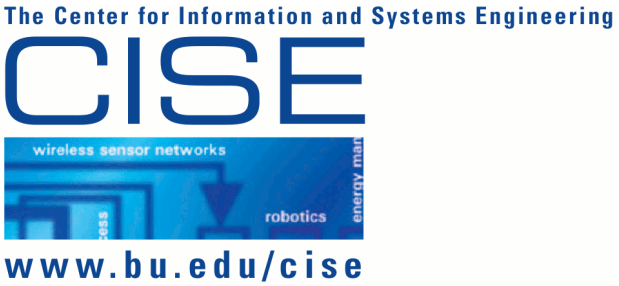

BOSTON UNIVERSITY

Research Publications

"Local Anomaly Detection"

V. Saligrama and M. Zhou

CISE Technical Report

CISE Technical Report #2011-IR-0014

CISE Technical Report Date: August 31, 2011

Center for Information and Systems Engineering
15 St. Mary's Street, Brookline, MA 02446

617.358.1295

Local Anomaly Detection

Venkatesh Saligrama
Boston University
Boston, MA 02215

Manqi Zhao
Google Inc.
Mountain View, CA

Abstract

Anomalies with spatial and temporal stamps arise in a number of applications including communication networks, traffic monitoring and video analysis. In these applications anomalies are temporally or spatially localized but otherwise unknown. We propose a novel graph-based statistical notion that unifies the idea of temporal and spatial locality. This notion lends itself to an elegant characterization of optimal decision rules and in turn suggests corresponding empirical rules based on local nearest neighbor distances. We compute a single composite score for the entire spatio-temporal data sample based on the local neighborhood distances. We declare data samples as containing local anomalies based on the composite score. We show that such rules not only asymptotically guarantee desired false alarm control but are also asymptotically optimal. We also show that our composite scoring scheme overcomes the inherent resolution issues of alternative multi-comparison approaches that are based on fusing the outcomes of location-by-location comparisons. We then verify our algorithms on synthetic and real data sets.

1 Introduction

We are motivated by local anomalies that arise in several applications that deal with spatial and temporal data. These applications include network anomaly [1], traffic monitoring [2], sensor network intrusion [3] and epidemics [4]. For instance, consider network anomaly detection. Here each data sample corresponds to a time-trace of packet counts from origin-to-destination over the course of a day. Different days correspond to different data samples. A local anomaly such as an outage, port-scan or a network-scan could occur anytime over a period of a 10-30 minutes. In this paper we formalize a statistical non-parametric notion of local anomalies to deal with such scenarios and develop algorithms with proven statistical guarantees. Our algorithms are based on local K-nearest neighbor distances, which is amenable to hashing [5].

Data driven approaches for detection of *localized* temporal anomalies has been described in a number of papers in the statistics and data mining. The focus in data mining is on algorithms and do not provide statistical guarantees[6, 7, 1]. Much of the focus in statistics literature is on parametric models, where the models are customized to specific applications. Nominal parametric models are first estimated from data and then anomalies are detected whenever a temporal data sequence deviates from that predicted by the nominal model. Nominal temporal data is usually modeled as a moving average or other autoregressive models [8, 9, 10, 11]. Anomalies are detected when a target data sequence deviates from the model.

A well studied problem related to this paper is the detection of *sparse* heterogenous mixtures for Gaussian processes [12, 13, 14, 15]. The nominal data is a realization of a zero mean, unit variance IID process. The anomaly has an unknown mean shift in a sparse set of components. These papers attempt to characterize fundamental tradeoff between sparsity and mean shift to asymptotically guarantee detection.

Our approach is also related to a number of other non-parametric data-driven approaches such as [16, 17, 3, 18] with important differences. Existing statistical approaches do not account for local anomalies, i.e., anomalies that are localized to a small time interval or spatial region. We rigorously define a statistical notion for local anomaly structure in Sec. 2. This notion of locality lends itself to optimal Bayesian and Neyman-Pearson characterization (see Sec. 3). While

Boston University Technical Report

this characterization requires knowledge of underlying likelihood models, it nevertheless motivates consideration of specific local statistics such as local K-nearest neighbor distances on the raw data (Sec. 4). In this sense our paper is closely related to the ranking method of [18], where scores for each data sample is computed by ranking global K-nearest neighbor distances. In contrast we develop several rules (entropy, sum and max statistics) that pool together all the local K-nearest neighbor distances and yet provide statistical guarantees. In Sec. 5 we not only establish asymptotic optimality of such rules in a number of interesting cases but also characterize the improvement over global anomaly methods such as [18].

In Sec. 5.1 we also describe the benefits of our approach over other approaches encountered in the context of multi-comparison tests. Specifically, we consider Bonferroni and the recently developed "Higher Criticism" as candidate approaches, where a score for each location is first estimated and then anomaly is detected by fusing these local scores. We notice that these approaches suffer from poor resolution in the low false alarm regime. This arises when the number of samples scale in proportion to dimension of each data sample. Finally we verify our theoretical analysis on both synthetic and artificial data set in Sec. 6. Proofs of our results appear in the supplementary section.

2 Local Anomaly Model

We consider discrete-valued data with each data sample $x = (x_v)_{v\in V}$ corresponding to a collection of random variables, which are indexed on a graph $G = (V, E)$. Note that x_v can be a random vector without loss of generality. The set V is endowed with the usual graph metric $d(u, v)$ defined for any two nodes v and u. We often use T to denote $|V|$. Our setup accounts for temporal (V is time), spatial (V indexes spatial locations) and spatio-temporal data.

We assume that baseline data $x = (x_v)_{v\in V}$ is drawn from the null hypothesis H_0:

$$H_0 : x \sim f_0(x) \tag{1}$$

We describe the anomalous distribution as a mixture of location-specific anomalous likelihood models. Let $f_v(x)$, P_v be the likelihood function and prior probability associated with location v respectively. Then,

$$H_1 : x \sim \sum_{v\in V} P_v f_v(x) \tag{2}$$

We next introduce notation to describe our local model. Let $\omega_v(s)$ be a ball of radius s around v:

$$\omega_v \triangleq \omega_v(s) = \{v' : d(v, v') \le s\}$$

With abuse of notation ω_v will generally refer to a ball of a fixed radius s at node v. The marginal distribution of f_0, f_v on a subset set $\omega \subset V$ is denoted as $f_0(x_\omega)$. We also denote by $\omega_{v,\epsilon}$ as the set that includes all points within an ϵ radius of ω_v, i.e.,

$$\omega_{v,\epsilon} = \{u \in V \mid d(u, v) \le \epsilon,\ v \in \omega_v\}$$

Definition 1. We say an anomaly is of *local structure* if the distributions f_0 and f_v satisfy the following Markovian and Mask assumptions.
(1) Markov Assumption: We say f_0 and f_v's satisfy the observation x forms a Markov random field. Specifically we assume that there is an ϵ-neighborhood such that x_v, $v \in \omega_v$ is conditionally independent of x_u, $u \notin \omega_{v,\epsilon}$ when conditioned on the annulus $\omega_{v,\epsilon} \cap \omega_v^c$.

(2) Mask Assumption: If the anomaly event happens at region ω_v, the marginal distribution of f_0 and f_v on ω_v^c is identical: $f_0(x_{\omega_v^c}) = f_v(x_{\omega_v^c})$.

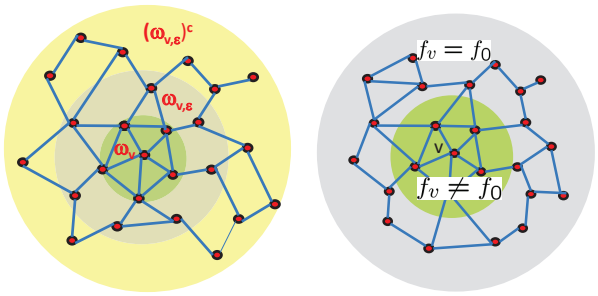

Figure 1: Illustration of Markov and Mask Properties. Markov implies random variables in region ω_v are independent of random variables in ω_v^c when conditioned on the annulus. Mask assumption means that the anomalous density and nominal density are identical outside ω_v.

