A Framework for Discovering Anomalous Regimes in Multivariate Time-Series Data with Local Models

Stephen Bay,[1] Kazumi Saito,[2] Naonori Ueda,[2] and Pat Langley[1]
[1] Center for the Study of Language and Information
Stanford University, Stanford, CA 94305, USA
[2] NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Soraku, Kyoto 619-0237 Japan
{sbay,langley}@apres.stanford.edu
{saito,ueda}@cslab.kecl.ntt.co.jp

ABSTRACT
In many modeling endeavors, researchers observe a physical system and collect data measuring its behavior over time. This data is often used to build models that predict the future behavior of the system and explain relationships between the measured variables. However, systems can change over time and discovering these changes is important for detecting abnormal behavior and identifying new phenomena. In this paper, we focus on the problem of discovering when an observed system has moved into a new anomalous regime. The regime may be defined by a change in the functional relationships between the variables or by the introduction of a previously unseen causal effect. We propose a general solution framework for time-series data based on comparing local models inferred on test data to those trained on a reference set. We show that our approach scales efficiently in the size of the data, and we demonstrate that our framework is able to detect anomalous regimes with high sensitivity and selectivity.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications—*data mining*

Keywords
anomaly detection, multivariate time series, regimes

1. INTRODUCTION
In many scientific and engineering endeavors, researchers observe a physical system and collect data measuring its behavior over time. This data is often used to build models that predict the future behavior of a system or explain how the measured variables relate to each other. However, systems change over time and our goal in this paper is to discover from observed data when the dynamics of the system has changed to a novel state. We refer to this problem as the task of discovering *anomalous regimes* in multivariate time-series data. By anomalous we mean a situation that is either extremely rare or that has not been encountered before. By regime, we refer to the hypothetical true model of the system that generates the observed data.

An anomalous regime could be caused by a change in the existing functional relationship between variables or by the introduction of a previously unseen causal effect. Clearly an algorithm able to detect these changes would be very useful as an exploratory tool to discover new phenomena from data, and as a monitoring device to detect when a system is behaving abnormally. For example, consider monitoring a physical device such as a battery from observations on voltage, current, and state-of-charge. For most batteries, the voltage depends on the state-of-charge, which in turn is a function of charge accumulation (and loss) through current. If after time, the state-of-charge accumulates more slowly for a given current flow this may indicate that the battery has degraded and is in need of replacement.

Note that our problem is different from the tasks of discovering *outliers* [27, 13] and *unusual patterns* [18]. In outlier detection the goal is to find individual data points that have extreme values, primarily to remove them from the analysis to prevent an undue effect on the inferred models. In anomalous pattern finding the goal is to find a set of contiguous points whose physical shape or expression of values is unusual. In contrast, our goal is to find regions where the generating process has changed.

To highlight the difference between anomalous patterns and regimes consider data from a random walk. A random walk can generate all possible patterns and two examples are shown in Figure 1. A pattern based approach would identify that the two plots are very different: the first pattern oscillates while the second is a linear downward trend. In contrast, an approach that models the variables relationships should identify that successive changes in the signal are independent and hence consider both curves from the same regime.

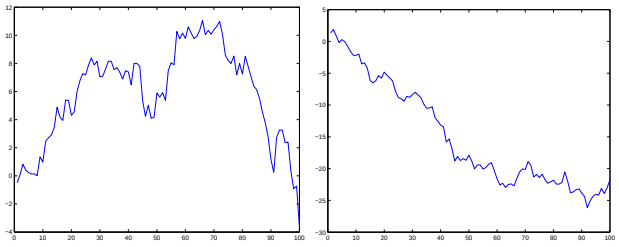

Figure 1: Two different sequences generated by a random walk.

In this paper, we propose a new framework for discovering anomalous regions in time-series data. The basic idea is simple: given a reference and test set of data, transform the time series into a set of local models where each is trained on a small time bounded set of data. Compare the models from the test set to those from the training set in the parameter space to find the anomalies. In this framework there is no need for an explicit characterization of normal behavior as the method automatically models the space of variable relationships from reference data.

The remainder of this paper is organized as follows. In the next section, we present our framework for mining anomalous regimes in time-series data along with a simple expository example. In Section 3, we discuss scaling issues and show that our framework can be implemented efficiently. In Section 4, we evaluate our approach by applying our algorithm to real and synthetic data sets. In Section 5, we discuss the limitations of our approach and directions for future work. In Section 6, we review the related work, and in the last section we present concluding remarks.

2. DISCOVERING ANOMALOUS REGIMES IN TIME SERIES

In this section we present our framework called DARTS which stands for Discovering Anomalous Regimes in Time Series. As input, we will assume that we are given a multivariate time series R that serves as a reference of normal behavior and a test series T. The goal is to find sections of T where there is a new relationship between the variables not present in R. The output will be a vector for each measured variable whose elements represent the anomaly score for a corresponding time point. The framework involves the following six steps:

1. Perform any needed preprocessing on the time series R and T. This may include normalization of variance, centering the mean values, deseasonalizing, differencing, and aggregation to reduce dimensionality on the time axis.

2. Choose a variable as a target and create a set of local models to predict it from R by using sliding (or disjoint) windows to select training data. The coefficients of the local model define a mapping from R into a new space $\mathcal{P}$.

3. Create a set of local models representing T and map them into $\mathcal{P}$ by using the same sliding window technique as in step 2.

4. Compare the local models of T to R in space $\mathcal{P}$ and calculate an anomaly score based on the density estimate for each model of T in $\mathcal{P}$.

5. Compute a null distribution of the anomaly scores one would expect to see if there were no changes by comparing R to itself. Points in T that have extreme values according to the null distribution are anomalous.

6. Repeat steps 2-5 for each possible target variable.

We discuss the main steps in greater detail by highlighting our instantiation of this framework and then follow with an expository example.

