Searching through composite time series

Kaustav Das
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
kaustav@cs.cmu.edu

Andrew Moore
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
awm@cs.cmu.edu

ABSTRACT

Recently there has been much interest in detecting anomalies in both categorical and real valued time series data. Many classical statistical methods deal with univariate data, and less has been done about multivariate data. The drastic decrease of data storage costs, availability of cheap sensors along with automation of systems have resulted in proliferation of time series data. In most cases the data is multivariate in nature, and the effect of an anomaly can potentially be observed across more than one of these series. A traditional method to apply univariate methods in these cases would be to reduce the dimension using some dimensionality reduction technique (eg PCA). But, an anomaly detected in a weighted linear combination of the data might not be meaningful to the end user. A majority of such techniques suffers from the lack of user interpretability of the results. This motivates our approach of search through simple arithmetic combinations of time series. We compare the performance of our algorithm with related methods such as Vector Autoregression on semi-synthetic health data.

1. INTRODUCTION

Automatic surveillance systems are becoming more popular and are increasingly using data mining methods to perform detection. The observation of industrial manufacturing processes is one traditional application of these systems. Another application is public health monitoring, which has the goal of detecting new disease outbreaks as early as possible. Searching for terrorist activity or attacks is also becoming important. Applications in that area include monitoring human health and behavioral data to detect a chemical or biological attack, or searching for signs of radiation to detect development or deployment of nuclear devices. The RODS lab at the University of Pittsburgh (see www.health.pitt.edu/rods/) is focused both on public health monitoring and detection of biological attacks. This paper is based on our work in the RODS lab and thus focuses on these applications, but the algorithms we present are not specific to them. They are appropriate for a variety of monitoring tasks.

Modern surveillance systems are characterized by the need to analyze many variables simultaneously. Because of this fact, the traditional method of setting upper and lower bounds for a single variable are no longer appropriate. Data mining methods are used that must address the complex interactions between variables, the dangers of multiple hypothesis testing, and the computational issues caused by large data sets. See [15] for an overview of detection methods.

We consider the problem of detecting an anomalous increase of values in multivariate time series data. The problem stems from the fact that the increase can be spread over multiple variables. As an example consider the time series of counts of patients visiting emergency departments every day. For each possible symptom we have a corresponding time series. A particular disease such as an influenza outbreak, will affect the count of multiple syndromes. In this case, we need to simultaneously consider all the variables to detect the presence of an anomaly. We are concerned with prospective surveillance, where we need to detect a disease outbreak as soon as possible.

To combine information from multiple time series we examine a novel technique which is simple but powerful. Composite time series are constructed by simple addition and subtraction of the individual time series. We search through all possible composite time series for an anomaly. Using just simple arithmetic operations like addition and subtraction provides an easy physical interpretation of the composite series. It is also able to detect anomalies sooner than other traditional methods.

2. RELATED METHODS

In this section we describe various multivariate techniques that can detect a shift in the data.

2.1 Vector Auto Regression

The time series is modeled as a standard VAR(p) model [1]. Let the number of variables be n. Let $\mathbf{X_t}$ denote the $(n \times 1)$ vector of values at time t.

$$\mathbf{X_t} = \mathbf{C} + \sum_{i=1}^{p} \mathbf{\Phi_i X_{t-i}} + \epsilon_\mathbf{t}$$

where, $\mathbf{C}$ denotes an $(n \times 1)$ vector of constants, $\mathbf{\Phi_i}$ are $(m \times m)$ coefficient matrices and the $(n \times 1)$ vector $\epsilon_\mathbf{t}$ is the residual vector. Here $E[\epsilon_\mathbf{t}] = 0$. The coefficients $\mathbf{\Phi_i}$ can be estimated from data using ordinary least squares (OLS) linear regression.

The expected value of $\mathbf{X_t}$ given the past p days' data is given by

$$E[\mathbf{X_t}] = \mathbf{C} + \sum_{i=1}^{p} \mathbf{\Phi_i X_{t-i}}$$

At each time step, we compare the actual and expected values of $\mathbf{X_t}$. We signal an alarm when $\mathbf{X_t}$ deviates significantly from $E[\mathbf{X_t}]$. Quantitatively, we compute the Mahalanobis distance:

$$D^2 = (\mathbf{X_t} - E[\mathbf{X_t}])^T \mathbf{\Sigma}^{-1} (\mathbf{X_t} - E[\mathbf{X_t}])$$

where, $\mathbf{\Sigma}$ is the sample variance-covariance matrix for the past p days' data.

An alarm is signaled when D exceeds a threshold h. Here h is the parameter which controls the number of false positives.

