Using Labeled Data to Evaluate Change Detectors in a Multivariate Streaming Environment

Albert Y. Kim,[1] Caren Marzban,[1] Donald B. Percival[2] and Werner Stuetzle[1]

Technical Report no. 534

Department of Statistics, University of Washington

Abstract

We consider the problem of detecting changes in a multivariate data stream. A change detector is defined by a detection algorithm and an alarm threshold. A detection algorithm maps the stream of input vectors into a univariate detection stream. The detector signals a change when the detection stream exceeds the chosen alarm threshold. We consider two aspects of the problem: (1) setting the alarm threshold and (2) measuring/comparing the performance of detection algorithms. We assume we are given a segment of the stream where changes of interest are marked. We present evidence that, without such marked training data, it might not be possible to accurately estimate the false alarm rate for a given alarm threshold. Commonly used approaches assume the data stream consists of independent observations, an implausible assumption given the time series nature of the data. Lack of independence can lead to estimates that are badly biased. Marked training data can also be used for realistic comparison of detection algorithms. We define a version of the receiver operating characteristic curve adapted to the change detection problem and propose a block bootstrap for comparing such curves. We illustrate the proposed methodology using multivariate data derived from an image stream.

Key words: Block bootstrap; Change point detection; Time series analysis

1 Introduction

We consider the problem of detecting changes in a multivariate data stream. We want to assess whether the most recently observed data vectors (the "current set") differ in some significant manner from previously observed vectors (the "reference set"). Change detection is of interest in a number of applications, including neuroscience [3], surveillance [7] and seismology [16] (see also [14, 15] and references therein for other applications).

The notion of change is often formalized in terms of distributions: vectors in the current set are assumed to be sampled from some multivariate distribution Q, whereas those in the reference set are assumed to come from a (possibly different) distribution P. The task of

[1]Department of Statistics, University of Washington

[2]Applied Physics Laboratory, University of Washington

a change detector then is to test the hypothesis $P = Q$ given the two samples. We obtain a new value of the test statistic every time a new observation arrives. We flag a change as soon as the test statistic exceeds a chosen alarm threshold [8, 10, 12].

In a concrete application of this recipe we face a number of choices. We have to pick a two-sample test that is sensitive toward changes of interest; we have to choose the sizes of the current and reference sets; and we have to choose an alarm threshold that results in the desired tradeoff between false alarms and missed changes.

More complicated schemes are possible. We can use multiple two-sample tests and multiple sizes of current and reference sets in parallel and summarize the resulting values of the test statistics. We can even adopt a more complex notion of "change". No matter what the details, ultimately we will end up with a univariate stream that we call the "detection stream". We flag a change whenever the detection stream exceeds a chosen alarm threshold. Abstracting away details, a change detector can be defined as a combination of a detection algorithm mapping the multivariate input stream $\mathbf{x}_t$ into a univariate detection stream d_t, and an alarm threshold τ. The only fundamental restriction is that d_t can only depend on input observed up to time t.

In this paper we focus on two problems: (i) choosing between different detection algorithms; and (ii) selecting an alarm threshold to obtain a desired false alarm rate. We assume that we have labeled training data, i.e., a segment of the stream where changes of interest have been marked. To quantify the performance of a detection algorithm, we propose an adaptation of the standard receiver operating characteristic (ROC) curve (Section 3). A resampling method similar to the block bootstrap lets us compare the ROC curves of different detection algorithms on the labeled data in a statistically meaningful way (Section 5). The labeled data also allow us to determine the alarm threshold for a desired false alarm rate without the usual assumption that vectors in the stream are observations of independent random variables. The independence assumption seems implausible when we are observing a time series. If the assumption is violated, estimates of the false alarm rate based on this assumption can be wildly off the mark (Section 4). We illustrate our main points using a multivariate data stream derived from a series of images of Portage Bay in Seattle (Sections 2 and 6). Section 7 with a summary and some ideas for future work concludes the paper.

2 Data

To illustrate the ideas in this paper, we created a multivariate data stream from a sequence of images recorded with a web camera operated by the Sound Recording for Education (SORFED) project at the Applied Physics Laboratory, University of Washington. The camera is mounted on a barge several feet above the water of Portage Bay, Seattle, and monitors natural (e.g., rain and wind) and man-made (e.g., boats) processes that can cause sound in the water. It takes images at two second intervals (usually). We use 5002 images recorded on June 27, 2007. To eliminate a bridge and a portion of the sky with little activity, we crop the tops of the images, leaving us with a sequence of 168×280 pixel images focused on the water of Portage Bay. We divide the pixels in each (cropped) image into a 14×20 grid of bins, with each of the 280 bins containing $12 \times 14 = 168$ pixels. We summarize each of the 280 bins of an image by its average grey level, resulting in a stream of 280-dimensional