2.1 Local Models

In this paper, we will examine local models which predict the present value of a variable as a linear combination of past values of itself and other variables. More specifically, we look at vector autoregressive (AR) models [20] which have the form,

$$X_i(t) = b_0 + \sum_{i=1..p} \sum_{j=1..l} b_{ij} X_i(t-j) \qquad (1)$$

Where X is the time-series data with p variables and N time points. The subscript i refers to the variable, l is the maximum lag in the AR model, and b_0 is a constant bias term. For p variables with lags up to order l there are $p*l+1$ parameters. The parameters b_0 and b_{ij} define the mapping into a space $\mathcal{P}$.

We advocate using local models for two reasons. First, although many systems may be governed by a globally non-linear function, in limited areas the function may be linear (or nearly so). For example, batteries have a complex non-linear function that governs it's behavior and depending on whether the battery is charging or discharging different equations apply. However, within each mode a simple linear model may suffice to model its behavior [3]. Second, many models are computationally expensive to learn and may be quadratic or worse in the number of data points. Thus, learning multiple local models can be much faster than learning one global model (e.g., if $w << N$ then $Nw^2 << N^2$).

For notational simplicity, we will assume the time series X has been flattened into its marginal form where $X_f(t) = \{X_1(t-1)X_1(t-2)...X_1(t-l)...X_p(t-1)X_p(t-2)...X_p(t-l)\}$ and that $y(t) = X_i(t)$. Doing so allows one to conveniently express the solution of the parameters in Equation 1 in matrix form. For instance, the model in Equation 1 can be estimated with least squares as follows:

$$\mathbf{b} = (\mathbf{X_f}'\mathbf{X_f})^{-1}(\mathbf{X_f}'\mathbf{y}) \qquad (2)$$

where $\mathbf{b}$ is the vector of coefficients corresponding to b_0 and b_{ij} in Equation 1. Note that the bias term b_0 has been incorporated into the matrix $\mathbf{X_f}$ by representing it as a constant variable [8].

Learning local models in this fashion requires estimating $p \cdot l+1$ parameters and this may cause difficulties if the window

size is small. One method of mitigating the data sparseness problem is a technique called *ridge regression* [14, 21], which can be viewed as a form of regularization or weight decay. In ridge regression the coefficients are computed as follows.

$$\mathbf{b} = (\mathbf{X_f}'\mathbf{X_f} + \lambda\mathbf{I})^{-1}(\mathbf{X_f}'\mathbf{y}) \quad (3)$$

The term $\mathbf{I}$ is the identity matrix and λ is the ridge parameter and it should be non-negative. A value of zero means that ridge regression is identical to linear regression. Increasing values of λ result in greater regularization. In this paper, we assume that λ is fixed for all local models. There are several automatic methods for determining λ [16, 15] and of course cross-validation may be used.

Finally, although we explore the use of vector AR models in this paper, other models are possible. For example, in some domains an appropriate local model might be a linear differential equation.

2.2 Scoring and Density Estimation

In step 4 of our framework, we estimate the probability of points located in the test set according to the distribution of the reference set. The two main methods for density estimation are parametric models (and mixtures thereof) and kernel methods. In this paper, we suggest using kernels because they have two important advantages over parametric methods. First, kernel based density estimates are asymptotically optimal and as the number of data points grows, will converge on the true value. In contrast, any model based method must necessarily be wrong as it depends on parametric assumptions that are never completely correct. Furthermore, in many domains massive amounts of time-series data are available and kernel based methods can take maximum advantage of this. Second, modeling the density with a parametric model, such as a mixture of Gaussians, will generally require structural identification to determine the number of components and this can be time consuming especially when interleaved with parameter estimation.

We define the anomaly score as the negative logarithm of the probability density. More formally, let $A(x)$ be the anomaly score of a point x in space $\mathcal{P}$ and can be computed as follows.

$$A(x) = -\log \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x_i - x}{h}) \quad (4)$$

where n is the number of points in $\mathcal{P}$ from R, K is the kernel function, and h is the bandwidth. There are several choices for the kernel function and a popular one is the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}u^2)$.

Our initial experiments indicated that kernel methods often tended to produce zero valued estimates for points corresponding to anomalous regimes. We believe this is because anomalous test points could have extremely low probability coupled with numerical errors. Thus our computed density was often not helpful in distinguishing between the more and less extreme anomalous regions of our signal. Hence, in this paper we did not use a strict probability measure for our anomaly score and instead we used the Euclidean distance to the k-th nearest neighbor for the reference points in $\mathcal{P}$: i.e., $A(x) = distance_R(x, k)$. The distance to the k-th nearest neighbor implicitly corresponds to a density estimate of k/nV where V is the volume of the hypersphere needed to encompass k points [9]. Under certain kernels, such as Gaussians with diagonal covariance, Equation 4 yields nearly identical results to the k-th neighbor when zeroing is not a problem.

Fast methods have been developed for estimating the density for vector spaces [12] based on index structures such as KD trees [4, 10]. However, kernel density estimation can be slow if applied to high dimensional data. In Section 3 we discuss possible solutions in this situation.

2.3 Determining a Null Distribution

In the density estimation step, sections of the time series are graded on a continuous scale with higher values assigned to more anomalous regions. However, in many situations users require a hard cutoff as a trigger to initiating action such as a manual review of the time series in a monitoring application. To make this binary decision we take a frequentist hypothesis testing approach and compare the score of observed test data to a null distribution.

The null distribution is important because it can be used to precisely set a cutoff that corresponds to the probability of falsely rejecting the null hypothesis (Type I error). Given a Type I error rate, we determine the appropriate cutoff and then segment the test series into anomalous and non-anomalous regions by discretizing accordingly.[1]

We use the data in R to define a distribution of anomaly scores one could expect when comparing normal data to itself. We develop an empirical distribution by using block cross-validation: We divide the data into n continuous blocks and select each block in turn to serve as the test data with all other non-adjacent blocks as the reference set. We do not use adjacent blocks to prevent shared data from the reference and test sets as the local models are inferred on a window of data. The anomaly scores from all the folds serve as the empirical distribution.

2.4 Example

To demonstrate our framework, we will present an expository example by analyzing a simple synthetic univariate time series. As shown in Figure 2A, the basic signal is a sinusoid with some noise added. Part B shows the test signal, which is also a sinusoid with the same period and amplitude, but has three anomalies inserted at roughly time points 60-80, 120-140, and 160-170.