2.2 Vector Moving Average

This method is a special case of the Vector Autoregression as described above. We assume that the expected value of $\mathbf{X_t}$ is the mean of the past p days' values.

$$E[\mathbf{X_t}] = \frac{1}{p} \sum_{i=1}^{p} \mathbf{X_{t-i}}$$

We compute the Mahalanobis distance as mentioned previously, and signal an alarm when $D > h$.

2.3 Hotelling T^2 Test

We model the distribution of the mean of the recent p days' data. Let

$$\mathbf{\bar{X}} = \frac{1}{p} \sum_{i=0}^{p-1} \mathbf{X_{t-i}}$$

and $\mathbf{\Sigma}$ be the sample variance-covariance matrix for the past p days' data.

The statistic T^2 is defined as [2]:

$$T^2 = n(\mathbf{\bar{X}} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{\bar{X}} - \mu)$$

T^2 is distributed as $\frac{p(n-1)}{n-p} F_{(p,n-p)}$, with $F_{(p,n-p)}$ representing the F distribution with p and $n-p$ degrees of freedom. We signal an alarm when $P(x \geq T^2) < \alpha$, where α controls the rate of false positives. Application of Hotelling T^2 in multivariate quality control has been investigated in [3].

3. DETECTION METHOD: CUSUM

Before presenting our algorithm, we describe a popular method used in detecting anomalies in time series. CUSUM was originally developed to detect changes in the quality of output of continuous production process. It can quickly detect a shift in the mean of a process. As the name suggests, CUSUM maintains a cumulative sum of deviations from a reference value r. Let us consider a time series where at time t we have measurement $X(t)$. The one-sided CUSUM calculation is as follows:

$$C(0) = 0 \tag{1}$$

$$C(t) = max(0, X(t) - (\mu_0 + L) + C(t-1)) \tag{2}$$

μ_0 is the in-control process mean. From the equations above, if the X_m values are close to the mean, then the $C(t)$ values will be some small value. However once a positive shift from the mean occurs, the $C(t)$ value will increase rapidly. L is known as the slack value or allowance. In the equation above, any values within L units of μ_0 will be effectively ignored. The allowance L is usually set to be the midpoint between the in-control process mean μ_0 and the out-of-control process mean μ_1.

Alerts are raised whenever $C(t)$ exceeds a threshold decision interval H. The cumulative sum is then reset to zero. The Average Run Length (ARL) is controlled by this parameter. The ARL is the average number of time steps before an alert is raised.

The CUSUM algorithm described here has been extensively used in biosurveillance systems. It has been used for influenza surveillance [14], detection of salmonella outbreaks [11] and in the Early Aberration Reporting system [10]. CUSUM algorithms have also been extended to incorporate spatial information such as [12] and [13].

3.1 Modified CUSUM

In this work we use a modified CUSUM as the detection method. We have found this method to be very effective in detecting upward shifts in time series.

We calculate the cumulative sum of deviation similar to equation 2. Instead of maintaining the cumulant starting at t=0, we consider only the last CW (Cumulant Window) number of time steps. This means that the current Cumulant at time t, will be independent of any data before the time T - CW. We signal an alarm if the current cumulant value is greater than H. This modification does not affect the performance of the algorithm significantly, and is actually desired in our case, as explained later. This also allows us to speed up the computation as described in section 7.

In the original algorithm, H is usually taken as a fixed threshold value. We have set H = hσ, a multiple of the standard deviation σ of the time series. We need to calculate and update the σ value at each time step. In our method σ is the sample standard deviation of the series calculated over a sliding window of the last N days. Thus, H is dynamically updated based on the behavior of the variable. Also, since we do not know the out-of-control process mean μ_1, we set $L = l\sigma$, for some constant l. L too gets updated at each time step. The in-control process mean μ_0 is taken as the moving average over the last N days. This dynamic updation of the parameters at each time step is a significant modification of the original CUSUM algorithm. This allows us to model non-stationary time series variables.

3.2 Multivariate CUSUM

An analogous Multivariate version has also been applied to surveillance data. Crosier's multivariate cumulative sum (MCUSUM) method [5] has been applied to syndromic data from multiple hospitals [6] and Pignatiello's MCUSUM [7] applied to yearly, spatially distributed counts of breast cancer incidence [8]. We have implemented the MCUSM method from [7] and compared it against our method.

4. PROPOSED METHOD: PARALLEL MONITORING OF COMPOSITE SERIES

A common feature of all the multivariate methods is that the statistic on which the alarm is set, does not have an intutive physical interpretation in terms of the variables. However, if we monitor the individual variables in parallel, we can identify the variable that has an anomalous behavior in case of an alarm.

As mentioned in [4] these multivariate methods are 'omnidirectional, a property that can be useful in detecting an earlier signal, but can also cause false alerts if a change in the covariance matrix occurs that is irrelevant to any outbreak signal of interest'. They do not specifically check for increases in individual series. In our experiments this causes them to perform worse than parallel monitoring of univariate series.