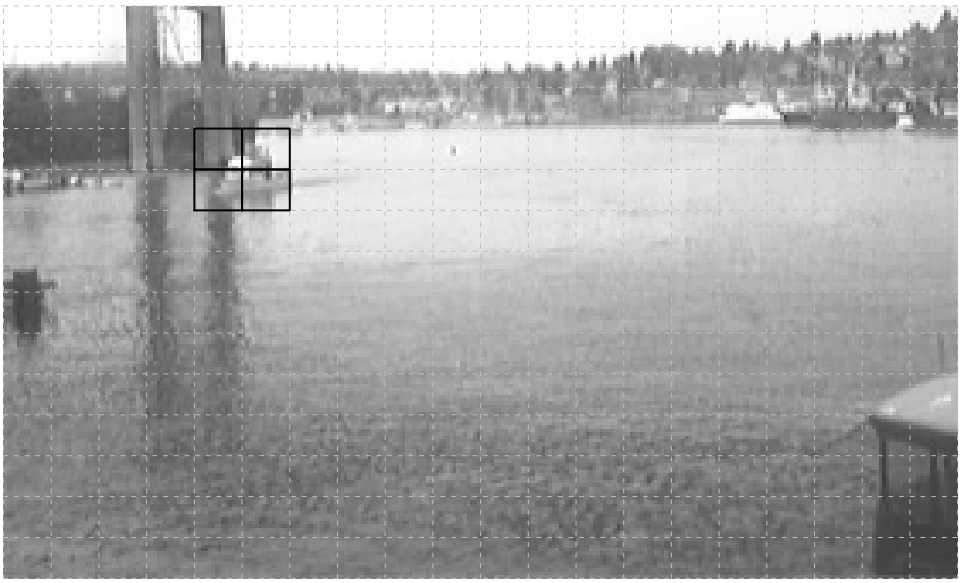

Figure 1: Portion of a picture taken by a web camera overlooking Portage Bay, Seattle. The picture has been divided into a 14×20 grid of rectangular bins, four of which are highlighted and contain a boat passing through the bay.

data vectors.

Motivated by potential applications of change detection to surveillance, we decided to regard the appearance of boats in the image stream as changes of interest. We looked at each of the 5002 images and manually marked the bins in each image containing a boat passing through Portage Bay. Figure 1 shows one such image, with four bins marked as containing a boat. Figure 2 shows the number of marked bins for each image plotted against image index. We define a boat event as a sequence consisting of two or more consecutive images with at least one marked bin. There are 19 boat events in all, with the shortest consisting of 6 images, and the longest, of 151 images. The black rectangles at the bottom of Figure 2 show where these events occur. There are 20 quiescent periods surrounding the boat events. The shortest (longest) consists of 2 (1030) images. The images during the quiescent periods are quite variable, due to light variations on the water from cloud movement, ducks moving around in the water close to the camera, wind-driven ripples in the water, wakes from boats no longer in view of the camera, and other sources of noise.

We emphasize that we use the images primarily as a means for constructing a multivariate data stream with characteristics one would expect in actual applications of change detection, but are not typically present in simulated data (e.g., correlated and heterogeneous noise). We do not make use of the fact that there is a neighborhood structure among the 280 variables; in fact, all of the results we present would be exactly the same if we were to randomly reorder the variables. In short, the methods we propose are not specific to image streams.

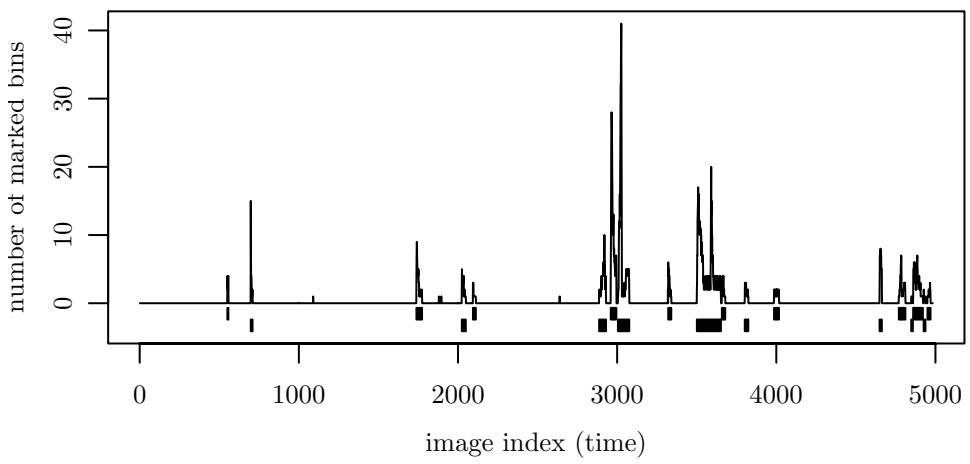

Figure 2: Number of bins (variables) marked as containing a boat versus image index (top part of plot), along with markers for the nineteen boat events (bottom).

3 Quantifying the Performance of a Change Detector

Defining a general measure quantifying the performance of a change detector for streams is a nontrivial problem. Generally there are two kinds of errors, missed changes and false alarms, but appropriate definitions for these errors can depend on the application. Even in the simplest scenario – a stream $\mathbf{x}_t$ consisting of stretches during which the vectors are independent and identically distributed (IID) – there appears to be no obvious way to distinguish between a late alarm, and a missed change followed by a false alarm. Another problem is that the detection stream d_t is typically correlated, even if observations in the input stream $\mathbf{x}_t$ are independent. This autocorrelation causes false alarms to occur in bursts, and we have to choose between counting individual false alarms or counting bursts. Moreover, the piecewise IID scenario might not be appropriate: in our boat detection problem, as presumably in other surveillance applications, it makes more sense to think of the stream as a concatenation of "quiescent periods" (without boats), interrupted by "events" (where boats appear and move across the scene). During the events, the distribution of the observation vectors is not constant because the boat (or boats) move.

In a surveillance context raising an alarm soon after a change caused by the transition from a quiescent period to an event is obviously crucial; if the delay is too long, the horse will have left the barn, and the alarm is no longer useful. Changes within events or transitions from events to quiescent periods are not of interest. We consider an event to be successfully detected if the detection stream exceeds the alarm threshold τ at least once within a tolerance window of width N_W after the onset of the event. We define the hit rate $h(\tau)$ as the proportion of events that are successfully detected. We define the false alarm rate $f(\tau)$ purely in terms of the quiescent periods: it is simply the proportion of times in those periods during which the detection stream exceeds the alarm threshold. There is no penalty for raising multiple alarms during an event. Our definitions for hit rate and false alarm rate are admittedly simple, and alternative definitions might be better adapted to scenarios not

involving surveillance. The method for comparing the performance of change detectors proposed in Section 5 is not critically dependent on the particular definitions.