Note that our assumptions are structural and we are still left with a non-parametric model for f_0 nor f_v. The local anomaly event can happen in any neighborhood ω_v, which is centered at node v and the range of its influence is dictated by radius s.

We present some of the limitations and extensions of our framework that are not explicitly addressed here:
(a) Unknown radius: While we assumed an upper bound on s here, we will later account for unknown s in our algorithm and simulations.
(**b**) **Complex Shaped Anomalies**: We have assumed that there is a sufficiently large ball encapsulating the anomaly.
(c) Latent Models: Our model does not appear to account for latent parameters. For example, consider traffic data where nominal weekend pattern can be different from weekday pattern. We can account for this phenomena by suitably extending our model to include a hidden state. The hidden state dictates whether the

time-series is from weekends or weekdays. The nominal pattern then corresponds to a mixture model and the Markovian assumption holds conditionally, i.e., it is Markovian when conditioned on the latent variable. We will present simulations for this setting in Sec. 6.
(d) Multiple Anomalies: While we primarily describe single anomalies, our method extends to multiple local independent anomalies as well. To see this we can let, P_{uv} be probability of joint occurrence of anomalies at well-separated locations at two nodes u, v. Then the term $P_v f_v(x)$ can be modified to $P_{uv} f_{uv}(x)$ and can be further factorized by appealing to the Markovian property. Our theoretical results extend to this situation. Evidently [12] single anomalies, due to the inherent sparsity, are often more difficult to detect in comparison to multiple anomalies.

2.1 Relevant Applications

Both the Markovian and Mask assumption are quite general and meaningful. We need the Markovian property to ensure local dependence otherwise the effect of what happens locally would propagate globally and the effect is no-longer local. The second assumption says that no information about anomaly can be revealed from the data extracted outside of the anomalous region. We have already described communication network anomalies [1] as an instance of local anomalies in Sec. 1. We further justify the Mask and Markovian assumptions further by describing other applications.

Detection in Sensor network. A wireless sensor network is deployed to detect intrusions in an area based on the field measurements [3]. The nodes of the graph G are sensors and their associated locations. The edges connect neighboring sensors. Sensors measure the received signal strength at various sensor locations. Intrusion is a localized phenomena since the sensor range is local and the intrusion itself occurs at a specific location and at a specific time instant. The receive signal strength decays as r^{α} with radius r. Note that Mask assumption holds because the intrusion is undetectable from sensor measurements that are not in the immediate neighborhood of the sensing radius. Markovianity holds because the sensor measurements are independent under the null hypothesis.

Bio-surveillance of epidemics. In modern bio-surveillance [4], the problem is the early detection of a disease epidemic in a certain region. During the early stages, the disease is contained in a small local region and the data can be mapped to a general Markov random field in which each node represents a certain neighborhood. The mask assumption is satisfied since regions outside the local anomalous region do not reveal information about the disease.

Detection in Traffic data. Each data sample consists of vehicle counts $x(t)$ at time t [2] over the course of a day. Markovianity in $x(t)$ follows if $x(t)$ follows a poisson counting process, a typical assumption made for this scenario. Anomaly can happen at any time window $[i-s, i+s]$. Note that in most cases the anomaly is local in time because the events (such as baseball game, car accident etc.) will only impact a certain period of time during the day. During the rest of the day, the time-series looks exactly like a normal day satisfying Mask property.

Surveillance. A camera records video footage in an urban scenario. The dataset can be broken into small video segments. Different video segments can be assumed to be statistically similar. Anomalies are usually motion anomalies such as dropped baggage, unusual direction of motion etc. These anomalies are local in space and time (see Fig. 2).

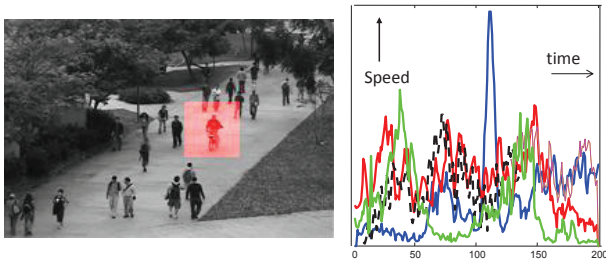

Figure 2: Illustration of local anomaly in time. Left: Illustrates frame of a video segment [19] with anomaly (bicycle). Right: Plot of speed (computed from optical flow) vs. time for nominal and anomalous video segments. Note that speed trace outside the anomalous window looks identical to nominal traces verifying Mask and Markov properties.

3 Problem Formulation and Optimal Characterization

An anomaly detector is a decision rule, π, that maps observations $x = (x)_{v \in V}$ to $\{0, 1\}$ with zero denoting no anomaly and one denoting an anomaly. Let $\Omega_\pi = \{x \mid \pi(x) = 1\}$. The optimal anomaly detector π minimizes the "Bayesian" Neyman-Pearson objective function.

$$\text{Bayesian:} \quad \max_{\pi} \int_{\Omega_\pi} \sum_{v \in V} P_v f_v(x) dx \qquad (3)$$

$$\text{subject to}$$

$$P_F \triangleq \int_{\Omega_\pi} f_0(x) dx \leq \alpha$$

The optimal decision rule can be characterized as

$$\sum_v P_v \mathcal{L}_v \underset{nominal}{\overset{anomaly}{\gtrless}} \xi \tag{4}$$

where the likelihood ratio function $\mathcal{L}_v$ is defined as $\mathcal{L}_v = f_v(x)/f_0(x)$ and ξ is chosen such that the false alarm probability is smaller than α. Lemma 1 (see Supplementary Section for the proof) shows that the likelihood ratio function $\mathcal{L}_v$ simplifies under our assumptions of Definition 1.

Lemma 1. *Let ω_v be a ball around v and $\omega_{v,\epsilon}$ be the ϵ-neighborhood set such that the Markovian assumption of Definition 1 is satisfied. Then we have,*

$$\mathcal{L}_v(x) = \frac{f_v\left(x_{\omega_{v,\epsilon}}\right)}{f_0\left(x_{\omega_{v,\epsilon}}\right)} \tag{5}$$

Several issues arises in applying this decision rule. Both P_v and the likelihood model f_v are unknown and we only have nominal training data. A uniform prior ($P_v = 1/|V|$) or a worst case prior are options for dealing with unknown P_v. The worst-case prior also has a well-known Bayesian interpretation and provides a link between the robust composite testing and optimal Bayes decision theory. Under symmetrizing location invariance [20] assumptions it turns out that the uniform prior is also the worst-case prior. The issue of unknown f_v is an important aspect in anomaly detection. Typically, the volume of Ω^c under the Lebesgue measure is used as a proxy for the missed detection rate. This requires that Ω^c be bounded and this is guaranteed if $f_0(\cdot)$ has bounded support. This issue is further complicated in our setting since $\mathcal{L}_v(x)$ is location dependent and so support of f_0 varies with location. So if at location v, $x_{\omega_{v,\epsilon}}$ lies in a set of diameter λ_v and $\omega_{v,\epsilon}$ has diameter $s+\epsilon$ Equation 5 reduces to:

$$\mathcal{L}_v(x) = \frac{\lambda_v^{-(s+\epsilon)}}{f_0\left(x_{\omega_{v,\epsilon}}\right)} \tag{6}$$

3.1 Composite Scores with Guarantees

While Equation 4 characterizes the optimal decision rule, it is unclear how to choose a threshold to ensure false alarm control. To this end we let $G(x)$ be a real-valued statistic of the raw data. Consider the score function:

$$R(\eta) = \mathbb{P}_{x\sim f_0}\left(x : G(x) \geq G(\eta)\right) \tag{7}$$

It is easy to show that this score function is distributed uniformly for a large class of statistics $G(x)$. This includes:

(1) NP detector: $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$.
(2) GLRT [21]: $G_{MAX}(x) = \max_v \mathcal{L}_v(x)$.
(3) Entropy: $G_{ENT}(x) = -\sum_v \log(\mathcal{L}_v(x))$.