The novel method that we suggest involves parallel monitoring of not only the individual variables, but also simple arithmetic combinations of them. This retains the advantage of easy interpretability while giving a better performance as shown in our experiments. This method of using combinations of time series is orthogonal to the univariate detection method used to monitor each series. We have chosen CUSUM as the detection algorithm because of its superior and robust performance in detecting slight increases over the normal value. In the following sections we describe this algorithm in more detail.

5. SEARCH SPACE

As mentioned previously, we perform a parallel monitoring of the time series variables and arithmetic combinations of them. Here we describe the composite series that are monitored in parallel for any increase from expected values.

Let X_1, ..., X_k be k random time series variables, and $X_i(t)$ denote the value of X_i at time step t.

Addition: We create time series of the form:

$$Y = X_{i_1} + X_{i_2} + \ldots + X_{i_m};\ i_1, \ldots, i_m \in \{1, 2, \ldots, k\}$$

This means that at each time step t,

$$Y(t) = X_{i_1}(t) + X_{i_2}(t) + \ldots + X_{i_m}(t);\ i_1, \ldots, i_m \in \{1, 2, \ldots, k\}$$

Here we can choose the indices $i_1, i_2, \ldots, i_m$ in $\binom{k}{m}$ ways. If we consider summations of up to k terms, the total number of such composite series = $\binom{k}{1} + \binom{k}{2} + \ldots + \binom{k}{m}$.

Similar to addition, we create time series of the form

$$Y = X_{i_1} - X_{i_2};\ i_1, i_2 \in \{1, \ldots, k\}$$

Here we consider combination of just 2 series. There are $\binom{n}{2}$ such composite series.

Motivation of the addition and subtraction operations:

1. Addition: We assume that an outbreak simultaneously causes an increase in the value of more than one variable. The detection accuracy of any anomaly detection method will depend on the signal to noise ratio (SNR) of the outbreak. The anomalous increase in the value is the signal we want to detect, and the standard deviation of the variable is the noise. Here we describe a situation where the composite additive series will have a better SNR than any of the individual series. Consider two random time series variables X_1 and X_2. Assume that they have equal standard deviations, $\sigma_{X_1} = \sigma_{X_2} = \sigma$. Let a be the actual anomalous increase in the values of X_1 and X_2.

   Let $Y = X_1 + X_2$. Now,

   $$\sigma_Y^2 = \sigma_{X1}^2 + \sigma_{X2}^2 - 2 * r * \sigma_{X1}\sigma_{X2} = 2 * \sigma(1 - r)$$

   where r is the Pearson correlation coefficient between X_1 and X_2. By definition, r≥-1. Hence, $\sigma_Y \leq$ 2*σ. The SNR of the individual variables is $\frac{a}{\sigma}$. The SNR of the composite series Y is $\frac{2a}{\sigma_Y} \leq \frac{a}{\sigma}$.

   We note that if there is a very strong positive correlation between the variables, then the noise(variance) will increase proportional to the signal (outbreak). In these cases, the false positive rate will increase because of multiple hypothesis testing. Hence in those situations, considering summation of series can give worse results.

2. Subtraction:

   Considering series of the form $Y = X_1 - X_2$ can be helpful if there is some positive correlation between X_1 and X_2. If these two random variables are positively correlated, then any anomalous increase present in X_1, but not in X_2, will be more pronounced in Y. This is because the noise will tend to cancel, whereas the signal will be left unaffected. The increase of false positive rate due to multiple hypothesis testing also applies in this case. Hence we expect an improvement using the subtraction operator only when there is a high positive correlation among the variables.

6. OUTBREAK SIMULATION

Because there were no known outbreaks in our datasets, we assumed artificial outbreaks by adding ramp increases. We call these outbreaks as attacks, since one of the motivations of this work is to detect bioterrorist attacks.

$$\begin{aligned} attack(t) &= atttack\_height * \frac{(t - t_{start})}{(t_{start} - t_{end})}; \\ &\quad \text{for } t_{start} \geq t \geq t_{end} \\ &= 0 \text{ otherwise} \end{aligned} \tag{3}$$

The attacks are spread through more than one time series. We randomly choose m of the k time series to add an attack. We choose m random weights $w_1, \ldots, w_m$ uniformly from the set $\{(w_1, w_2, ..., w_m) | 0 \leq w_i \leq 1, \Sigma w_i = 1\}$. We then add a weighted attack to each of these m time series:

$$X_i^{attack}(t) = X_i(t) + w_i * attack(t); \text{for i = 1, ..., m}$$

We spread the attack to more than one time series so that it becomes difficult to detect it from any individual variable. The effect of attack becomes more evident when we combine more than one variable.