We can summarize the performance of a change detection algorithm by plotting the hit rate $h(\tau)$ versus the false alarm rate $f(\tau)$ as we increase the alarm threshold τ. Both $h(\tau)$ and $f(\tau)$ are monotonically non-increasing functions of τ. The graph of the curve $\tau \longrightarrow (f(\tau), h(\tau))$ is a monotonically non-decreasing function of $f(\tau)$. We call this curve the ROC curve of the detection algorithm, because of the obvious parallels to the standard ROC curve used to summarize the performance of binary classifiers [6].

It is sometimes useful to compare the performance of a detection algorithm with a reference algorithm that completely ignores the data and simply signals an alarm with probability α whenever a new input observation arrives. The ROC curve of this "null" detector is $\alpha \longrightarrow (\alpha, 1-(1-\alpha)^{N_W})$ as α varies from 0 to 1; it depends on the width N_W of the tolerance window.

4 Setting the Alarm Threshold

A critical parameter of a change detector is the alarm threshold τ, whose variation controls the tradeoff between false alarms and missed changes. Without training data for which changes of interest have been marked, there is no way of realistically assessing the hit rate $h(\tau)$ for a given alarm threshold τ. The commonly proposed approach to setting τ is therefore to choose a false alarm rate α considered acceptable and then determine the corresponding τ. If we are willing to accept the "piecewise IID" model, then the appropriate value of τ can sometimes be determined analytically. If an explicit calculation is infeasible, we can resort to a computational approach based on a permutation argument [1, 5]. Assuming there is no change at or before the current time T, then $\mathbf{x}_1, \ldots, \mathbf{x}_T$ would be IID (under the piecewise IID model). To test the IID hypothesis, we compare the current value d_T^{orig} of the detection stream to values $d_T^1, \ldots, d_T^M$ obtained by applying the detection algorithm to M random permutations of $\mathbf{x}_1, \ldots, \mathbf{x}_T$. If d_T^{orig} is the k-th largest among $\{d_T^{orig}, d_T^1, \ldots, d_T^M\}$, then we can reject the IID hypothesis at level $k/(M+1)$. If the level is less than the desired false alarm rate, we signal a change and "reset the clock" by discarding $\mathbf{x}_1, \ldots, \mathbf{x}_T$. (Note that in this case the detection threshold will vary with time.)

The problem with both the analytical and the permutation-based approaches is that their validity depends critically on the piecewise IID assumption. This assumption seems inherently implausible, given that we are observing a time series. If it is violated, the results can be wildly off the mark. We now illustrate the problem for a simple detection algorithm based on a two-sample test. Detection algorithms based on two-sample tests have been discussed previously in the literature (see, e.g., [8] and references therein). The idea is to use a two-sample test for comparing the distribution P of the most recently observed data with the distribution Q of a reference set observed earlier. The value d_T of the detection stream is the test statistic of the two-sample test, and the (nominal) false alarm rate for detection threshold τ is the probability that $d_T \geq \tau$ under the null hypothesis $P = Q$. The qualifier "nominal" is a reminder that the significance level is derived under the IID assumption.

For our illustration we assume that the data stream is one dimensional. We define the current set as the N_C most recent observations, and the reference set as the N_R observations

immediately preceding the current set. We use the square of a two-sample t-test to form the detection stream at the current time T:

$$d_T = \frac{(\bar{x}_C - \bar{x}_R)^2}{\left(\frac{1}{N_C} + \frac{1}{N_R}\right)\hat{\sigma}^2},$$

where $\bar{x}_C$ and $\bar{x}_R$ are the sample means of the current and reference sets, and

$$\hat{\sigma}^2 = \frac{1}{N_C + N_R - 2}\left[\sum_{n=0}^{N_C-1}(x_{T-n} - \bar{x}_C)^2 + \sum_{n=0}^{N_R-1}(x_{T-N_C-n} - \bar{x}_R)^2\right]$$

is the pooled variance estimate. Although the t-test is designed to test the null hypothesis that the observations in the current and reference sets have the same mean, we can still use it as a test of the IID hypothesis, recognizing that it might have little or no power for detecting changes other than mean shifts.

If we are willing to assume that the observations in the current and reference sets are realizations of IID Gaussian random variables, then the threshold τ for false alarm rate α is the square of the $\alpha/2$ quantile of the t distribution with $N_C + N_R - 2$ degrees of freedom. If we do not want to make the Gaussianity assumption, we can use the permutation approach described above. The problem in either case is that the actual false alarm rate can be vastly different from the desired (nominal) rate if the independence assumption is violated.

As an example, choose $N_C = 4$, $N_R = 16$, and let $X_1, \ldots, X_{20}$ be a segment of a Gaussian first-order univariate autoregressive (AR) process $X_t = \phi X_{t-1} + \epsilon_t$, where $|\phi| < 1$ can be interpreted as the correlation between X_{t-1} and X_t, and the random variables ϵ_t are IID Gaussian with zero mean and unit variance. If $\phi = 0$, then $X_1, \ldots, X_{20}$ are IID. If $\phi \neq 0$, then the X_t variables are still Gaussian and identically distributed, but no longer independent. The alarm threshold for false alarm rate $\alpha = 0.1$ under the assumption of independence ($\phi = 0$) is $\tau \doteq 3$ (the square of the 5th percentile for a t distribution with 18 degrees of freedom). For five selected values of ϕ, we simulate $10,000$ independent realizations of $X_1, \ldots, X_{20}$ and compute d_{20} for each realization. We then estimate the actual false alarm rate α as the fraction of times when $d_{20} > 3$ in the $10,000$ realizations. The results are shown in Table 1. Note that, as expected, the false alarm rate is close to 0.1 when $\phi = 0$, but is dramatically off the mark otherwise.