For now we assume no preprocessing (step 1). In our framework, we generate a set of local models by sliding a window across R to select a subset of the training data. For each window we infer an AR model (order = 3) and its parameters define a new space into which segments of the time series are mapped. We repeat this for the training data as well. Figure 3 shows the mapping of the reference and test data into parameter space. The reference data forms a compact elliptical cloud. The test data has many points at the same location, but also many points spread out.

The next step is to score the points from T given the dis-

[1] For multiple hypotheses there are standard corrections [25].

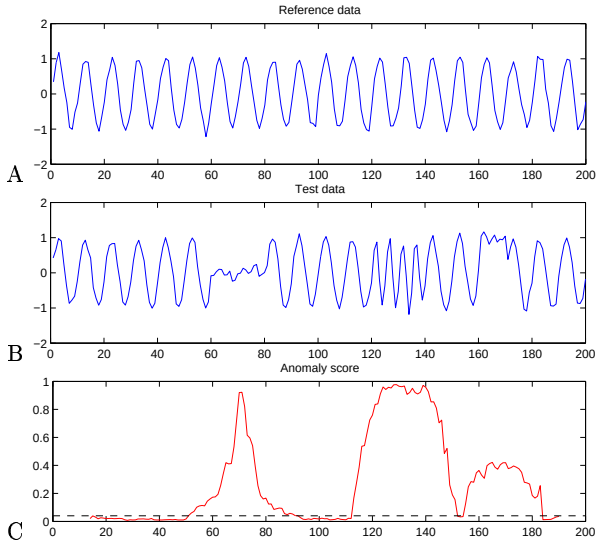

Figure 2: Artificial sine data with noise and anomalies added. (A) Reference data with no anomalies, (B) test data with three anomalous regions, (C) the anomaly score computed by DARTS. The cutoff is the dashed horizontal line.

tribution of R. Figure 2C shows the anomaly score plotted along the time axis based on the distance to the k-th nearest neighbor ($k = 5$). Note that each point represents the score of a window centered on that point. Higher values correspond to lower density estimates.

We also estimate a null distribution by comparing R to itself. The cutoff is shown by the dashed horizontal line in Figure 2C. The method has clear peaks greater than the cutoff magnitude at the locations of the three anomalies. The peaks are somewhat wider than the anomaly because of the sliding window. Finally, in normal regions the score is less than the cutoff.

2.5 Alternative Methods

In the remainder of this paper, we will present experiments that evaluate and compar DARTS to other methods. In choosing alternative methods, we focussed on those that could be applied to multivariate time series and whose purpose was to detect when the underlying process has changed.

In particular we will compare DARTS with Hotelling's T^2 statistic [17, 28, 29] a widely used method for multivariate statistical process control. The T^2 statistic is a measure of the distance between two groups of instances and is defined as follows:

$$T^2 = \frac{n_1 n_2 (\mu_1 - \mu_2)\mathbf{S^{-1}}(\mu_1 - \mu_2)}{n_1 + n_2} \quad (5)$$

where n_1 and n_2 are the sample sizes, μ_1 and μ_2 are the vector sample means with the subscripts 1 and 2 referring to the group membership. The term $\mathbf{S}^{-1}$ is the pooled sample covariance matrix. The T^2 statistic is the same as Mahalanobis distance for two groups. The T^2 is distributed as F(p,$n_1 + n_2 - p - 1$) and thus allows hypothesis tests of a

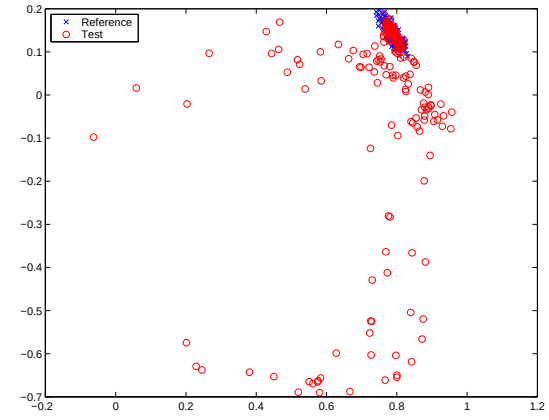

Figure 3: Local models of R and T mapped into the parameter space (2 of 4 dimensions shown).

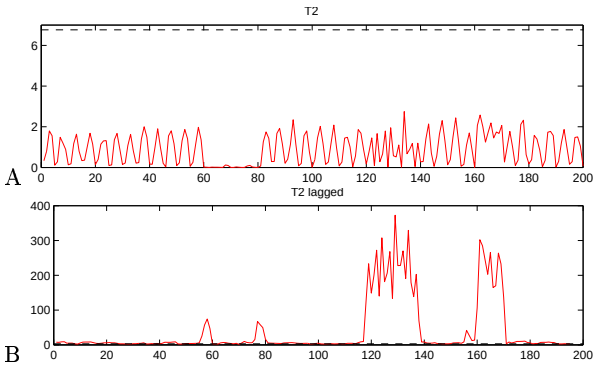

Figure 4: Hotelling's T^2 statistic: (A) without lagged variables, (B) with lagged variables.

similar form as suggested in section 2.3. Gather et al. [11] presented an extension to Hotelling's T^2 statistic for monitoring online data by incorporating time delays where lagged variables are included in the vector space in the same fashion as discussed in Section 2.1.

Figure 4 shows the results of applying the T^2 statistic to the sinusoidal data we examined in the previous section. In this situation the reference data is used to estimate μ_1, and μ_2 is simply the current point in the test set. The cutoff defined by the F distribution is shown by the dashed line.

The basic T^2 statistic fails to detect any anomalies. Although the entire series is below the cutoff, the first anomaly shows up as the least anomalous region so lowering the cutoff would not help. The time-delayed version fairs much better and clearly detects the last two anomalies. It also detects the transition points of the first anomaly, but does not register the central portion as unusual.

Finally, it is common to smooth the T^2 score with an exponentially weighted moving average, i.e., $EWMA_i = T_i^2\gamma + (1-\gamma) * EWMA_{i-1}$. This can also be applied to the scores generated by DARTS.