7. SEARCH ALGORITHM

As mentioned in section 5, the number of composite time series can be very large. Let C_i denote the C value of the composite time series TS_i. One approach to monitor all these series individually would be to store the C_i values corresponding to each of these series and update them at each time step. At each time step, we signal an alarm if the C_i value of any of the composite time series exceeds the corresponding hσ_i. We also need to store and update each σ_i value at each time step.

Let m be the maximum number of individual series in a composite series. In cases where k is large this method will require an exponential amount of memory depending on m. We now describe a branch and bound approach that does not require us to store all the C_i and σ_i values.

The main idea is to determine whether a composite series can possibly signal alert without explicitly calculating the C_i value. If we are able to eliminate a majority of the series by using an appropriate bound, then we need only calculate the C_i and σ_i values only for a small fraction of them.

First we note that at a particular time step, if $X_i(t)$ - (μ_0^i + K) < 0, then we can ignore the composite series i. This is because at this time step, the C_i value will decrease, and it cannot signal a new alert. μ_0^i is taken as the moving average of X_i over a past window of N days. σ_i^2 is calculated as the sample variance of the last N days. For simplicity we assume that the mean μ_0^i has been subtracted from X_i for each i, as a preprocessing step. We have fixed N = 21 days in all our experiments.

7.1 Searching through the additive space

We search through additions of all possible combinations of m time series from the k series. The search is done in a depth first manner. We find a lower bound on the standard deviation of the sum of two random variables. Let X_1 and X_2 be two random variables, and σ_{X_1} and σ_{X_2} be the corresponding standard deviations. Let Y = X_1 + X_2. The standard deviation of Y is given by σ_Y^2 = $\sigma_{X_1}^2$ + $\sigma_{X_2}^2$ - 2*r*σ_{X_1} σ_{X_2}. Here r is the Pearson correlation coefficient of X_1 and X_2. We can obtain a lower bound for σ_Y when r=-1. We can better this bound if we can assume that r is lower bounded by a higher value.

Now, let $\hat{\sigma}_{X_1} \leq \sigma_{X_1}$ and $\hat{\sigma}_{X_2} \leq \sigma_{X_2}$, where $\sigma\hat{_{X_1}}$ and $\hat{\sigma}_{X_2}$ are lower bounds on the standard deviation of X_1 and X_2. Let $\hat{r}$ be a lower bound on the correlation coefficient of X_1 and X_2. Define, $\hat{\sigma}_Y^2$ = $\hat{\sigma}_{X_1}^2$ + $\hat{\sigma}_{X_2}^2$ - 2*$\hat{r}$*$\hat{\sigma}_{X_1}$ $\hat{\sigma}_{X_2}$. Under these assumptions it can be shown that $\hat{\sigma}_Y^2 \leq \sigma_Y^2$, ie $\hat{\sigma}_Y$ gives a lower bound on the standard deviation of $Y = X_1 + X_2$.

Our depth first search algorithm is as follows. We describe our search algorithm as a recursion:

```
For each time step t:

Initialize

  1. Update the standard deviations $\sigma_1$, $\sigma_1$, ..., .$\sigma_k$.
  2. S ← $\phi$.
  3. DfsRecur(S,0)

DfsRecur(S, $\hat{\sigma}_{X_S}$)

  1. Let $max\_index$ = the maximum index number among the series present in S.
  2. $X_S = X_{i_1}$ + ... + $X_{i_p}$, where $X_{i_1}$ , ... , $X_{i_p}$ ∈ S
  3. If $X_S \leq h\hat{\sigma}_{X_S}$, then goto step 8
  4. Calculate the value of $\sigma_{X_S}$. This step requires O(N) time, where N is the moving-average window size.
  5. If $X_S \leq h\sigma_{X_S}$, then goto step 8
  6. Calculate the value of $C_S$, the cumulative sum for the composite series S. We need only consider $CW$ days in the past to calculate this value.
       $C(0) = 0$                                                      (4)
       $C(i) = max(0, X_S(t - CW + i) - (\mu_0 + L) + C(i-1))$,
              for i = 1 to CW                                          (5)
       $C_S = C(CW)$                                                   (6)
     If $C_S \geq h\sigma_{X_S}$, then signal an alert.
  7. If | S | = m, return
  8. For each i such that $max\_index < i \leq k$
     (a) $S' = S \cup X_i$
     (b) if $|S'| > m$ then return
     (c) $\hat{\sigma}_{X_{S'}}$ = sqrt($\hat{\sigma}_{X_S}^2$ + $\hat{\sigma}_i^2$ - 2$\hat{r}\hat{\sigma}_{X_S}$ $\hat{\sigma}_i$).
     (d) $DfsRecur(S', \hat{\sigma}_{X_{S'}})$
```

Here m is the maximum number of series that are considered in one composite series X_S. It first calculates a lower bound of the standard deviation of a composite series without explicitly calculating it from the past data. This lower bound allows us to determine if the current value of the composite series can possibly signal an alert. We can avoid calculating the exact standard deviation and cumulative sum by this bounding procedure. In a fraction of cases we actually need to perform the exact calculations.