Lemma 2. *Suppose statistics $G(x)$ has the nestedness property, that is, for any $t_1 > t_2$ we have $\{x : G(x) > t_1\} \subset \{x : G(x) > t_2\}$. Then $R(\eta)$ is uniformly distributed in $[0,1]$ when $\eta \sim f_0$.*

This lemma implies that we can control false alarms via thresholding the statistic $R(\eta)$.

Theorem 3. *If G satisfies the nestedness property, by setting the detection rule as $R(\eta) \leq \alpha$, we control the FA at level α. Furthermore, if $R(\eta)$ is computed with $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$, then it is optimal solution to Equation 3 for the uniform prior.*

4 LCS based on Nearest-Neighbors

The goal in this section is to empirically approximate $R(\cdot)$ given training data $(x^{(1)}, \cdots, x^{(n)})$, a test point η and a statistic $G(\cdot)$. Consider the empirical score function:

$$R_n(\eta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{G_n(x^{(i)}) \geq G_n(\eta)\}} \tag{8}$$

where G_n is a finite sample approximation of G and $\mathbb{I}_{\{\cdot\}}$ is the indicator function. Here we propose local nearest neighbor based statistics. We denote it as a local neighborhood based composite scores (LCS). This is because $G_n(\cdot)$ as described in the previous section combines statistics over local neighborhoods of a data sample and the ranking function produces a composite score for an entire time series or random field.

Definition 2. We define the d-statistic $d_{\omega_{v,\epsilon}}(\eta)$ for window $\omega_{v,\epsilon}$ at an arbitrary point η as the distance of $\eta_{\omega_{v,\epsilon}}$ to its k-th closest point in $\left(x^{(1)}_{\omega_{v,\epsilon}}, \cdots, x^{(n)}_{\omega_{v,\epsilon}}\right)$.

We generally choose Euclidean distance for computing the distances. In general, we can apply any distance metric customized to specific application. To approximate $G(x)$ for different cases we need to determine the support parameter λ_v. To this end we let $d^{(j)}_{\omega_{v,\epsilon}}$ as the ordered the distances of $d_{\omega_{v,\epsilon}}(x^{(j)})$ $(j = 1, 2, \cdots, n)$ in decreasing order and we approximate the support as an ξ percentile:

$$\lambda_v = d^{(\lfloor n\psi \rfloor)}_{\omega_{v,\epsilon}} \tag{9}$$

where $\lfloor n\psi \rfloor$ denotes the integer part of real number and can be tuned (in simulations we usually use the median). Now $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$ can be approximated by SUM LCS, $G_{n,SUM}$:

$$G_{n,SUM} = \sum_v \left(\frac{d_{\omega_{v,\epsilon}}(\eta)}{\lambda_v}\right)^s \tag{10}$$

Similarly, we can take a max statistic to obtain MAX LCS:

$$G_{n,MAX} = \max_v \frac{d_{\omega_{v,\epsilon}}(\eta)}{\lambda_v} \tag{11}$$

Observe that when s is equal to dimension of x the two statistics max and sum coincide. This resulting statistics is identical to the K-nearest-neighbor ranking (KNN-Ranking) scheme of [18].

Practical Issues with $G_{SUM}(\cdot)$: Recall from Section 3, $G_{SUM}(x)$, appears to be optimal for uniform priors and minimax optimal under symmetrizing assumptions. However, it is difficult to reliably approximate $G_{SUM}(x)$ for several reasons. (1) Sum is no longer optimal if the prior is not uniform. (2) Errors can accumulate for the summation but max is relatively robust. (3) The additional s exponent term in the expression of SUM LCS (which compensates for the dimension) leads to sensitivity to parameters such as λ_v. (4) For large values of s, since max distance is a dominant term in $G_{SUM}(x)$ the theoretical difference between the two statistics maybe negligible. Therefore, we pursue MAX-LCS in this paper and prove some its properties in the next section.

Algorithm: For concreteness we list the steps below.
(Input) Nominal Data, $(x^{(1)}, \cdots, x^{(n)})$, test sample, η, false alarm rate, α and maximum anomaly size, s, if known else select initial size.
(Step 1) Compute k nearest distances $d_{\omega_{v,\epsilon}}(\cdot)$ for each v and for all the nominal data and test sample.
(Step 2) Compute MAX-LCS using Eq. 11 for each nominal sample, (i) and test sample, η.
(Step 3) Compute the rank for η using Eq. 8. Declare anomaly if rank below threshold α and window size s is known. If s unknown iterate over geometrically increasing window sizes. Declare anomaly if the minimum rank falls below $\alpha/\#\{\text{windows}\}$.

5 Theoretical Properties of Composite Scores

First, we show that $G_{n,MAX}$ asymptotically converges to the true statistic G_{MAX}. This which implies that the score function Equation 7 converges to uniform distribution, and thus establishes false alarm control for thresholding MAX-LCS.

Theorem 4. *Assume smoothness condition on f_0 as in [18] and the regularity condition $k \to \infty, n \to \infty$ and $k/n \to 0$, we have,*

$$R_{n,MAX}(\eta) \overset{n\to\infty}{\longrightarrow} R_{MAX}(\eta)$$
$$= \mathcal{P}_{x\sim f_0}\left(\max_v \mathcal{L}_v(x) \geq \max_v \mathcal{L}_v(\eta)\right)$$

A similar property holds for SUM LCS. Consequently, SUM LCS scheme is asymptotically optimal.

The proof of the asymptotic optimality of SUM LCS follows by combining the convergence result above with Theorem 3. However, we do not pursue SUM LCS due to the serious practical issues described in Sec. 4.

Corollary 5. *Given the same technical conditions as Theorem 4, the MAX-LCS score $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{G_{n,MAX}(x^{(i)}) \geq G_{n,MAX}(\eta)\}}$ with $G_{n,MAX} = \max_v d_{\omega_{v,\epsilon}}(\eta)/\lambda_v$ asymptotically converges to uniform distribution.*

The above corollary tells us that MAX-LCS asymptotically controls false alarm rate. Next we derive the properties of MAX-LCS to deduce its detection power. First, we consider a simple case when f_0 is Gaussian which leads to a precise sharp characterization.

Theorem 6. *Assume the normal pattern is IID standard Gaussian $\mathcal{N}(0,1)$ and the ground truth anomaly pattern is:*

$$H_1 : \begin{cases} x_i \sim \mathcal{N}(\mu_i, 1) & i = 1, \cdots, s \\ x_i \sim \mathcal{N}(0,1) & i = s+1, \cdots, T. \end{cases}$$

and this structure of anomaly is not revealed. Consider the regime when both s and T go to infinity and s/T goes to zero. It follows that for miss probability P_M of KNN-Ranking scheme to vanish, we need $\sum_{i=1}^{s} \mu_i^2$ to scale as $\Theta\left(\sqrt{T}\right)$. In comparison for the MAX-LCS we need $\sum_{i=1}^{s} \mu_i^2$ to scale as $\Theta\left(\sqrt{s\log(T/s)}\right)$.

This result establishes a fundamental dichotomy, namely, for values of s, i.e., the radius of local anomaly such that $s = o(T/\log(T))$, locally based MAX-LCS is provably superior to KNN-Ranking of [18].

Finally, given further assumptions on the d-statistics, we can prove a more general result about the detection power of MAX-LCS. We need two assumptions about d-statistics of normal points and anomalous point.