To illustrate the failure of the permutation approach, we generate an additional 1000 independent realizations of $X_1, \ldots, X_{20}$ for our selected values of ϕ. For each of these realizations, we generate 1000 random permutations, compute d_{20} and keep track of the proportion of times that $d_{20} > 3$ — this proportion is what a permutation test would declare the false alarm rate to be for $\tau = 3$. When averaged over all 1000 realizations of the AR process, this proportion is very close to 0.1 for all five values of ϕ: the permutation approach gives the correct false alarm rate when $\phi = 0$ (the IID case) but it underestimates (overestimates) the correct rate α when $\phi > 0$ ($\phi < 0$), with the discrepancy becoming more serious as ϕ approaches 1 (-1). We conclude that the permutation-based approach for setting the alarm threshold is not viable in the presence of correlated data (presumably this is almost always the case when dealing with time series).

ϕ	-0.9	-0.5	0	0.5	0.9
α	0.008	0.018	0.098	0.282	0.537

Table 1: False alarm rate α for the squared two-sample t-test using a threshold level of $\tau = 3$ and data generated from a Gaussian first-order autoregressive process with a unit-lag autocorrelation of ϕ.

5 Comparing Change Point Detectors

In this section we propose a method for evaluating the relative performance of change detectors that takes into account sampling variability.

Suppose we have two change detectors with ROC curves $\tau \longrightarrow (f_1(\tau), h_1(\tau))$ and $\tau \longrightarrow (f_2(\tau), h_2(\tau))$. There are two obvious ways to use these curves for assessing the relative performance of the detectors. For a given hit rate $h_1(\tau_1) = h_2(\tau_2) \equiv h$, we can compare the false alarm rates $f_1(\tau_1)$ and $f_2(\tau_2)$ and declare the first detector to be better if $f_1(\tau_1) < f_2(\tau_2)$; alternatively, for a given false alarm rate, we can compare hit rates. More elaborate comparison schemes are possible [13]. In our boat example, we define the hit rate $h(\tau)$ in terms of the onset of a small number of events, so it is easier to compare the false alarm rates for a given hit rate. This approach yields false alarm rates for the two detectors that are functions of h. We denote these functions as $f_1(h)$ and $f_2(h)$ and compare them by either their difference $\Delta_{1,2}(h) = f_1(h) - f_2(h)$ or the ratio

$$r_{1,2}(h) = \frac{\max(f_1(h), \epsilon)}{\max(f_2(h), \epsilon)}, \tag{1}$$

where ϵ (a small number) allows for the false alarm rate to be zero (note that $\Delta_{2,1}(h) = -\Delta_{1,2}(h)$ and $r_{1,2}(h) = 1/r_{2,1}(h)$).

We use a modified version of the block bootstrap to assess if $\Delta_{1,2}(h)$ is significantly different from zero or if $r_{1,2}(h)$ is significantly different from one. Block bootstrapping is an adaptation of the standard bootstrap that is appropriate for time series [4, 9, 17]. In the standard bootstrap, the basic unit for resampling is an individual observation; in a block bootstrap, the basic unit is a block of consecutive observations, with each block having the same size. The block size is selected such that, within a block, the dependence structure of the original time series is preserved, while values at the beginning and end of each block are approximately independent of each other. Our input stream is naturally broken up into blocks of unequal size, namely, boat events and quiescent periods. We use these blocks to define the basic unit in two modified block bootstraps. The first modification is an "uncoupled" scheme. Given n_e boat events and $n_q = n_e + 1$ quiescent periods, we resample (with replacement) n_e boat events and n_q quiescent periods to form a bootstrap sample with the same overall structure as the original input stream (i.e., n_q quiescent periods separated by n_e events). The second modification is a "coupled" scheme, in which the basic unit is taken to be an event and its preceding quiescent period. The motivation for the second scheme is to preserve any dependence between the quiescent period preceding an event and the event itself.

The method for comparing detectors is the same for the coupled and the uncoupled schemes. For a given bootstrap sample, we evaluate $f_1(\tau)$, $h_1(\tau)$, $f_2(\tau)$ and $h_2(\tau)$ over a grid of thresholds τ, from which we calculate the curves $\Delta_{1,2}(h)$ and $r_{1,2}(h)$. We repeat this procedure n_b times, yielding n_b bootstrap replicates of $\Delta_{1,2}(h)$ and $r_{1,2}(h)$. We then construct $(1-\alpha)$ two-sided non-simultaneous confidence intervals for the difference $\Delta_{1,2}(h)$ and for the ratio $r_{1,2}(h)$ based upon the empirical distribution of the bootstrap replicates. (The "matched pair" design by which we evaluate the difference between detectors for each bootstrap sample and then compute confidence intervals for the difference will lead to sharper comparisons than an unmatched design in which bootstrap samples are generated separately for each detector.) As we vary h, the end points of these confidence intervals trace out confidence bands. If the confidence interval for $\Delta_{1,2}(h)$ at a given hit rate h does not include zero, we have evidence at the $(1-\alpha)$ confidence level that one change detector outperforms the other in that it has a smaller false alarm rate for hit rate h. The analogous inference can be made if the confidence interval for $r_{1,2}(h)$ does not include one.