7.2 Searching through difference series

```
  1. For each i = 1 to k:
     (a) For the time series $X_i$ find the corresponding series $CS_i$ that is most correlated with it.
     (b) Create random variable $D_i = X_i - CS_i$.
     (c) Update $\sigma_i^D$ and $C_i^D$.
     (d) If $C_i^D \geq h\sigma_i^D$, then signal an alert
```

8. DATASETS

We use three datasets in our experiments.

1. Over the Counter Sales Data (OTC) in US. Each sale belongs to one of the following categories:
   (a) Baby/Child Electrolytes
   (b) Cough/Cold
   (c) Internal Analgesics
   (d) Stomach Remedies
   (e) Thermometers

   We have 5 time series corresponding to each of the above categories for a period of about 2 years.

2. Emergency department dataset from the regions around Pittsburgh. The data spans 668 days.

   It has the following attributes:

   (a) ADMIT_DAY_INDEX: Date on which the patient was admitted.

(b) PRODROME: The main category of the patient's complaint upon arrival at the emergency department. It can have 7 possible values. We get 7 time series of the count of patients each day.

3. Stock Prices Dataset: We consider the daily stock prices of the following 12 companies: Dell, Sun, GE, IBM, Microsoft, GM, Nissan, Toyota, Sony, Ford, BP and Exxon Mobil for a period of 4 years.

9. RESULTS

To measure the performance of the algorithms, we need to measure their false positive rate and the corresponding detection lag. Detection lag is the time difference between the start of the attack and the first instance when an alert is signaled with the attack underway. A plot of the number of false positives vs the detection lag is called an AMOC (Activity Monitoring and Control Chart) curve.

To get a point on the AMOC curve we do the following:

1. Fix a value of h, where, $H = h\sigma$, is the CUSUM threshold.
2. For i = 1 to 50,
   (a) Inject a random attack of duration 15 days in the data. The attack is spread over at most three individual variables.
   (b) Estimate the baseline trend values using Moving Average with a slide window of length 21 days.
   (c) Run the modified CUSUM algorithm on the residues. Keep track of the number of false positives and the detection lag. If no alert is signaled within the duration of the attack, the detection lag is taken as the duration of attack.
3. Calculate the average number of false positives and the average detection lag over the 50 random attack simulations.

This gives us a point on the AMOC curve. We then vary h to obtain the entire curve.

We ran our algorithm on each dataset, with different values of m (the maximum number of series in a composite series). We compared the CUSUM algorithm with VAR, Vector Moving Average, Hotelling T^2 and MCUSUM. Both VAR and Vector Moving Average used a 3-day slide window (p=3). Hotelling T^2 used the last 10 day's values for calculating the mean.

9.1 OTC Dataset

Fig 1 shows the comparison between CUSUM and the other related methods as explained in section 2 for the OTC dataset. We run CUSUM on the individual series independently for the Simple CUSUM method (m=1). We see that CUSUM significantly outperforms the other methods. For the same False Positive rate, it gives a much lower Detection Lag.

Fig 2 shows the curves for CUSUM where m varies as 1, 2 and 3. The fourth curve corresponds to considering the difference series as explained in section 7. We see that there is an improvement in the detection lag time when we consider summation of two or more series. The performance of the two series and three series algorithms are similar. But the