- For the normal training points, $d_{\omega_{v,\epsilon}}(x^{(j)})$ decays exponentially $\Pr\left(d_{\omega_{v,\epsilon}}(x^{(j)}) > d\right) \leq e^{-\lambda d}$. This is true when f_0 is sub-gaussian [22].
- For the anomalous test point η, the measure of $d_{\omega_{v,\epsilon}}(\eta)$ is not concentrated around zero: $\Pr\left(d_{\omega_{v,\epsilon}}(\eta) < \log(T)\frac{\rho}{\lambda}\right) \leq \epsilon$.

Theorem 7. *Consider the case when the nodes $(x_v)_{v\in V}$ for the nominal data are independent. Given the above two assumptions, the expectation of MAX-LCS (over the training set and the test point) is upper bounded by $\frac{1}{2}\epsilon + \frac{1}{2}(1 - e^{-1/\rho})$.*

Theorem 7 can be extended in several cases where the nodes $(x_v)_{v\in V}$ are correlated. Specifically, for so called negative associations (see [23] and Chapter 3 of [24]) it follows that,

$$\Pr\left(d_{\omega_{v,\epsilon}}(x^{(j)}) < d, \forall i\right) \geq \prod_v \Pr\left(d_{\omega_{v,\epsilon}}(x^{(j)}) < d\right) \quad (12)$$

and in these cases Theorem 7 continues to hold. In other cases such as in Markovian settings a similar bound can be derived if the sub-gaussianity of certain conditional distributions is assumed. In these cases Theorem 7 points to the fact that the average missed detection rate of MAX-LCS is small. This result in combination with Theorem 5 provides a sharp characterization of false alarm and missed detection rates.

5.1 LCS vs. Multiple Comparisons

Multiple comparison(MC) arises whenever two groups (such as a control group & treatment group) are compared over many attributes. In our context each attribute can be identified with a local neighborhood and two groups (nominal and test sample) can be compared for each local neighborhood. A score or significance value can be assigned for each local neighborhood. These local neighborhood scores can be fused to determine the presence or absence of anomaly. It has long been recognized that MC [25, 12] leads to poor false alarm control unless one *corrects* for multiple comparisons. Traditionally, much of the MC literature [12, 26, 13, 15] has focused on Gaussian setting and so the issue of empirical computation of local scores does not arise and the p-value for each location is computed. In our context local p-values can be computed by the KNN-ranking scheme [18] localized to a specific local neighborhood $\omega_{v,\epsilon}$. These p-values can then be fused using methods developed in the MC literature. For the purpose of comparison we describe two well-known methods:

Bonferroni Detector: Here we take a minimum of the p-values (estimated) or scores corresponding to each window [25] and declare anomaly if it is below a Bonferroni corrected threshold $\alpha/\#\{windows\}$.
Higher Criticism (HC): The *higher criticism* statistic [12] computes the following quantity: $HC = \max_{1\le i\le \alpha_0 l} \sqrt{l}[i/l - \hat{p}_{(i)}]/\sqrt{\hat{p}_{(i)}(1-\hat{p}_{(i)})}$. It has been argued that this detector is asymptotically optimal for sparse heterogenous mixture of Gaussian processes [12].

Resolution Issues: The main problem is that when empirical estimation of p-values is combined with fusion of p-values, resolution issues arise whenever the data dimension, T, scales in proportion to the number of data samples, n. To see this note that p-values at each location cannot be expected to be smaller than $1/n$ since we only have n data samples. Now if we were to apply the Bonferroni correction and we wish to control false alarms at level α, we need to test each location against a corrected threshold α/T. This quantity makes sense (for $s = 1$) when $T/n \le \alpha$ which requires that the number of samples grow faster than the data dimension. HC does not necessarily suffer from this problem. Nevertheless, it is known [12] that in the limit when the anomalous locations are sparse both methods, HC and Bonferroni, result in similar performance. Note that in contrast to the above fusion strategies our MAX-LCS algorithm first pools together information from all of the windows (via $G(x)$ statistics) and then computes a composite statistic for each sample. A direct impact of this type of processing is that we can control false alarms at any desired level.

6 Simulation

In this section, we test our MAX LCS algorithm on both synthetic and real data set on time-series data of length T, size n and anomaly radius s. The computational cost of MAX LCS is linear in dimension, linear in the total number of local regions and quadratic in the data size. Standard techniques such as k-d trees and locality sensitive hashing [5] can be used to speed up the computation but we do not pursue this aspect here. We apply the algorithm described in Sec. 4 for known and unknown window sizes.

6.1 Synthetic data

We experiment with synthetic data because ground truth information is necessary to compute false positives (FP) and true positives (TP). For time-series data set, the ground truth is usually not available. Therefore, synthetic data is more appropriate for accurate comparison in terms of FP and TP.

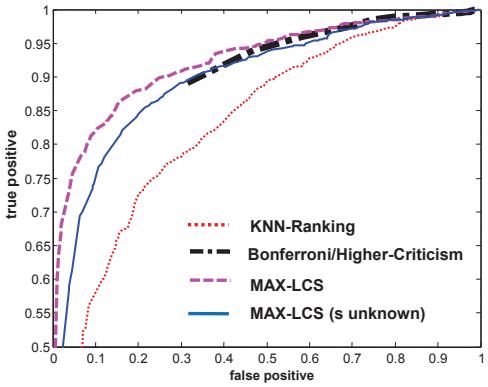

Figure 3: ROC curves for KNN-ranking, Bonferroni and Higher Criticsm (HC), MAX-LCS, and the variant of MAX-LCS when s is unknown. Length of time-series is 100; Duration of anomaly pattern is $s = 5$; Training sample size $n = 200$. For normal sequence the value of each time instant is IID Gaussian $\mathcal{N}(0, 1)$. Note that both HC and Bonferroni do not produce estimates for FP < 0.3.

For all experiments we set the time series length to be $T = 100$ and data size to be $n = 200$; and anomaly radius to be $s = 5$. The anomaly pattern is generated from a mixture model where we first randomly choose

a window $[i, i+s-1]$. Outside of the window x_t still follows the nominal pattern while in the window $[i, i+s-1]$ we generate $x_t \sim U[-4, 4]$. We choose $m = 2000$ test time-series where $m_0 = 800$ of them are drawn from H_0 and $m_1 = 1200$ are drawn from H_1.

IID Gaussian: In the first synthetic experiment, we generate the normal training time-series as IID standard Gaussian vector $\mathcal{N}(0, I_T)$. In the left panel of Figure 3 we plot the empirical ROC curve for KNN-ranking, Bonferroni, Higher Criticism, MAX-LCS and the variant of MAX-LCS where s is unknown. We can see that KNN-ranking performs much worse than our MAX-LCS. Bonferroni and Higher Criticism perform similarly when the FA $\alpha > 30\%$. However, when FP $< 30\%$, the empirical ROC curve is missing for these two algorithms(see Sec. 5.1 for explanation).

Inhomogeneous Gaussian: The nominal pattern follows the equation $x(t) = 3\sin(t/10+1) - 5\sin(3t/40) - 3$ and the variance is larger near the two ends compared to the variance in the middle (see Fig. 4). Again we see that HC and Bonferroni suffer from resolution at FP < 0.3.