6 An Illustrative Example

In this section we illustrate the methodology presented in the previous sections by considering two different change detectors that are designed to detect the boat events described in Section 2. We define the two-sample tests behind the change detectors in Section 6.1, after which we demonstrate the pitfalls of using a permutation approach to determine the false alarm rate (Section 6.2). We then illustrate how we can compare the performance of the two change detectors in a manner that takes into account sampling variability (Section 6.3).

6.1 Definition of Detection Streams based on Two-Sample Test Statistics

The two detectors we use to illustrate our methodology are quite different in their intent, but both are based on two-sample tests. The first detector is designed to be sensitive to mean changes, while the second uses a nonparametric test with power against all alternatives. To simplify notation, we define the tests for samples $\mathbf{c}_1, \ldots, \mathbf{c}_n$ (the current set) and $\mathbf{r}_1, \ldots, \mathbf{r}_m$ (the reference set), with the understanding that we would obtain the values of the corresponding detection streams at the current time T by comparing the $n = N_C$ most recent observations with the $m = N_R$ observations immediately preceding them.

The first detection stream, denoted as $d_T^{(\max)}$, is based on the largest squared element of the vector $\bar{\mathbf{c}} - \bar{\mathbf{r}}$, where $\bar{\mathbf{c}}$ is the average of $\mathbf{c}_1, \ldots, \mathbf{c}_n$, and $\bar{\mathbf{r}}$ is similarly defined. The detection stream will be large if there has been a recent large change in one or more of the 280 variable in the input stream, i.e., a large change in mean grey level for one or more of the bins in the image. Boats are small and their appearance changes the mean grey level for a small number of bins; therefore we want a test that is sensitive to large changes in a few bins, rather than to small changes in a large number of bins. The top pane of Figure 3 shows the detection stream $d_T^{(\max)}$ plotted against time for the case $N_C = 4$ and $N_R = 16$.

The second change detector we consider is based on a so-called "energy" test statistic that has been advocated as a nonparametric test for equality of two multivariate distributions [2,

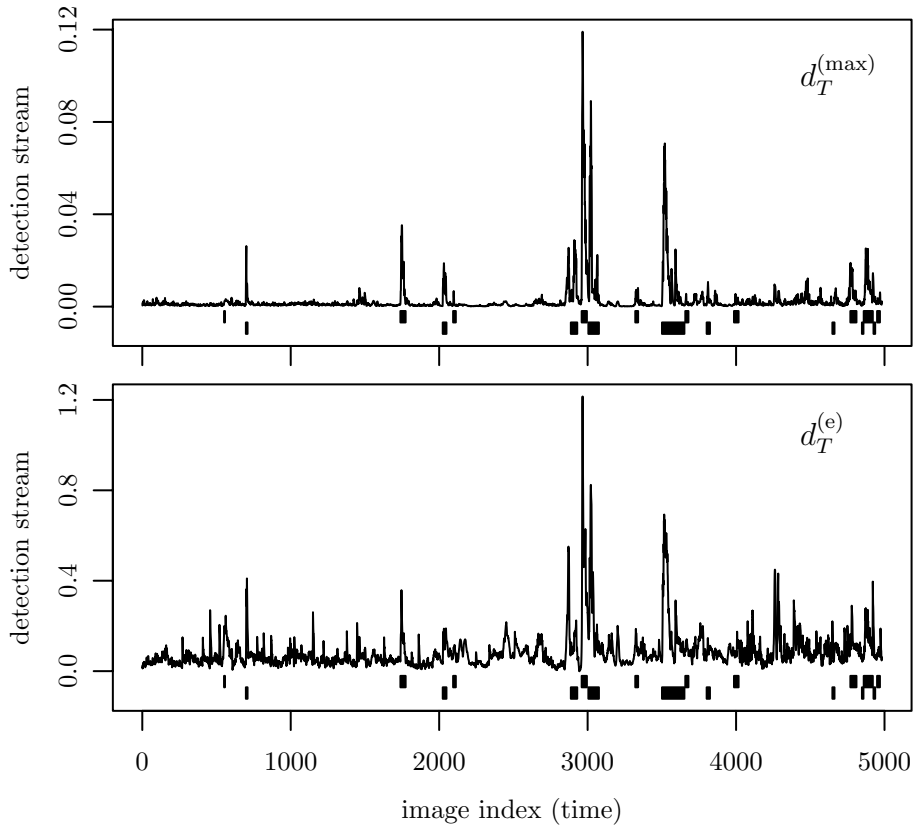

Figure 3: Two detection streams plotted versus image index T, with boat events marked as in Fig. 2. The top plot shows $d_T^{(\max)}$, which is based upon the maximum squared difference in means; the bottom is for $d_T^{(e)}$, which is based upon the energy test statistic. The settings $N_C = 4$ and $N_R = 16$ are used for both detectors at each current time T.

18, 20, 21, 22]. This statistic is given by

$$d_T^{(e)} = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{c}_i - \mathbf{r}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{c}_i - \mathbf{c}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{r}_i - \mathbf{r}_j\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. This test is consistent against all alternatives to H_0 and hence is not focused on any particular aspect of the difference in distribution between the current and reference sets [22]. Because it is an omnibus test, it cannot be expected have as much power for detecting a change in means as a test specifically designed for that type of change. The bottom pane of Figure 3 shows the detection stream $d_T^{(e)}$ plotted against time.

6.2 Pitfalls of Setting the Alarm Threshold via Permutation Tests

To complement the simulated example of Section 4, we now present an empirical demonstration of our assertion that we cannot expect to get reasonable estimates of the false alarm rate using a permutation argument.