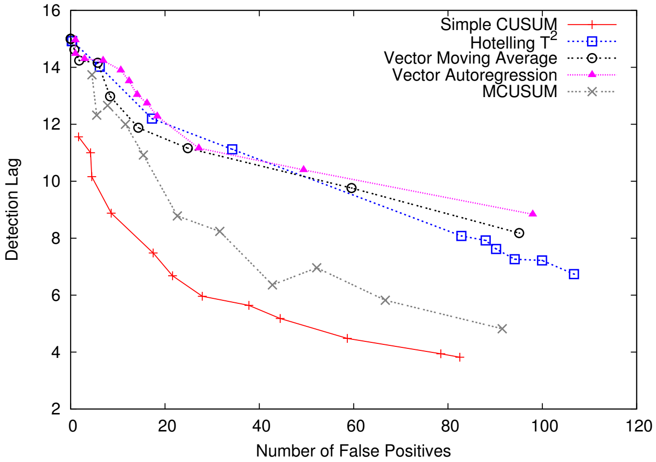

Figure 1: OTC dataset: AMOC Curves comparing Related Methods

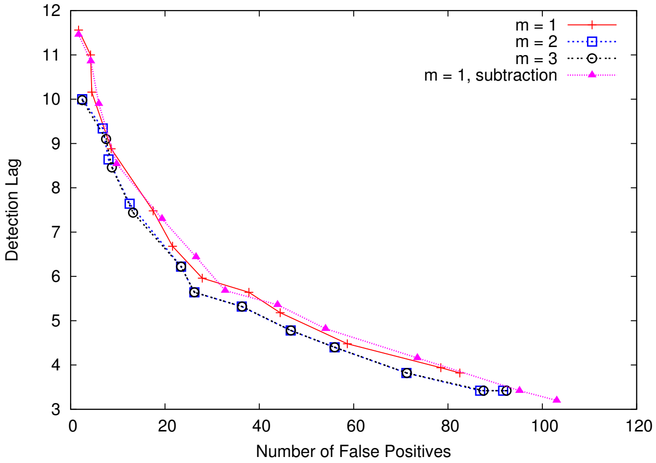

Figure 2: OTC dataset: AMOC Curves comparing different combinations of time series

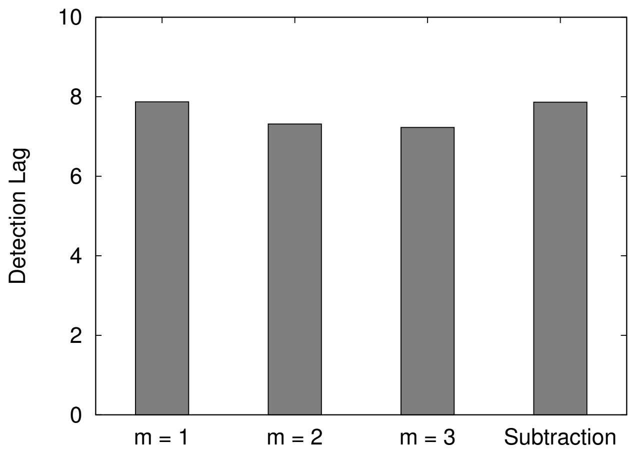

Figure 3: OTC dataset: Improvement in Detection lag using the proposed methods corresponding to 15 false positives over the duration

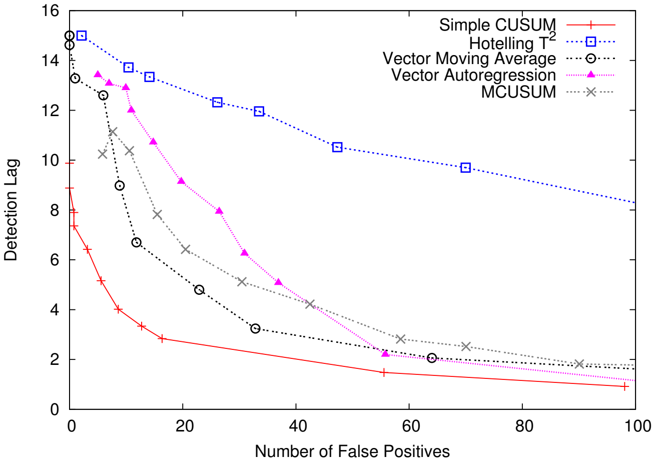

Figure 4: Emergency Department dataset: AMOC Curves comparing Related Methods

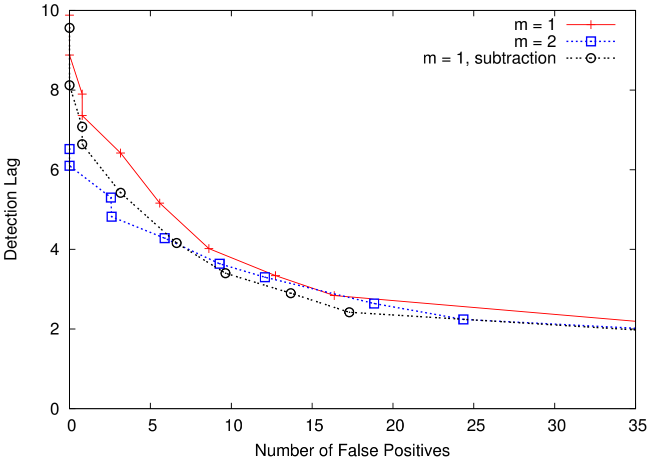

Figure 5: Emergency Department dataset: AMOC Curves comparing different combinations of time series

difference operation does not seem to give an improvement. For a fixed false positive rate of 15 for the entire duration, Fig 3 shows the corresponding Detection Lags. The detection lag is 7.87 days for m = 1. It improves by about 8% to 7.23 days for m = 3.