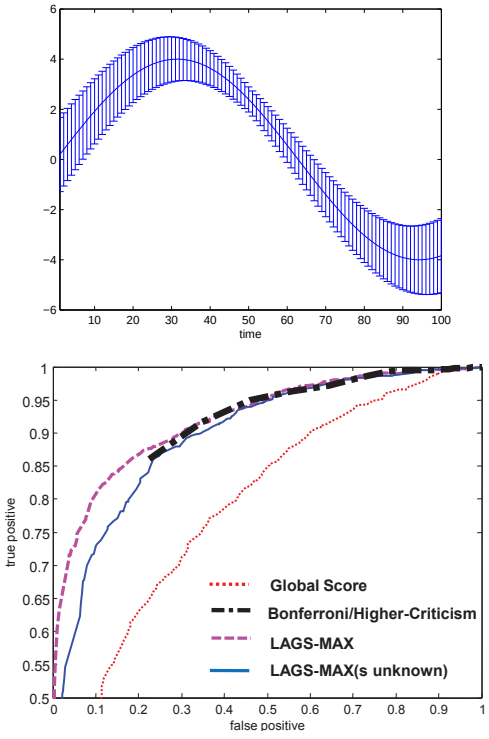

Figure 4: The length of time-series is 100; Anomaly Duration is $s = 5$. Training sample size $n = 200$. Top: Illustrates Inhomogenous Gaussian Process; The variance at time t varies as $\mathcal{N}(S(t), \sigma_t^2)$ where the error bar plot of parameter $(S(t), \sigma_t^2)$'s is shown. Bottom: ROC curves for different algorithms. Note that both HC and Bonferroni do not produce estimates below FP < 0.3.

Latent Model with Gaussian Mixtures: Here the normal pattern is drawn from a mixture model. There exist two normal patterns with equal prior probability: $x(t) = 3\sin(t/10 + 1) - 5\sin(3t/40) - 3$ and $x(t) = 4\sin(t/20) + 2$. We also have varying noise variance over time (Fig. 5). Note that this is a latent model (Section 2) since the markovianity is only satisfied conditionally. However, note that d-statistics can sometimes handle latent models. This is because the samples corresponding to the different mixtures are each uniformly distributed in the unit interval. Consequently, test sample is an anomaly if it is distant from each of the mixture components. In this example of a Gaussian mixture, the corresponding d-statistics turns out to be essentially unimodal. We can see that the Bonferroni-type correction and HC perform similarly and the ROC curves are missing when FP is $< 25\%$.

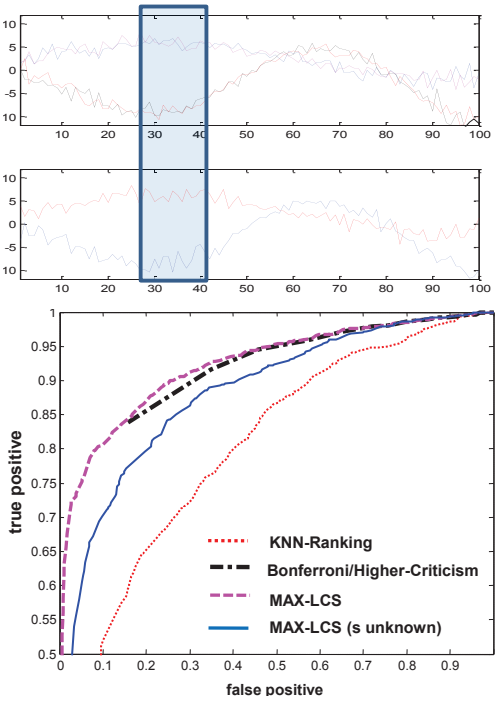

Figure 5: The time-series of length 100; Duration of anomaly pattern is $s = 5$. Training sample size $n = 200$. The nominal time-series is generated from a mixture model. Top: two sample paths of nominal training time-series. Bottom: ROC curve comparison. Note that HC and Bonferroni do not produce estimates below FP < 0.25.

6.2 Real data

We use the power consumption data set to verify our MAX-LCS algorithm. The power data set records the power consumption for a Dutch research facility for the whole year of 1997 [27]. Measurements are made every 15 minutes and there are totally $96 \times 365 = 35040$ measurements in the year 1997. We regard the mea-

surements of each day (of length 96) as an individual time-series. Hence we have $T = 96$ and we choose the window size $s = 16$. We set the threshold $\xi = 20\%$.

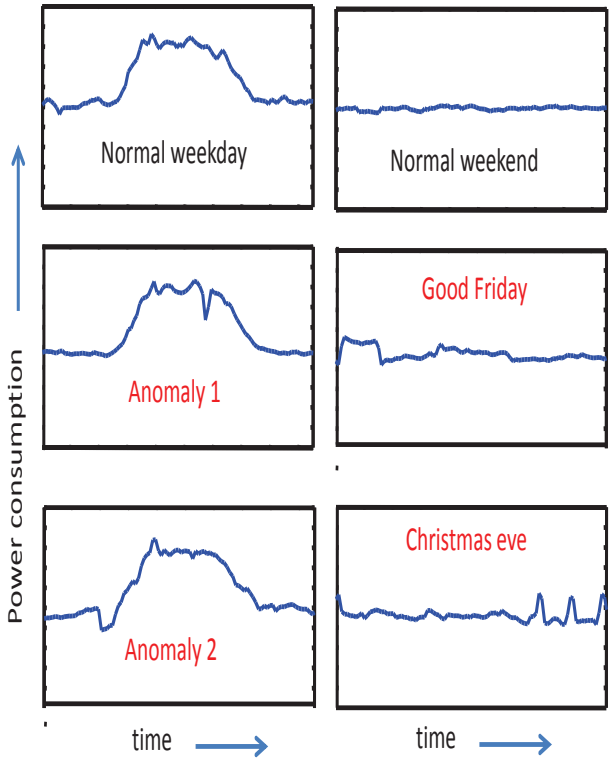

Figure 6: **Left Column**: the first plot shows a typical normal weekday. The other two plots are two anomalous weekdays (1997-05-12, 1997-10-19) returned by our MAX-LCS algorithm but not detected by Bonferroni, higher criticism or an algorithm based on KNN-ranking. **Right Column**: The first plot shows a typical normal weekend. The other two plots are two anomalous weekends or holidays (1997-03-02, Good Friday, Christmas Eve) returned by our MAX-LCS algorithm but not detected Bonferroni, higher criticism or an algorithm based on KNN-ranking.

Typical normal weekday and normal weekend time-series patterns are shown in the first column of Figure 6. The other three plots on the upper (lower) panel of Figure 6 show three anomalous weekdays (weekends or Holidays) that are returned by our MAX-LCS algorithm and that are not detected using the KNN-ranking. Interestingly, the last two plots in the lower panel actually correspond to the good Friday and Christmas Eve of 1997 (During most time of these days, the curve is indistinguishable from a normal pattern). We can see visually that actually the anomaly is local in time. Note that we cannot detect such subtle patterns of anomalies using KNN-ranking. Note that in our algorithm, we don't discriminate between weekdays or weekends and we never use this piece of information in the algorithm. The distinction between weekends and weekdays are just for the clarity of illustration. For illustration purposes, we also plot in Figure 7 the original time series of the last three weeks of 1997 (upper panel) and their score function via the $G_{MAX}(x)$ statistics (lower panel). Due to lack of

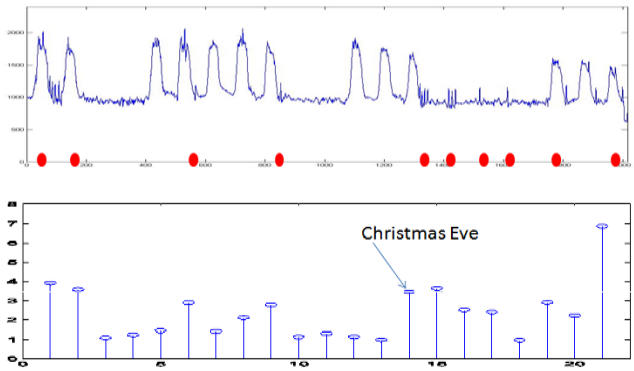

Figure 7: **Upper panel:** the original last three week time-series of 1997 (Power data set). The days with red dots correspond to the anomalous days detected by MAX-LCS. **Lower panel:** Score function for each day for the same period. Anomalies are now obvious.

ground truth information for the power demand data set, we cannot plot the ROC curve of our algorithm. Here what we choose to do is to inject some artificial anomalies to some of the days and compute the detection power. To be more precise, we inject artificial anomalies to all the 50 days between 1997-04-07 and 1997-05-26 by choosing a random 4-hour period and add Gaussian noises to these periods. Here we vary the threshold ξ and for different choice of ξ we get different detection power (see Figure 8). For MAX-LCS and KNN-ranking, ξ is exactly the proportion of declared anomalies. For the Bonferroni procedure, it is hard

Threshold	3%	10%	20%	30%	40%	50%
MAX LCS	6	16	26	35	45	49
KNN-Rank	0	2	7	21	29	37
Bonferroni	1	16	26	34	42	44

Figure 8: Detection Power of MAX LCS, KNN-Rank, Bonferroni on the Power data set with injected anomalies

to explicitly control the number of declared anomalies. For a fair comparison, one reasonable choice is to correct the threshold by a factor of $T/s = 6$, that is, the threshold $\xi' = \xi/6$ where ξ is used in MAX-LCS and KNN-ranking. We can see that MAX-LCS is consistently better than KNN-ranking. Bonferroni is comparable with MAX-LCS in a certain range of ξ but performs poorly when ξ is very small.