We apply the change detector based on the energy test statistic with $N_C = 4$ and $N_R = 16$ to the longest quiescent period in our boat data (1030 images). For each of the 1011 segments

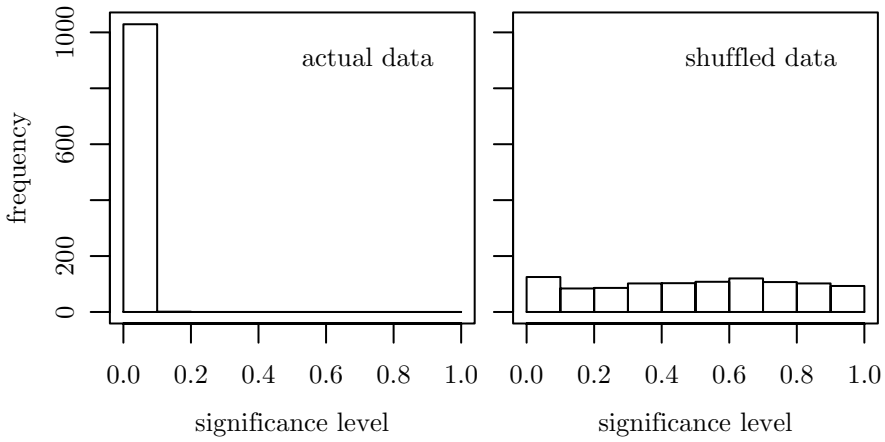

Figure 4: Histograms of levels of significance (p-values) as empirically determined by a permutation test based upon data from a quiescent period (left-hand plot) and upon data from the same period but randomly shuffled (right-hand).

of length 20 we calculate the permutation-based p-value of the energy test statistic: we compare the original value of the test statistic for the segment with a reference set of 500 "permuted" values obtained by applying the test to a randomly shuffled version of the segment. If the original value is the k-th largest amongst these 501 values, then the p-value (the level of significance of the test) is $\hat{\alpha} = k/501$ [1].

Since we are dealing with a quiescent period, the distribution of $\hat{\alpha}$ across all 1011 values in the detection stream should be uniform over the interval $[0, 1]$ (see Lemma 3.3.1 of [11]). The left-hand pane of Figure 4 shows a histogram of the p-values, which clearly is not consistent with a uniform distribution. To demonstrate that it is indeed the correlated nature of the input stream that is causing the problem, we reran the entire procedure using the same 1030 images, but shuffling the order of the images at random. This shuffling removes the correlation between images that are close to one another. We now obtain the histogram in the right-hand pane, which is clearly much more consistent with a uniform distribution. This demonstrates that we can use a permutation argument to determine the false alarm rate if indeed the IID assumption is valid.

6.3 Comparison of Change Point Detectors

Here we compare the two change detectors whose detection streams are based on the two-sample test statistics defined in Section 6.1 (again using $N_C = 4$ and $N_R = 16$). As discussed in Section 3, we declare that a change detector has successfully identified a boat event if the detection stream exceeds the alarm threshold at least once during a tolerance window of width N_W. For this example, we let N_W be the same as the current set $N_C = 4$, but other choices could be entertained (i.e., there is no compelling reason to couple N_W with N_C).

Figure 5 shows the ROC curves for the change detectors based on $d_T^{(\max)}$ and $d_T^{(e)}$ along

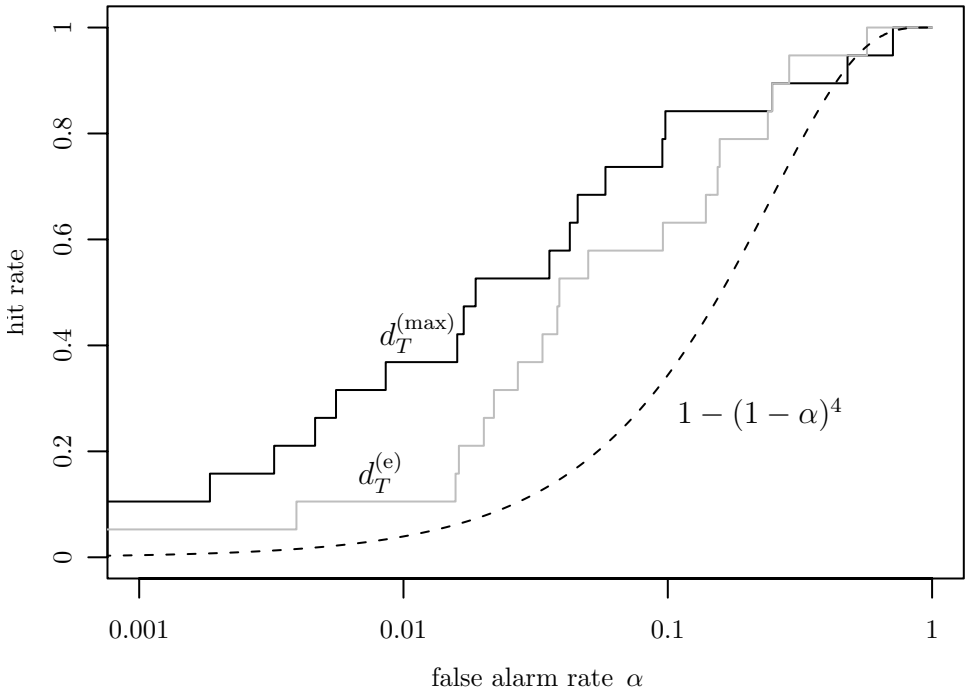

Figure 5: ROC curves for the detection streams $d_T^{(\max)}$ and $d_T^{(e)}$.