9.2 Emergency Department Dataset

Fig 4 shows the comparison between CUSUM and the other related methods. Similar to the OTC dataset, we see that CUSUM significantly outperforms the other methods.

The AMOC curves for this dataset are shown in Fig 5. There is a significant difference in the detection lag time for very low (<10) false positive rate. For example, for no false positives over the entire duration, the detection lags are 8.88, 6.1 and 6.46, for m = 1, 2 and the difference operator respectively. This is illustrated in the bar chart Fig 6. We see an improvement of 2.78 days or 31% in detection lag when considering more than one series. In applications such as disease outbreak detection, we need to have a low false positive rate. Having a high false positive rate makes the system almost useless because it becomes infeasible to investigate each alarm. Hence, our result in the low false positive range is significant.

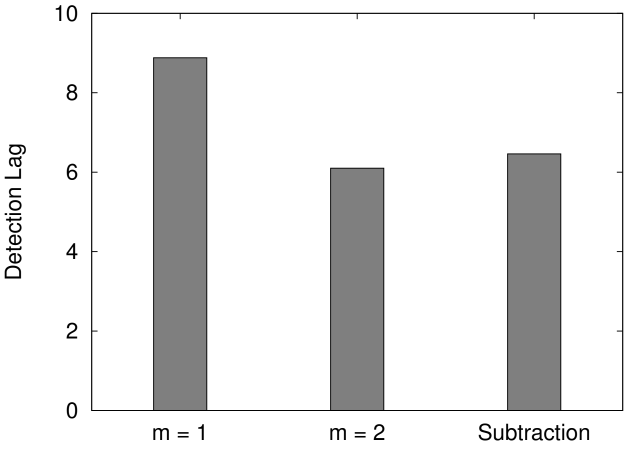

Figure 6: Emergency Department dataset: Improvement in Detection lag using the proposed methods corresponding to no false positives over the duration

9.2.1 *Effect of Cusum Window(CW)*

CW denotes the number of previous days that are considered to calculate the cumulative for the current day. In eqn 5, when $CW = 1$ and L = 0, C_S measures the deviation of the current value from the expected mean. The CUSUM test in this case becomes identical to the one sample Gaussian test (computing the p-value of a sample). In our experiments, we have set L $= \sigma$, which empirically give the best results. Hence for $CW = 1$, our test is similar to the simple Gaussian test, except for the effect of L. L defines a threshold such that we are concerned only about increases that are above that threshold.

Another advantage of CUSUM over the Gaussian testing is that it considers samples from CW past days. If there is a gradual increase in the time series, it can utilize past information to make a better decision. It can be expected that higher CW values will be helpful when the expected detection lag is long. But if the expected detection lag is close to one day, then higher CW values won't be helpful. This is because in this case the attack mostly gets detected on the first day, and the data from previous days do not provide any helpful information.

Fig 7 shows the AMOC curves for m=1 (considering individual series), with different values of CW. We see that for large (>70) false positive rate, $CW = 1$ performs best. But, portion of the curves that correspond to lower false positive rates show that higher CW values perform better. Most applications in practice, including disease detection require a very low false positive rate. Hence having a larger CW value is preferable in these conditions.

9.2.2 *Computational Speedup*

Table 1 gives an indication of the advantage of using a lower bound on the standard deviation of the composite series. The first column 'Num Series Considered' corresponds to the number of composite time series that are tested for

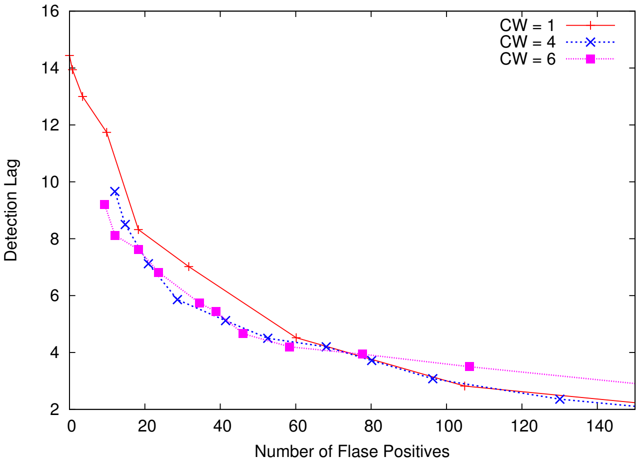

Figure 7: Plot of Detection Time vs False Positives for ED dataset with varying *CW*

Table 1: Number of instances that required exact calculation of σ in the Emergency Department Dataset

	Num Series considered	Num Calculated
m = 2	93,923	886
m = 3	428,571	5,587

anomaly over the entire time period. The column 'Num Calculated' corresponds to the cases where we actually needed to perform the exact computation of σ. We see that for m = 2 and 3, we need to perform the expensive computation of σ in only a small fraction of the cases considered.