7 Appendix

To avoid cumbersome notation we consider several simplifying assumptions.

The proofs here consider the case of time series data. The data length is T and the number of samples are n and the anomaly radius is s. The underlying graph is just a Markov chain. In this case, a node v is reduced to an ordered index i. We also assume first order Markovianity, i.e., conditioned on the random variable, x_t at time t the random variables x_v, $v > t$ and x_u, $u < t$ are independent. The results can be easily generalized to the general Markov random graph. Consider balls of size s at location i. These are time windows

$$\omega_i = \{i - s,\, i - s + 1,\, \ldots,\, i + s - 1\}.$$

Let the 1-neighborhood window be

$$\omega_{i,1} = \{i - s - 1,\, i - s,\, i - s + 1,\, \ldots,\, i + s\}$$

8 Proof of Lemma 1

By the formula of the conditional probability,

$$\mathcal{L}_i(x) \quad = \quad \frac{f_i(x)}{f_0(x)} = \frac{f_i\left(x_{\omega_{i,1}}\right) f_i\left(x_{\omega_{i,1}^c}|x_{\omega_{i,1}}\right)}{f_0\left(x_{\omega_{i,1}}\right) f_0\left(x_{\omega_{i,1}^c}|x_{\omega_{i,1}}\right)}$$

By Markovian property, the second term in the numerator can be simplified to,

$$f_i\left(x_{\omega_{i,1}^c}|x_{\omega_{i,1}}\right) = f_i\left(x_{\omega_{i,1}^c}|x_{i-1}, x_{i+s}\right)$$

Note that all the variables in the RHS is in the normal window $[i, i+s-1]^c$, and by the mask property, it should equal to

$$f_i\left(x_{\omega_{i,1}^c}|x_{i-1}, x_{i+s}\right) = f_0\left(x_{\omega_{i,1}^c}|x_{\omega_{i,1}}\right)$$

which cancels out the second term of the denominator. The second equality follows from decomposing the likelihood ratio in another way:

$$\mathcal{L}_i(x) = \frac{f_i\left(x_{\omega_i^c}\right) f_i\left(x_{\omega_i}|x_{\omega_{i,1}^c}\right)}{f_0\left(x_{\omega_i^c}\right) f_0\left(x_{\omega_i}|x_{\omega_{i,1}^c}\right)}$$

The first terms of the numerator and the denominator cancel out due to the mask property. Moreover, due to Markov property, the second terms of numerator and the denominator can be simplified to be only dependent on the boundary x_{i-1}, x_{i+s}.

9 Proof of Lemma 2

Denote $y_0 = R(\eta_0)$ for a fixed η_0 and $y = R(\eta)$ for $\eta \sim f_0$. Then we have

$$\begin{aligned}
\Pr(y \le y_0) \quad &= \quad \mathbb{P}_{\eta\sim f_0}\left(\mathbb{P}_{x\sim f_0}\left(x : G(x) \ge G(\eta)\right) \le \mathbb{P}_{x\sim f_0}\left(x : G(x) \ge G(\eta_0)\right)\right) \\
&\overset{(a)}{=} \quad \mathbb{P}_{\eta\sim f_0}(\{x : G(x) \ge G(\eta)\} \subset \{x : G(x) \ge G(\eta_0)\}) \\
&\overset{(b)}{=} \quad \mathbb{P}_{\eta\sim f_0}(G(\eta) \ge G(\eta_0)) = y_0
\end{aligned}$$

where both (a) and (b) follow from the nestedness property of G and the last equality follows from the definition of y_0.

10 Proof of Theorem 3

The first statement follows directly from Lemma 2. The second statement follows from the fact that $G_{SUM}(\cdot)$ satisfies nestedness property and so is uniformly distributed. The optimality follows from the fact that for distributions that are not flat, the corresponding score $R(x)$ is a one-to-one monotonically increasing transformation of the likelihood ratio. The result now follows by noting that thresholding the likelihood ratio is itself an optimal detector.

11 Proof of Theorem 4

The following lemma establishes that $G_{n,MAX}$ converges to G_{MAX} in the limit.

Lemma 8. *Given the regularity condition $k \to \infty, n \to \infty$ and $k/n \to 0$, we have,*

$$\left(d_{\omega_{i,1}}(\eta)/\lambda_i\right)^s \to \frac{\xi_{\alpha,i}}{f_0(\eta_{\omega_{i,1}})} \quad \text{when } n \to \infty \tag{13}$$

where λ_i is set as Equation (13) and $\xi_{\alpha,i}$ satisfies $\mathbb{P}_{x\sim f_0}\left(f_0(x_{\omega_{i,1}}) \le \xi_{\alpha,i}\right) = \alpha$.

Proof. The convergence result is based on the asymptotic moments expression for k-NN distance distributions developed in [22, 28]. Given the regularity condition: $k \to \infty, n \to \infty$ and $k/n \to 0$, we have that [22, 28],

$$\frac{n}{k}\left(d_{\omega_{i,1}}(\eta)\right)^s \to \frac{1}{c_s f_0(\eta_{\omega_{i,1}})}$$

where c_s is some constant which only depends on the value of s. Then we can derive the limit expression of the compensation factor:

$$\frac{n}{k}(\lambda_i)^s = \frac{n}{k}\left(d_{\omega_{i,1}}^{(\lfloor n\alpha \rfloor)}\right)^s$$

Therefore $\frac{k}{n}(1/\lambda_i)^s$ will converge to the $\lfloor n\alpha \rfloor$'s smallest value of $c_s f_0(\eta_{\omega_{i,1}})$.

On the other hand, when n goes to infinity, the solution of $\mathbb{P}_{x\sim f_0}\left(f_0(x_{\omega_{i,1}}) \le \xi_{\alpha,i}\right) = \alpha$ can be approximated by the solution of $\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\left\{f_0(x_{\omega_{i,1}}^{(j)}) \le \xi_{\alpha,i}\right\}}$, which is exactly the the $\lfloor n\alpha \rfloor$'s smallest value of $f_0(\eta_{\omega_{i,1}})$. □

Based on the above lemma, we can easily establish Theorem 4.

12 Proof of Theorem 6

In this section, we want to quantitatively compare the detection power of LPE with MAX LPE when H_0 is Gaussian:

$$H_0 : x_i \sim \mathcal{N}(0,1), i = 1, \cdots, T$$

and the ground truth anomaly pattern H_1 is:

$$H_1 : \begin{cases} x_i \sim \mathcal{N}(\mu_i, 1) & i = 1, \cdots, s \\ x_i \sim \mathcal{N}(0,1) & i = s+1, \cdots, T. \end{cases}$$

that is, the anomaly always happens at the very beginning. We should understand that in the following calculation we only reveal H_0 to the detector (via training sample) and the detector is unaware of H_1.