3. COMPUTATIONAL EFFICIENCY

Computationally efficient algorithms are important because in many domains researchers have large data sets to analyze. Furthermore, in online settings where a data stream is being monitored one must obviously process the data as fast as it arrives which is not trivial given the state of current measuring devices which can easily record hundreds or thousands of variables with high sampling rates.

We divide the complexity analysis of our algorithm into two parts. We first discuss the cost of constructing the local models, and then performing density estimation. To begin, let N_R be the size of the reference data set and w be the window size (the number of points used for inferring local models). In step 2, we place a window over the reference time series and data that falls within is used to train a local model. The window is advanced by one time point and the process repeats a total of $N_R - w$ times. For the reference data then, training the local models will take $O(N_R f(w, v))$ where $f(\cdot)$ is the complexity of inferring a local model on w time points with v variables. Similarly, generating models for the test set will take $O(N_T f(N_w, v))$ where N_T is the size of the test data.

The complexity of the function $f(w, v)$ for constructing local models will obviously depend on both the structure and the learning algorithm. For our regression models, solving Equation 3 for can be done in $O(v^3 + v^2 w)$ [30]. However, since we are using sliding windows one can reduce the cost for multiple inferences by iteratively updating the terms $\mathbf{X_f'X_f}$ and $\mathbf{X_f'y}$: simply add and subtract the contributions of the entering and leaving points.

In step 4, DARTS performs density estimation in the parameter space of the local models. Fortunately, density estimation can be done quickly in moderate dimensional spaces using techniques such as KD-trees. Creating the KD tree is efficient and can be done in $O(N_R \log N_R)$; finding the k nearest neighbors of a test point can be done in $\log N_R$ time. Assuming logarithmic retrieval performance, processing the entire test set will take time $O(N_T \log N_R f(N_w, v))$.

Experiments conducted by Moore [22] suggest that when the underlying dimensionality of the data is low (roughly 10 dimensions or less) KD-trees are very efficient. However, in higher dimensions the performance suffers and the KD-tree search can be much slower. There are several solutions: (1) apply a dimensionality reduction technique such as PCA or feature selection, (2) use alternative index structures such as metric-trees [23] that have better performance in high dimensions, and (3) use randomization and pruning techniques to identify the most extreme points [2].

To summarize, DARTS can be implemented extremely efficiently and has good scaling properties in terms of the size of the reference and test data sets. Processing the training set, including creating the KD tree, takes $O(N_R f(w, v) + N_R \log N_R)$. Applying DARTS to process the test data takes $O(N_T \log N_R f(w, v))$ with logarithmic retrieval.

4. EXPERIMENTAL EVALUATION

Evaluating an anomaly detection algorithm is a conceptually simple task. One applies the algorithm to a data set with known anomalies and measures the algorithm's sensitivity and selectivity, which are respectively the ability to detect true anomalies and prevent spurious discoveries. However, in practice objective evaluation of anomaly detection algorithms is very difficult because ground truth is seldom known for most data sets.

Because of this problem, we will evaluate the ability of DARTS to detect anomalous regions on the following four data sets, each chosen to test different aspects of sensitivity and selectivity.

- **CD Player**. In this data set, we artificially inject an anomaly whose detection requires examining the relationship between multiple variables.
- **Random Walk**. This data set is synthetically generated and has no anomalous regimes.
- **ECG Arrhythmia**. This is a well understood domain (from the point of the anomalies that can occur) and has been annotated by experts. It provides an objective test of sensitivity and selectivity.
- **Financial Time-Series**. This multivariate domain serves as a subjective test by a domain expert of the discovered anomalies.

Unless otherwise noted, our experiments were run on a Windows PC in Matlab. The k-th nearest neighbor was found with a KD-tree as implemented by the Kernel Density Estimation Toolbox developed by Alex Ihler and Mike Mandel[2]. In general, we divide signal into two equal sized parts and use the first part as the reference set and the second as the test set. We use a window size of 21, order 3, and k = 5 and no smoothing of the anomaly signal. The cutoff was set to a level covering 98% of the empirical null distribution.

4.1 CD Player Data

The CD player data represents measurements of a mechanical arm for a CD player. There are four measured variables $\{u_1, u_2, y_1, y_2\}$ where u_1 and u_2 are input forces to the mechanical actuator and y_1 and y_2 relate to the arm's tracking accuracy. We obtained this data from the DAISY library of system identification data sets [1].

Figure 5A shows a plot of the reference signal and Figure 5B shows the test data for variable y_2. We added a synthetic anomaly by copying the time series for variable y_2 from points 800 to 900 onto points 300 to 400. By replicating the signal itself we preserve the same marginal distribution and patterns of the variable's expression. Except for the boundary points, this duplication should be undetectable by methods that examine signals in isolation.

Figure 5C shows the results for DARTS which clearly identifies the anomalous region. Hotellings T^2 without (D) and T^2 with lagged variables (E) also correctly identified the anomaly although the cutoff score computed from the T^2 distribution appears to be too low for both variations. Applying DARTS and the T^2 statistic to variable y_2 alone, fails to detect any anomalous regions.