9.3 Stock Prices Dataset

The AMOC curves for this dataset are shown in Fig 8. We see that m=2 and 3 performs similar or worse than m=1. This is not very surprising since there is a high positive correlation between the variables. As noted earlier, in presence of positive correlation, considering summation two or more series can cause the false positive rate to increase without producing a significant decrease in the detection lag. We see that in this case, when we consider the difference operator, the AMOC curve is significantly better. This shows that the difference operator is able to exploit the positive correlation present in the dataset.

10. FUTURE WORK

Apart from using addition and subtraction, other arithmetic operations such as division can be used to create composite series. We will need to find an efficient way to compute the standard deviation of the composite series since the combinations would no longer be linear.

A main advantage of our method is the easy interpretability of an alert. But, not all combinations of time series are meaningful to the end user. We can have an user interface that can specify which combinations to consider. Alternatively it might be possible to learn meaningful combinations through a more interactive system.

11. REFERENCES

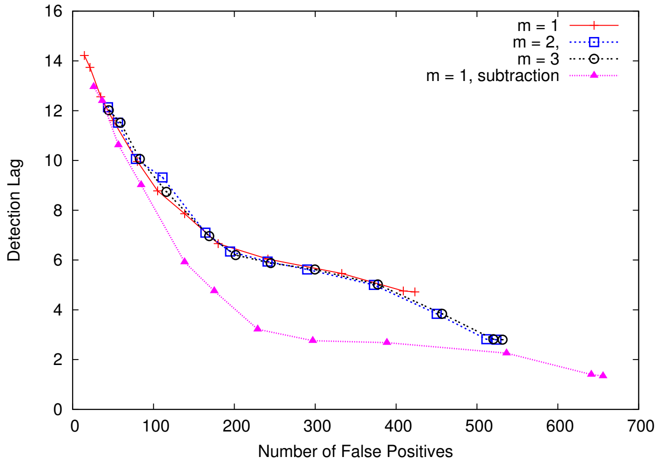

Figure 8: Plot of Detection Time vs False Positives for Stock Prices dataset

[1] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, p257-350, 1994.

[2] H. Hotelling. Multivariate Quality Control. In: C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. *Techniques of Statistical Analysis*. New York: McGraw-Hill, p111-184, 1947.

[3] B. Hong, M. Hardin. A report of the properties of the multivariate forecast-based processing scheme. *In: Proceedings of the Joint Statistical Meetings* Toronto, Canada: American Statistical Association; August 2004.

[4] H. Burkom, J. Coberly, S. Murphy, Y. Elbert, K. Hurt-Mullen. Public Health Monitoring Tools for Multiple Data Streams. *In: Proceedings of the 2004 National Syndromic Surveillance Conference* [CD-ROM] Boston, MA, 2004.

[5] R. B. Crosier. Multivariate generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* 30:291–303, 1988.

[6] M. Stoto. Multivariate methods for aberration detection: a simulation report using the District of Columbia's syndromic surveillance data. *In: Proceedings of the 2004 National Syndromic Surveillance Conference* [Oral Presentation] Boston, MA, 2004.

[7] J. J. Pignatiello , G. C. Runger. Comparisons of multivariate CUSUM charts. *J Qual Technol*, 22:173–86, 1990.

[8] P. A. Rogerson, I. Yamada. Monitoring change in spatial patterns of disease: comparing univariate and multivariate Cumulative Sum Approaches. *Stat Med*, 23:2195–214, 2004.

[9] D. Hawkins. Multivariate quality control based on regression-adjusted variables. *Technometrics*, 33:61–75, 1991.

[10] L. Hutwagner, W. Thompspn, G. M. Seeman, and T. Treadwell. The bioterrorism preparedness and response early aberration reporting system(ears). *Journal of Urban Health*, 80:i89–i96, 2003.

[11] L. C. Hutwagner, E. Maloney, N. H. Bean, L. Slutsker, and S. Martin. Using laboratory-based surveillance

data for prevention: An algorithm for detecting salmonella outbreaks. *Emerging Infectious Diseases*, 3:395–400, 1997.

[12] R. F. Raubertas. An analysis of disease surveillance data that uses the geographic locations of reporting units. *Statistics in Medicine*, 8:267–271, 1989.

[13] P. A. Rogerson. Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine*, 16:2081–2093, 1997.

[14] H. E. Tillett and I. L. Spencer. Influenza surveillance in England and Wales using routine statistics. *Journal of Hygine*, 88:83–94, 1982.

[15] W. K. Wong. *Data Mining for Early Disease Outbreak Detection*. PhD thesis, Carnegie Mellon University, 2004.