with a plot of $1-(1-\alpha)^{N_W}$ versus α. (As noted in Section 3, this is the ROC curve for a statistic that rejects H_0 based upon a "coin flip" with false alarm rate α). Except at the very highest hit and false alarm rates (upper right-hand corner), the $d_T^{(\max)}$ detector (sensitive to mean changes) generally outperforms the $d_T^{(e)}$ detector (sensitive to arbitrary changes) in the sense of having a smaller false alarm rate for a given hit rate. To assess whether this difference between the detectors is statistically significant, we use the bootstrap procedures discussed in Section 5 to determine a 90% (non-simultaneous) confidence band for the difference $\Delta_{\max,e}(h)$ and the ratio $r_{\max,e}(h)$ defined in Equation (1) with $\epsilon = 0.001$. The uncoupled and coupled bootstrap procedures yield basically the same results, so we only present results from the uncoupled scheme. Figure 6 shows the confidence bands based upon 100 uncoupled bootstrap samples. Except for a limited range of hit rates around 0.2 to 0.3, the intervals for $\Delta_{\max,e}(h)$ include 0, and the intervals for $r_{\max,e}(h)$ include 1, indicating that for most hit rates the difference between the two detectors is not significant. A possible explanation for this inconclusive result is the small number of events in our training data.

7 Summary and Discussion

We have proposed a method for comparing two change detectors. The method is based on labeled data, i.e., a segment of the input stream in which we have identified events and quiescent periods. The key element is an adaptation of the block bootstrap. The adaptation constructs bootstrap streams by piecing together events and quiescent periods randomly chosen (with replacement) from those making up the original stream. The bootstrap allows

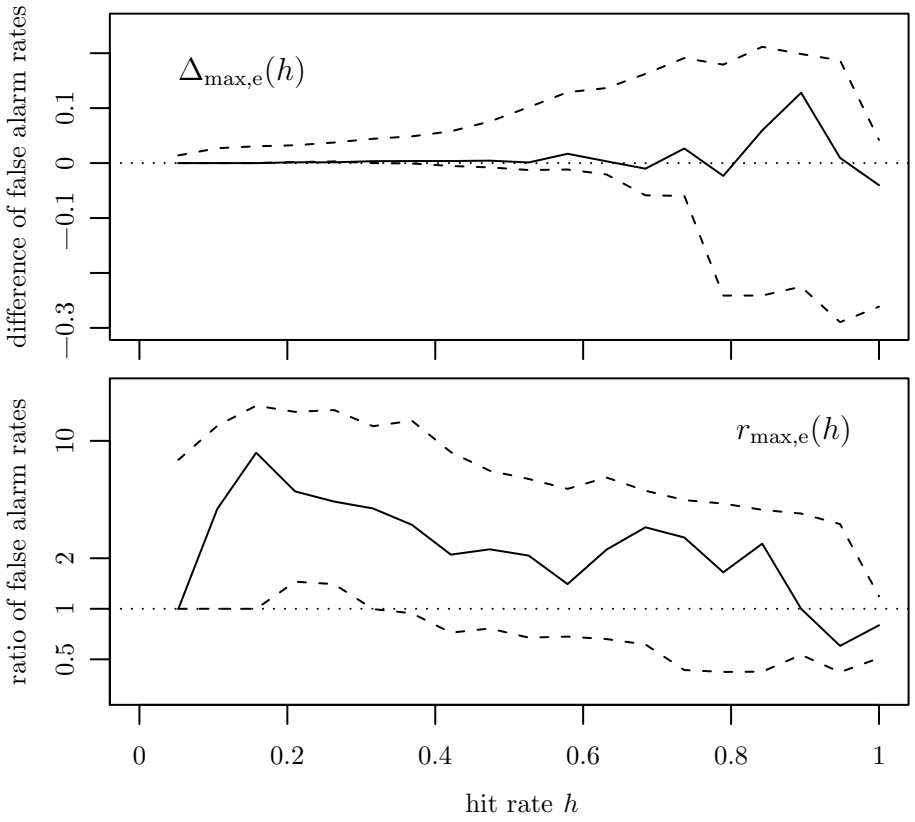

Figure 6: Comparing ROC curves using $\Delta_{\max,e}(h)$ (solid curve, top plot) and $r_{\max,e}(h)$ (bottom). The dashed curves in each plot indicate non-simultaneous 90% empirical confidence intervals based upon 100 bootstrap samples.

us to assess the effect of sampling variability on pairwise comparisons of ROC curves, and thereby determine whether a particular change detector is significantly better than another. Our example compared two change detectors whose detection streams are constructed using two-sample tests, but our method is not dependent upon this particular construction and can be applied to other kinds of change detectors (e.g., the output of cumulative sum statistics [19], which are not based on the two-sample notion).

In addition, we have demonstrated the pitfalls of using a permutation approach to assess the false alarm rate in a streaming environment. A permutation approach is appropriate for data that is reasonably modeled as IID, which is not likely to be the case for streams. Our experiments show that it is not possible to assess the false alarm rate to any reasonable degree of accuracy using a permutation approach unless the IID assumption holds.

Our proposed method can be extended to compare the performance of $K > 2$ change detectors. Change detectors can arise in many different ways. For example, we could use several different choices for the sizes of the current and references windows, as proposed by Kifer et al. [8]. We might also be interested in time-varying geometries. One obvious choice is to allow the reference window to grow monotonically in time, while the size of the current window is kept constant. This geometry has the potential benefit of a large sample size in the reference window and leads to a comparison of the data in the current window with

previous long-term averages.