[2]available at http://ssg.mit.edu/~ihler/code/kde.shtml

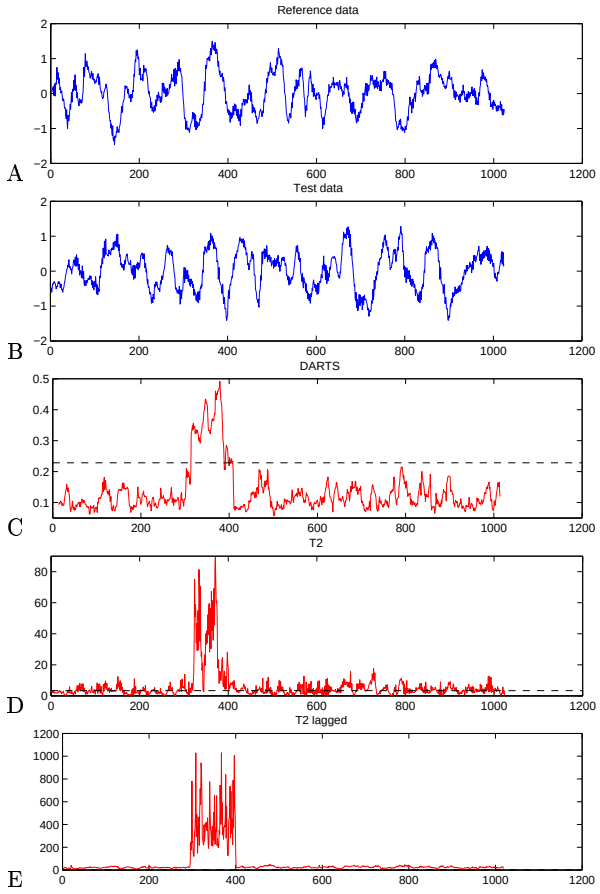

Figure 5: CD Player data: (A) y_2 of the reference signal, (B) y_2 of the test data with an artificial anomaly introduced at time points 300-400 by replicating data from time points 800-900, (C) DARTS anomaly score, (D) Hotelling T^2 statistic, (E) Hotelling T^2 statistic with lagged variables.

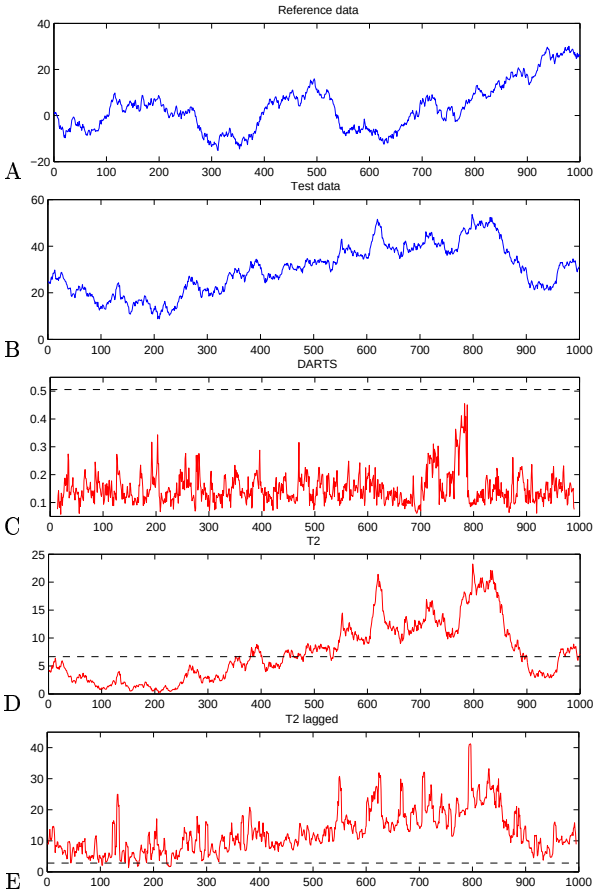

Figure 6: Random walk data: (A) the reference signal, (B) the test data, (C) DARTS anomaly score, (D) Hotelling T^2 statistic, (E) Hotelling T^2 statistic with time lagged variables.

4.2 Random Walk Data

We choose random walk data as a test of the selectivity of our approach. Since random walk data is generated by a single process, there are no anomalous regimes and algorithms for detecting anomalies should not return any results.

We generated univariate random walk data. Figure 6A and 6B show the reference and test signals. Figure 6C shows the results for DARTS. There were no anomalous regions that scored higher than the cutoff. Figure 6 D and E show the results for the T^2 statistic without and with variable lags, and both falsely flagged large portions of the signal as anomalous.

The T^2 statistic fails on this problem because it makes the implicit assumption that the process can be described by a single mean vector with a Gaussian distribution. As the random walk drifts further from the reference mean, the computed statistic increases and eventually exceeds the alloted boundaries. DARTS avoids this problem because the local models can condition their prediction on past values so large scale drifts do not affect it.

4.3 Arrhythmia Data

We obtained electrocardiogram (ECG) traces from the MIT-BIH Arrhythmia Database[3] which contains many recordings originally collected to evaluate arrhythmia detectors and to aid research in cardiac dynamics.

This database is significant from the perspective of evaluation because every beat is carefully annotated with a label which categorizes it as normal or as one of approximately a dozen abnormal beats. Each record was independently annotated by at least two cardiologists and any disagreements were resolved. Although there could be errors in the annotation, this data set provides an objective test to evaluate the selectivity and sensitivity of DARTS.

We selected one trace corresponding to the subject that had the widest variety of abnormal beat types.[4] The data itself consisted of a 30 minute recording sampled at 360 Hz and in total there were 650,000 data points. We used points 100-3000 as the reference set (Figure 7A) and 3001-650,000 as the test set. All beats in the reference set had a normal

[3] http://www.physionet.org/physiobank/database/mitdb/

[4] Record 201 contained seven different abnormal beat types.

beat label. In addition, because of the high resolution of the signal we used exponentially weighted moving averaging smoothing with $\gamma = 0.9$.

We evaluated DARTS by tabulating the number of true and false positives and negatives. However, measurement is somewhat complicated because although every beat is annotated with a label, the exact temporal extent of the beat is not indicated. Thus we defined a true positive having an anomaly score greater than the cutoff within 150 points of an abnormal beat's time stamp (on average the time points are approximately 300 time points apart). Similarly we defined a true negative as having no score greater than cutoff within 150 time points of a normal beat. Furthermore, we did not include scores for normal beats that were adjacent to an anomalous beat.

The results for DARTS are shown in Table 1. Sensitivity is computed as $TP/(TP+FN)$ and selectivity as $TP/(TP+FP)$. These results are very encouraging as DARTS was able to identify many of the anomalies with low false positive rates. Figure 7B and C show examples of three anomalous beats that were correctly detected by DARTS. From left to right the beats were classified by the cardiologists as a premature ventricular contraction followed by two aberrated atrial premature beats.

In contrast, Hotelling's T^2 statistic performed extremely poorly and exhibited no selectivity for anomalous regions. For example, as shown in Figure 7D and 7E, both versions of the T^2 statistic assigned the highest scores to the main beat impulse which is a normal part of the signal.