We first derive the result for the extreme case $s = 1$. Later the proof techniques can be easily generalized to any sparsity level s. Here we adopt a Neyman-Pearson criteria to evaluate the two detectors, that is, we fix the false alarm at level α and compare their detection power. When H_0 is standard Gaussian, LPE detector (in limit) reduces to ANOVA analysis [29] and the detection boundary is $\sum_{i=1}^{T} x_i^2 = \xi$ where ξ is chosen such that $\mathbb{P}_{x_i\sim\mathcal{N}(0,1)}\left(\sum_{i=1}^{T} x_i^2 \le \xi\right) = 1-\alpha$. Also for Gaussian H_0 the boundary of MAX LPE (in limit) is simplified to $\max_{i=1}^{T} |x_i| = \xi'$ where ξ' is chosen such that $\mathbb{P}_{x_i\sim\mathcal{N}(0,1)}\left(\max_{i=1}^{T} |x_i| \le \xi'\right) = 1-\alpha$. The remaining job is to compute the detection power of these two detectors.

In computing the detection power of LPE, we will need the tail probability of non-central chi-square distribution [30].

Lemma 9 (Sankaran)**.** *Define non-central χ^2 distribution $Y = \sum_{i=1}^{n}\left(\frac{X_i}{\sigma_i}\right)^2$ where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and we also define $\lambda = \sum_{i=1}^{n}\left(\frac{\mu_i}{\sigma_i}\right)^2$. We have the following approximation formula for the tail probability of Y:*

$$\Pr(Y \ge y) \approx Q\left(\frac{(\frac{y}{n+\lambda})^h - (1 + hp(h-1-0.5(2-h)mp))}{h\sqrt{2p}(1+0.5mp)}\right) \tag{14}$$

where the Q-function $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-x^2/2} dt$ *and*

$$h = 1 - \frac{2}{3} \frac{(n+\lambda)(n+3\lambda)}{(n+2\lambda)^2}, \quad p = \frac{n+2\lambda}{(n+\lambda)^2}, \quad m = (h-1)(1-3h).$$

Proposition 10. *Consider the regime when* $s = 1$ *and* T *goes to infinity. To make the miss detection rate* P_M *of LPE vanish,* μ^2 *should at least scale as* $\Theta(\sqrt{T})$.

Proof. For LPE detector with decision boundary $\sum_{i=1}^T x_i^2 = \xi$, the false alarm rate is

$$\mathbb{P}_{x_i \sim \mathcal{N}(0,1)} \left(\sum_{i=1}^T x_i^2 \geq \xi \right) = \alpha$$

The RV $\sum_{i=1}^T x_i^2$ is chi-square distributed with freedom of degree T. By applying the tail probability approximation in Lemma 9, we have

$$Q\left(\frac{(\frac{\xi}{T})^{1/3} - 1 + \frac{2}{9T}}{\frac{1}{3}\sqrt{\frac{2}{T}}} \right) \approx \alpha$$

Solving this equation gives the estimate of ξ:

$$\xi \approx T + \sqrt{2T} Q^{-1}(\alpha) + o(\sqrt{T})$$

We can also evaluate the miss detection probability

$$P_M = \mathbb{P}_{x_1 \sim \mathcal{N}(\mu,1), x_i \sim \mathcal{N}(0,1), i \geq 2} \left(\sum_{i=1}^T x_i^2 \leq \xi \right)$$

In this case, we can regard $\sum_{i=1}^T x_i^2$ as a non central chi-square distributed random variable with $\lambda = \mu^2$. Again by applying the tail probability approximation in Lemma 9, we have

$$P_M = 1 - Q\left(\frac{(\frac{\xi}{T+\lambda})^{1/2} - 1 + \frac{1}{T}}{\frac{1}{2}\sqrt{\frac{8}{T}}} \right)$$

A necessary condition for P_M to vanish is the argument in the Q-function is at least negative:

$$\left(\frac{\xi}{T+\lambda} \right)^{1/2} - 1 + \frac{1}{T} < 0$$

which implies $\xi < T + \lambda$. Since we have already derived that ξ is $T + \sqrt{2T} Q^{-1}(\alpha)$, we finally get that $\mu^2 = \lambda$ should be at least on the order of $\sqrt{T}$. □

Proposition 11. *Consider the regime when* $s = 1$ *and* T *goes to infinity. To make the miss detection rate* P_M *of MAX LPE vanish,* μ^2 *should at least scale as* $\Theta(\log(T))$.

Proof. For MAX LPE detector with decision boundary $\max_{i=1}^T |x_i| = \xi$, to control FA at level α, we have

$$\mathbb{P}_{x_i \sim \mathcal{N}(0,1)} \left(\max_{i=1}^T |x_i| \leq \xi \right) = 1 - \alpha$$

which implies $(1 - 2Q(\xi))^T = 1 - \alpha$ and this gives the solution

$$\xi = Q^{-1}\left(\frac{1 - (1-\alpha)^{1/T}}{2} \right) \approx Q^{-1}\left(\frac{\alpha}{2T} \right) \approx \sqrt{2 \log(2T/\alpha)}$$

where the last step from the Chernoff bound approximation $Q(t) \approx \frac{1}{2}e^{-t^2/2}$. Now we can evaluate the miss detection probability

$$\begin{aligned} P_M &= \mathbb{P}_{x_1\sim\mathcal{N}(\mu,1),x_i\sim\mathcal{N}(0,1),i\geq 2}\left(\max_{i=1}^{T}|x_i|\leq\xi\right) && (15)\\ &= \mathbb{P}_{x_1\sim\mathcal{N}(\mu,1)}\left(|x_1|\leq\xi\right)\prod_{i=2}^{T}\mathbb{P}_{x_i\sim\mathcal{N}(0,1)}\left(|x_i|\leq\xi\right) && (16)\\ &= (1-\alpha)^{\frac{T-1}{T}}\left(Q\left(\mu-\xi\right)-Q(\mu+\xi)\right) && (17)\end{aligned}$$

To drive P_M to zero, a necessary condition is that $\mu-\xi$ is negative, i.e., μ^2 should be at least as large as $\xi^2 = 2\log(2T/\alpha)$. □

The above propositions address the simple case $s = 1$, but the derivation can be easily adapted to any sublinear sparsity level. Now we are ready to prove to prove Theorem 6.

The miss detection rate of LPE follows exactly from the same line of argument except that now $\lambda = \sum_{i=1}^{s}\mu_i^2$. The analysis of MAX LPE requires a bit more effort. The definition of G_{MAX} is

$$G_{MAX}(\eta) = \max_{i=1,2,\cdots,T-s+1}\prod_{t=i}^{i+s-1}\left(f_0(\eta_t)\right)^{-1}$$

where $f_0(\eta_t)$ is assumed to be standard Gaussian. When $s > 1$, neighboring products $\prod_{t=i}^{i+s-1}(f_0(\eta_t))^{-1}$ are correlated because they share common terms and this makes the analysis more involved. To simplify the analysis we assume T/s is an integer and we analyze the slightly modified G_{MAX}:

$$G'_{MAX}(\eta) = \max_{i/s \text{ is integer}}\prod_{t=i}^{i+s-1}\left(f_0(\eta_t)\right)^{-1}$$

that is, we just consider the max over the non-overlapping time windows. To control the false alarm at α, now we have

$$\mathbb{P}_{x_i\sim\mathcal{N}(0,1)}\left(\max_{i/s \text{ is integer}}\sum_{t=i}^{i+s-1}x_t^2\leq\xi\right) = 1-\alpha$$

Again by applying the tail probability of chi-square distribution we can solve ξ and get $\xi \approx s + \sqrt{4s\log\left(\frac{2T}{\alpha s}\right)}$. Also, similar to the argument in the proof of Theorem 10, P_M vanishes only if $\xi < s+\lambda$, or,

$$s + \sqrt{4s\log\left(\frac{2T}{\alpha s}\right)} < s + \sum_{i=1}^{s}\mu_i^2$$

and this proves the second part of the theorem.