A problem with comparing multiple detectors is that none of them might emerge as uniformly best for all hit rates. Even for the two detectors in our boat example, Figure 5 shows that, if we ignore the question of statistical significance, the $d_T^{(\max)}$ detector generally outperforms the $d_T^{(e)}$ detector, but not at the highest hit/false alarm rates. We could focus on a single hit rate and then order the K detectors by their false alarm rates. The natural generalization of the matched pairs design for comparing the false alarm rates of two detectors is a blocked design where each bootstrap sample is a block, and the detectors are the treatments. Detectors can then be compared using standard multiple comparison procedures.

If we have labeled data, we can do more than just evaluate the performance of predefined change detectors – we can use the data to design new detectors. One possibility is to look for linear or nonlinear combinations of existing change detectors that outperform any single detector. For example, suppose that a change of interest is associated with a change in both the mean and the variance of a distribution, and suppose that we have two change detectors, one of which has power against changes in means, and the other, against changes in variance. Then some combination of these two detectors is likely to be superior to either individual detector in picking up on the change of interest, and the particular combination that is best can be determined using the labeled data. Using labeled data to construct new "targeted" change detectors is an interesting area for future research.

We have assumed all along that labeled data have been collected before the change detector is put into operation and therefore can be used in the design of the detector. In the context of on-going surveillance, an human operator who is responding to alarms raised by a change detector could in principle provide feedback indicating whether an alarm is false or does indeed identify an event of interest. This would provide additional labeled data; however, these would be qualitatively different from the training sample in that they would provide information on false alarms but not on missed hits. Determining how best to make use of feedback for assessing and improving the performance of change detectors is another problem worthy of additional research.

Acknowledgments

The authors would like to thank the personnel behind the SORFED project (Kevin Williams, Russ Light and Timothy Wen) for supporting us in the use of their data. This work was funded by the U.S. Office of Naval Research under grant number N00014–05–1–0843.

References

[1] T.W. Anderson, "Sampling Permutations for Nonparametric Methods", in: B. Ranneby (editor), Statistics in Theory and Practice: Essays in Honour of Bertil Matérn, Swedish University of Agricultural Sciences, Umeå, Sweden, 1982, pp. 43–52.

[2] B. Aslan, G. Zech, "New Test for the Multivariate Two-Sample Problem Based on the Concept of Minimum Energy", Journal of Statistical Computation and Simulation, Vol. 75, 2005, pp. 109–119.

[3] P. Bélisle, L. Joseph, B. MacGibbon, D.B. Wolfson, R. du Berger, "Change-Point Analysis of Neuron Spike Train Data", Biometrics, Vol. 54, 1998, pp. 113–123.

[4] P. Bühlmann, "Bootstraps for Time Series", Statistical Science, Vol. 17, 2002, pp. 52–72.

[5] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[6] T. Fawcett, "An Introduction to ROC Analysis", Pattern Recognition Letters, Vol. 27, 2006, pp. 861–874.

[7] M. Frisén, "Statistical Surveillance. Optimality and Methods", International Statistical Review, Vol. 71, 2003, pp. 403–434.

[8] D. Kifer, S. Ben-David, J. Gehrke, "Detecting Change in Data Streams", Proceedings of the 30th Very Large Data Base (VLDB) Conference, Toronto, Canada, 2004, pp. 180–191.

[9] S.N. Lahiri, Resampling Methods for Dependent Data, Springer, New York, 2003.

[10] T.L. Lai, "Sequential Changepoint Detection in Quality Control and Dynamical Systems", Journal of the Royal Statistical Society, Series B (Methodological), Vol. 57, 1995, pp. 613–658

[11] E.L. Lehmann, J.P. Romano, Testing Statistical Hypotheses (Third Edition), Springer, New York, 2005.

[12] F. Li, G.C. Runger, E. Tuv, "Supervised Learning for Change-Point Detection", International Journal of Production Research, Vol. 44, 2006, pp. 2853–2868.

[13] S.A. Macskassy, F. Provost, "Confidence Bands for ROC Curves: Methods and an Empirical Study", First Workshop on ROC Analysis in AI, ECAI-2004, Spain, 2004.

[14] M. Markou, S. Singh, "Novelty Detection: A Review – Part 1: Statistical Approaches", Signal Processing, Vol. 83, 2003, pp. 2481–2497.

[15] M. Markou, S. Singh, "Novelty Detection: A Review – Part 2: Neural Network Based Approaches", Signal Processing, Vol. 83, 2003, pp. 2499-2521.

[16] A. Pievatolo, R. Rotondi, "Analysing the Interevent Time Distribution to Identify Seismicity Phases: A Bayesian Nonparametric Approach to the Multiple-Changepoint Problem", Applied Statistics, Vol. 49, 2000, pp. 543–562.

[17] D.N. Politis, J.P. Romano, M. Wolf, Subsampling, Springer, New York, 1999.

[18] R. Rubinfeld, R. Servedio,"Testing Monotone High-Dimensional Distributions", Proceedings of the 37th Annual Symposium on Theory of Computing (STOC), 2005, pp. 147–156.

[19] G.C. Runger, M.C. Testik, "Multivariate Extensions to Cumulative Sum Control Charts", Quality and Reliability Engineering International, Vol. 20, pp. 587–606.

[20] G.J. Székely, "Potential and Kinetic Energy in Statistics", Lecture Notes, Budapest Institute of Technology (Technical University), 1989.

[21] G.J. Székely, "E-statistics: Energy of Statistical Samples", Technical Report No. 03–05, Bowling Green State University, Department of Mathematics and Statistics 2000.

[22] G.J. Székely, M.L. Rizzo, "Testing for Equal Distributions in High Dimension", InterStat, 2004.