Table 1: Statistics for DARTS.

Threshold	TP	TN	FP	FN	Sensitivity	Selectivity
97%	313	984	127	66	82.6%	71.1%
98%	285	1040	71	94	75.2%	80.1%
99%	205	1088	23	174	54.1%	89.9%

Finally, although DARTS performed well on this data set, in some ways it was a difficult test as the local models may not be the most appropriate formalism for capturing ECG signals which are non-stationary and have much longer time dependencies than the local windows.

4.4 Japanese Financial Data

In the previous section, we analyzed ECG data which allowed for objective evaluation of our anomaly discovery algorithm because every heart beat was labelled by cardiologists. However, the data set was univariate and did not test the multivariate capabilities of our framework. Thus, in this section, we analyze a multivariate Japanese financial time-series data set. Unfortunately this data does not allow for precise measurements of sensitivity and selectivity because it is unannotated. However, we will present a subjective evaluation of the discovered anomalies by a domain expert.

We obtained monthly financial time-series data from Japan covering the period 1983-2003. There are a total of 240 points and the measured variables are: monetary base, national bond interest rate, wholesale price index, index of

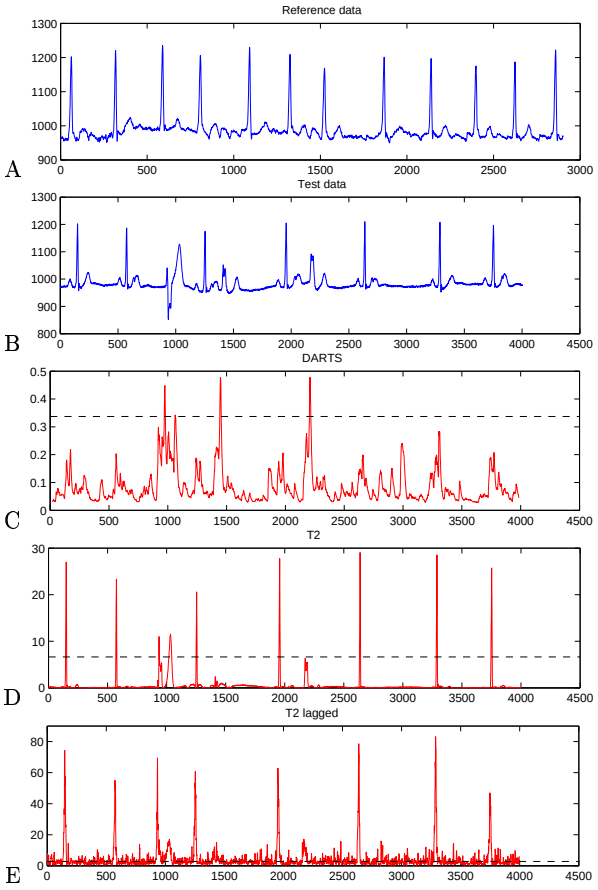

Figure 7: Arrhythmia data: (A) the reference signal, (B) the test data, (C) DARTS detects three anomalies, (D) Hotelling T^2 statistic, (E) Hotelling T^2 statistic with time lagged variables. The time axis is realigned to begin at zero for all plots.

industrial produce, machinery orders, and exchange rate yen to dollar. The monetary base represents the average outstanding balances stored with the Bank of Japan. The wholesale price index represents an average of the selling prices received by domestic producers (does not include services). The index of industrial produce is a measure of the goods produced by the economy.

We used the data from 1983-1991 as the reference set and the remainder as the test set. The raw data was non-stationary and was transformed to a stationary series through several operations including adjustments for seasonality and trends. The normalization procedure is described in [24].

DARTS finds clear anomalies in four variables which we list below along with the domain expert's opinions. The corresponding graphs are shown in Figures 8 to 11:

- **Machinery orders, 1992-1993**. The expert believed that anomaly was caused by an undocumented change in the preprocessing of the raw data.
- **National bond interest rate, late 1998**. In Au-

gust, 1998, there was a substantial default in Russian bonds. This caused a number of large hedge-fund companies to fail and the Bank of Japan rapidly raised interest rates in September, 1998. This is clearly an anomalous period.

- **Monetary base, 2001-2003**. The Bank of Japan rapidly raised funds in this period. The domain expert thought this might be the result of a policy change.

- **Index of industrial produce, 2001-2003**. The expert considered this to be the most interesting anomaly. In this period, the index of industrial produce greatly increased (after a drop) and the expert attributed this to a rapid increase in exports to China and other Asian countries.

Our expert's evaluation indicated that with the exception of the anomaly in machinery orders, all could be considered true regime changes. The anomaly in machinery orders did not correspond to an economic event, but it's discovery is understandable given that the preprocessing method changed in the middle of the series.

Hotelling's T^2 approach also identified similar time periods to DARTS as anomalous, although it does not make a distinction for different variables.

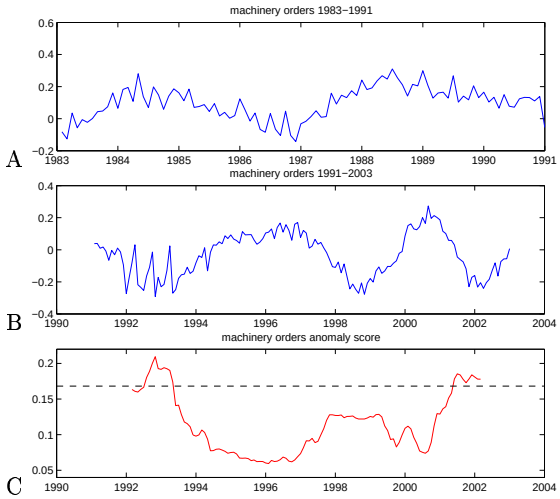

Figure 8: Machinery Orders: (A) reference data, (B) test data, (C) DARTS anomaly score.

5. LIMITATIONS AND FUTURE WORK

In our experiments we demonstrated that the DARTS framework was able to successfully detect anomalous regimes in multivariate data. However, the framework has several limitations and we enumerate them here.