Theorem 6 implies that MAX LPE outperforms LPE in the regime $T \geq \Theta(s\log(T/s))$. The smaller the sparsity s is relative to T, the more striking the performance difference is.

13 Proof of Theorem 7 and Extension to Correlated Case

We want to upper bound the expectation of $\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq \max_i d_{\omega_{i,1}}(\eta)\}}$. Without loss of any generality we assume that the anomaly happens at window $[1, s+1]$ for η. Then we have,

$$\begin{aligned}
&\mathbb{E}_{\eta,x^{(1)},\cdots,x^{(n)}}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq \max_i d_{\omega_{i,1}}(\eta)\}}\right)\\
=&\mathbb{E}_{\eta,x^{(1)},\cdots,x^{(n)}}\left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\}}\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq \max_{i>1} d_{\omega_{i,1}}(\eta)\}}\right)\\
\leq&\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left(\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\}}\right)\mathbb{E}\left(\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq \max_{i>1} d_{\omega_{i,1}}(\eta)\}}\right)
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality. The terms in the second expectation are all normal. Hence this term converges to 1/2 limit. Therefore the last expression simplified to

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left(\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\}}\right)\mathbb{E}\left(\mathbb{I}_{\{\max_i d_{\omega_{i,1}}(x^{[j]})\geq \max_{i>1} d_{\omega_{i,1}}(\eta)\}}\right)\\
=&\frac{1}{2n}\sum_{j=1}^{n}\Pr\left(\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\right)\\
=&\frac{1}{2}\Pr\left(\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\right)
\end{aligned}$$

Due to symmetry, the superscript $[j]$ can be any value. We just put a $[j]$ here. For the simplicity of notation, we denote $d_{\max}=\max_i d_{\omega_{i,1}}(x^{[j]})$ and the distribution of η as $q_\eta(\cdot)$. Now we have

$$\begin{aligned}
&\frac{1}{2}\Pr\left(\max_i d_{\omega_{i,1}}(x^{[j]})\geq d_{\omega_{1,1}}(\eta)\right)\\
=&\frac{1}{2}\Pr\left(d_{\max}\geq d_{\omega_{1,1}}(\eta)\right)\\
=&\frac{1}{2}\int_0^\infty q_\eta(w)\Pr(d_{\max}\geq w)dw\\
=&\frac{1}{2}\int_0^{\rho\log(T)/\lambda} q_\eta(w)\Pr(d_{\max}\geq w)dw+\frac{1}{2}\int_{\rho\log(T)/\lambda}^{\infty} q_\eta(w)\Pr(d_{\max}\geq w)dw
\end{aligned}$$

We bound the two integral separately. For the first integral, we make use of the assumption and get

$$\frac{1}{2}\int_0^{\rho\log(T)/\lambda} q_\eta(w)\Pr(d_{\max}\geq w)dw\leq\frac{1}{2}\epsilon$$

For the second integral, we have

$$\begin{aligned}
&\frac{1}{2}\int_{\rho\log(T)/\lambda}^{\infty} q_\eta(w)\Pr(d_{\max}\geq w)dw\\
\leq&\frac{1}{2}\left(1-\left(1-\frac{1}{T\rho}\right)^{T-s+1}\right)
\end{aligned}$$

and this term converges to $\frac{1}{2}(1-e^{-1/\rho})$ in limit.

References

[1] A. Lakhnia, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of ACM SIGCOMM*, Philadephia, PA, USA, 2005.

[2] PeMS., "Freeway performance measurement system." [Online]. Available: http://pems.eecs.berkeley.edu/.

[3] K. Sricharan, R. Raich, and A. O. H. III, "Empirical estimation of entropy functionals with confidence," December 2010, preprint, arXiv:1012.4188v1 [math.ST].

[4] T. H. Lotze, "Anomaly detection in time series: Theoretical and practical improvements for disease outbreak detection," Ph.D. dissertation, University of Maryland, College Park, 2009.

[5] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, pp. 117–122, January 2008. [Online]. Available: http://doi.acm.org/10.1145/1327452.1327494

[6] A. Vahdatpour and M. Sarrafzadeh, "Unsupervised discovery of abnormal activity occurrences in multi-dimensional time series, with applications in wearable systems," in *Proceedings of the SIAM International Conference on Data Mining*, Columbus, Ohio, USA, April 29 - May 1 2010, pp. 641–652.

[7] P. K. Chand and M. V. Mahoney, "Modeling multiple time series for anomaly detection," in *ICDM*, 2005, pp. 90–97.

[8] B.-K. Yi, N. D. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris, "Online data mining for co-evolving time sequences," in *Proceedings of the 16th International Conference on Data Engineering*, Washington, DC, USA, 2000.

[9] R. Fujimaki, T. Yairi, and K. Machida, "An anomaly detection metho for spacecraft using relevance vector learning," in *PAKDD*, 2005, pp. 785–790.

[10] B. Pincombe, "Anomaly detection in time series of graphs using ARMA processes," *ASOR Bulletin*, vol. 24, no. 4, pp. 2–10, 2005.

[11] I. V. Nikiforov and M. Basseville, *Detection of abrupt changes: theory and applications*. Prentice-Hall, New Jersey, 1993.

[12] D. Donoho and J. Jin, "Higher criticism for detectin sparse heterogeneous mixtures," *The Annals of Statistics*, vol. 32, no. 3, pp. 962–994, 2004.

[13] J. Jin, "Detecting a target in very noisy data from multiple looks," *IMS Lecture Notes Monograph*, vol. 45, pp. 1–32, 2004.

[14] E. A-Castro, E. J. Candès, and Y. Plan, "Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism," July 2010, preprint, arXiv:1007.1434v1.

[15] E. A-Castro, E. J. Candès, and A. Durand, "Detection of an anomalous cluster in a network," January 2010, preprint, arXiv:1001.3209v2.

[16] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, July 2001.

[17] A. O. Hero, "Geometric entropy minimization(GEM) for anomaly detection and localization," in *Neural Information Processing Systems Conference*, vol. 19, 2006.

[18] M. Zhao and V. Saligrama, "Anomaly detection with score functions based on nearest neighbor graphs," in *Neural Information Processing Systems Conference*, vol. 22, 2009.

[19] "Ucsd anomaly detection dataset." [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/

[20] F. Nicolls and G. de Jager, "Optimality in detecting targets with unknown location," *Signal Processing*, vol. 87, no. 5, pp. 841–852, 2007.

[21] S. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998.

[22] Y. P. Mack and M. Rosenblatt, "Multivariate k-nearest neighbor density estimates," *Journal of Mulivariate analysis*, vol. 9, pp. 1–15, 1979.

[23] J. Glaz, "Extreme order statistics for a sequence of dependent random variables," *IMS Lecture Notes*, vol. 22, Stochastic Inequalities, pp. 100–115, 1993.

[24] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[25] B. J. Becker, "Combining significance levels," in *The Handbook of Research Synthesis.* Russell Sage Foundation, New York, 1994.

[26] P. Hall and J. Jin, "Innovated higher criticism for detecting sparse signals in correlated noise," *The Annals of Statistics*, vol. 38, no. 3, pp. 1686–1732, 2010.

[27] E. Keogh, J. Lin, and A. Fu, "HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications," in *Fifth IEEE International Conference on Data Mining*, Houston, Texas, USA, 2005.

[28] D. Evans, A. J. Jones, and W. M. Schmidt, "Asymptotic moments of near neighbor distance distributions," in *Proceedings of the Royal Society of London A 458*, 2002, pp. 2839–2849.

[29] R. A. Fisher, *Statistical methods for research workers.* Oliver and Boyd, London, 1932.

[30] M. Sankaran, "Approximations to the non-central chi-square distribution," *Biometrika*, vol. 50, no. 1-2, pp. 199–204, 1963.