First, DARTS relies on local models and to the extent that these models capture the important dependencies and invariants in the domain, DARTS should perform well. In this paper we explored autoregressive models and these seemed effective across a variety of problems, however there may

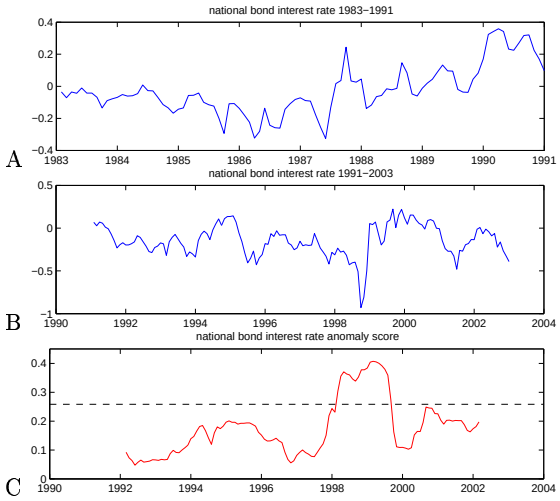

Figure 9: National Bond Index: (A) reference data, (B) test data, (C) DARTS anomaly score.

be domains for which they are inappropriate. Thus, we intend to explore other models such as state-space or linear differential equations.

Second, there is a trade-off between window size and sensitivity to the length of the anomaly region. Larger windows should result in better parameter estimates for the local models (and better density estimates), however larger windows also result in greater averaging and thus it may be difficult to identify small regions of anomalous behavior. Although our experiments indicate that DARTS can detect anomalous regions of very short durations, we have not systematically explored the tradeoff.

Third, although handling multivariate data is straightforward, the number of parameters can grow rapidly with additional variables and in consequence the window size will need to increase so that there is enough data to support the local models. To some extent, one can mitigate the sparsity of data with regularization; another alternative is to explore local models that use as few parameters as possible.

Fourth, the density estimate is created from local models with overlapping data. This potentially violates the independence assumption for points in the kernel density estimate and may lead to overly confident estimates. Clearly, one can enforce disjoint windows, but even this may not guarantee independence as time points are temporally linked.

Fifth, although the KD-tree is highly efficient in low dimensional spaces, its performance degrades as the dimensionality increases. In Section 3 we discussed some possible solutions and we have begun experimenting with dimensionality reduction techniques, such as PCA, which has worked well in initial experiments.

Finally, upon detecting an anomaly, a natural reaction is to explain why the anomaly occurred. In practice, we have found this can be very difficult in multivariate data. We believe one possible explanation method is to relate the con-

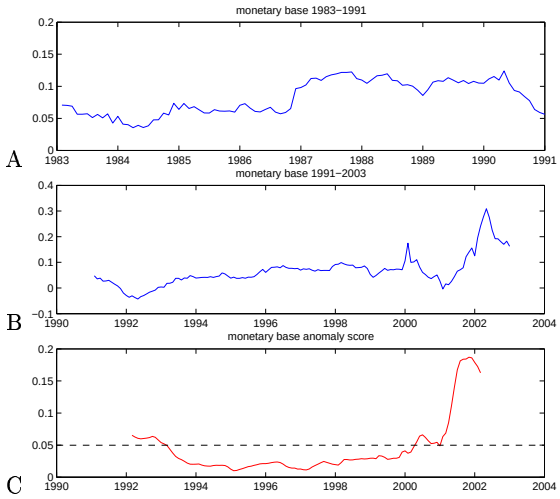

Figure 10: Monetary Base: (A) reference data, (B) test data, (C) DARTS anomaly score.

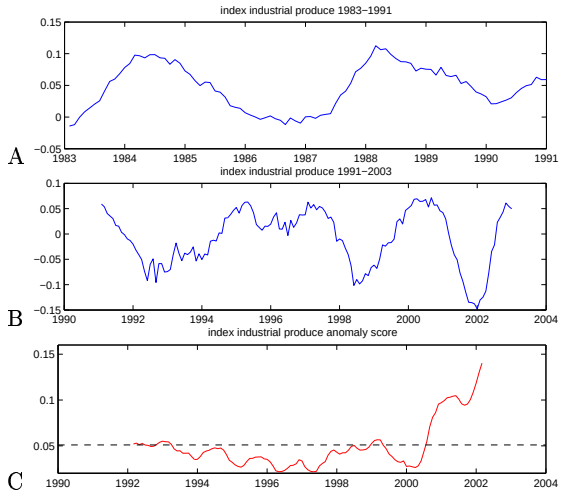

Figure 11: Index Industrial Produce: (A) reference data, (B) test data, (C) DARTS anomaly score.

tribution of the dimensions in the parameter space to the anomaly score.

6. RELATED WORK

Our work on detecting anomalous regimes falls into the broader class of finding irregularities in time-series data. In this section, we discuss this work in in order to position our contributions.

There has been much work on the topic of detecting outliers, which are points with extreme values. There are two broad classes of methods: limit checking and discrepancy checking. In limit checking one places a restriction on the maximum and minimum bounds of variable. For example, in Shewhart control charts the upper and lower limits are defined by a multiple of the process standard deviation. In discrepancy checking, one compares the observed time series to the predictions of a model and large deviations are flagged as outliers (e.g., [5, 30]). For example, Brutlag [5] applied a Holt-Winters forecasting model to detect network traffic anomalies and large prediction errors would trigger an alarm. The main drawback of discrepancy checking is that the prediction problem is extremely difficult.

In the context of autoregressive models, there have been a number of methods developed for detecting individual point outliers in time series as well as disturbances with longer term effects [6, 27]. Typically four types of outliers are considered: (1) additive outliers that represent a disturbance to a single point after which the series returns to it's normal values, (2) innovational outliers represent a disturbance to a single point whose effects persist, (3) temporary changes are disturbances with exponentially decaying effects, and (4) level shifts which are a permanent change in the values of the time series. These outliers are typically identified with a joint estimation procedure [6] where there is an explicit parametric model for the outlier. Detecting the outlier is an exercise in model selection.

Recently, several researchers have proposed methods for discovering unusual patterns of a variables expression. Keogh et al. [18] developed the TARZAN algorithm for enumerating surprising patterns in linear time. Their approach is based on discretizing the data and then using suffix trees to encode the frequency of all observed patterns (unobserved patterns estimated with a Markov model). TARZAN flags as surprising those patterns that occur significantly more or less frequently than expected. Dasgupta and Forrest presented IMM [7] which is based on the idea of generating "antibodies" which are pattern matchers specifically trained not to match the reference data but could potentially match novel sequences.

There has been a limited amount of work on finding anomalous regimes. As discussed earlier Gather et al. [11] presented an extension to Hotelling's T^2 statistic by incorporating time delays for monitoring online data. Smyth [26] proposed a hidden Markov model (HMM) to detect unknown states in an antenna monitoring problem. In his setup, the number of states of the HMM was considered known and then one additional state was added to cover all possible unknown states. For the known states, the observable conditional probabilities were modeled with a three component mixture of Gaussians (the structure was determined manually). Smyth's approach has similar goals but is not fully automated and may have scaling issues as the parameters are estimated with the EM algorithm.

Finally, we briefly mention that a related problem to finding irregularities in time series is tracking drifting concepts (e.g., [19]). Here the goal is to correctly classify objects with a static feature vector subject to the class description evolving over time.

7. CONCLUSIONS

We proposed the problem of detecting anomalous regimes, which are periods in time-series data where the observed system is being governed by a new and previously unseen set of relations between the variables. We presented a framework for finding anomalous regimes based on transforming

the time series to the parameter space of models learned on local windows of data.

Our framework was highly effective in determining anomalous regimes. In experiments, it achieved good sensitivity while maintaining selectivity on several real domains. Furthermore, the approach is efficient in important problem dimensions.

Acknowledgments

This work was supported by funding from NTT Communication Science Laboratories.

8. REFERENCES

[1] Daisy: Database of the identification of systems. http://www.esat.kuleuven.ac.be/sista/daisy.

[2] S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2003.

[3] S. D. Bay, D. Shapiro, and P. Langley. Revising engineering models: Combining computational discovery with knowledge. In *Proc. of the Thirteenth European Conf. on Machine Learning*, pages 10–22, 2002.

[4] J. L. Bentley. Multidimensional divide and conquer. *Communications of the ACM*, 23(4):214–229, 1980.

[5] J. D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *Proc. of the XIV Systems Administration Conf.*, pages 139–146, 2000.

[6] C. Chen and L. Liu. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88:284–297, 1993.

[7] D. Dasgupta and S. Forrest. Novelty detection in time series data using ideas from immunology. In *Proc. of the Int. Conf. on Intelligent Systems*, 1999.

[8] N. R. Draper and H. Smith. *Applied Regression Analysis*. John Wiley & Sons, second edition, 1981.

[9] E. Fix and J. L. Hodges. Discriminatory analysis: Nonparametric discrimination: Small sample performance. Technical Report Project 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolf Field, Texas, 1952.

[10] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software*, 3:209–226, 1977.

[11] U. Gather, R. Fried, V. Lanius, and M. Imhoff. Online monitoring of high dimensional physiological time series: A case-study. *Estadistica*, 53:259–298, 2001.

[12] A. G. Gray and A. W. Moore. Nonparametric density estimation: Toward computational tractability. In *SIAM Int. Conf. on Data Mining*, 2003.

[13] D. Hawkins. *Identification of outliers*. Chapman and Hall, 1980.

[14] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(3):55–67, 1970.

[15] A. E. Hoerl and R. W. Kennard. Ridge regression: iterative estimation of the biasing parameter. *Communications in Statistics*, 5:77–78, 1976.

[16] A. E. Hoerl, R. W. Kennard, and R. F. Baldwin. Ridge regression: Some simulations. *Communications in Statistics*, 4:105–123.

[17] H. Hotelling. Multivariate quality control. In C. Eisenhart, M. W. Hastay, and W. A. Wallis, editors, *Techniques of Statistical Analysis*. McGraw-Hill: New York, 1947.

[18] E. Keogh, S. Lonardi, and W. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 550–556, 2002.

[19] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: A new ensemble method for tracking concept drift. In *Proc. of the Third Int. IEEE Conf. on Data Mining*, pages 123–130, 2003.

[20] H. Lutkepohl. *Introduction to Multiple Time Series Analysis*. Spinger-Verlag, 1993.

[21] D. W. Marquardt and R. D. Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[22] A. Moore. An introductory tutorial on kd-trees. Technical Report Technical Report No. 209, University of Cambridge Computer Laboratory, 1991.

[23] A. W. Moore. The anchors hierarchy: Using the triangle inequality to survive high dimensional data. In *Proc. of the Twelfth Converence on Uncertainty in Artificial Intelligence*, pages 397–405, 2000.

[24] S. Sato. Stepwise prediction for economic time series by using vector autoregressive model. *Science of Modeling (AIC2003), ISM Report on Research and Eductation*, (17):225–233, 2003.

[25] J. P. Shaffer. Multiple hypothesis testing. *Annual Review Psychology*, 46:561–584, 1995.

[26] P. Smyth. Markov monitoring with unknown states. *IEEE Journal on Selected Areas in Communications*, 12(9):1600–1612, 1994.

[27] R. S. Tsay, D. Pena, and A. E. Pankratz. Outliers in multivariate time series. *Biometrika*, 87(4):789–804, 2000.

[28] W. H. Woodall and M. M. Ncube. Multivariate CUSUM quality control procedures. *Technometrics*, 1985(3):285–292, 1985.

[29] N. Ye, Q. Chen, S. M. Emran, and S. Vilbert. Hotellings T^2 multivariate profiling for anomaly detection. In *Proc. of the 2000 IEEE Workshop on Information Assurance and Security*, 2000.

[30] B.-K. Yee, N. D. Sidiropoulos, T. Johnson, H. V. Jagadish, C. Faloutsos, and A. Biliris. Online data mining for co-evolving time sequences. In *Proc. of the Int. Conf. on Data Enginerring*, 2000.