Concise Papers

Change Detection in Streaming Multivariate Data Using Likelihood Detectors

Ludmila I. Kuncheva, *Member*, *IEEE*

Abstract—Change detection in streaming data relies on a fast estimation of the probability that the data in two consecutive windows come from different distributions. Choosing the criterion is one of the multitude of questions that need to be addressed when designing a change detection procedure. This paper gives a log-likelihood justification for two well-known criteria for detecting change in streaming multidimensional data: Kullback-Leibler (K-L) distance and Hotelling's T-square test for equal means (H). We propose a semiparametric log-likelihood criterion (SPLL) for change detection. Compared to the existing log-likelihood change detectors, SPLL trades some theoretical rigor for computation simplicity. We examine SPLL together with K-L and H on detecting induced change on 30 real data sets. The criteria were compared using the area under the respective Receiver Operating Characteristic (ROC) curve (AUC). SPLL was found to be on the par with H and better than K-L for the nonnormalized data, and better than both on the normalized data.

Index Terms—Change detection, multidimensional data streams, Hotelling's T-square, log-likelihood detector

1 Introduction

Change detection in data streams has been extensively studied due to its vast application potential in all walks of science and technology, for example, fraud detection, market analysis, medical condition monitoring, and network traffic control [1]. The most notable application is engineering where *control charts* have been used for process quality control [2]. Classical examples of control charts are Shewhart's method, CUmulative SUM (CUSUM) and Wald's Sequential Probability Ratio Test (SPRT) [3], [4], [5]. There is a large collection of change detection methods for monitoring a single variable [6], [7], [8], [9], [10], [11], e.g., proportion of defective items, classification error rate, network traffic volume, or market indices. One of the main assets of the univariate change detection methods is their statistical soundness. Advanced as they are, these methods cannot handle directly multidimensional streaming data. Here, we are interested in detecting a change in the unlabeled multidimensional data stream.

Changes can be short-term "blips" or sustained shifts in the streaming data distribution. Anomaly/outlier detectors are used for blips or rare events [12], [13], providing a basis of applications such as intrusion and fraud detection. The second type of changes (steady changes) subdivides into abrupt and gradual. All change types can be detected by comparing two consecutive windows from the streaming data.

The list below helps to position this study within the grand landscape of change detection methods:

- We assume that the two consecutive windows of data, W_1 and W_2, are given. Many studies have been devoted to developing strategies of choosing, sampling, splitting, growing, and shrinking the windows for optimal change detection [6], [9], [11], [14], [15].
- This paper is about the *criterion* of detecting change in the distributions of the data in windows W_1 and W_2. While many criteria have been already discussed [6], [7], [8], [10], [16], [17], some in relationship with the windows definition and change signaling procedure, there is no general consensus about a single criterion. This paper looks at change detection from the perspective of likelihood as a general framework, and demonstrates that two popular change detection criteria are special cases thereof.
- While we give a simple recommendation about the threshold on the criterion value for declaring a change, we are not proposing a theoretical guarantee. Several solutions to this problem have already been proposed, for example, bootstrap Monte Carlo sampling, permutation sampling, or statistical significance levels [7], [10], [16]. To evaluate the criteria from the likelihood approach, we use the area under the Receiver Operating Characteristic (ROC) curve (AUC).

The contribution of this study is twofold. First, we show that the change detection through K-L distance and Hotelling's t^2 test can be accommodated within a common log-likelihood framework. Second, we propose a computationally simple criterion for change detection, called *semiparametric log-likelihood* (SPLL) detector.

Song et al. [10] propose a rigorous log-likelihood change detection criterion, called *density test*, which relies on kernel density approximation. SPLL is a simpler log-likelihood detector, whereby the density approximation is replaced by a single round of k-means clustering. The theoretical grounds for this simplification give rise to a slightly different test statistic. Compared to the density test, the simplicity of SPLL is paid by adopting a cruder semiparametric assumption about the density before the change, and using an upper bound on the log likelihood instead of the true value.

The rest of the paper is organized as follows. Section 2 positions the K-L distance and the Hotelling's t^2 within the log-likelihood detection framework for change detection. The SPLL detector is proposed and explained there, in relation to the density test [10]. In Section 3, we compare SPLL to the K-L distance and Hotelling's t^2 detectors using 30 real data sets with simulated change. Section 4 offers our conclusions.

2 Likelihood Change Detectors for Multivariate Data

2.1 K-L Distance for Qualitative Data

Let $\mathbf{x} = [x_1, \ldots, x_n]^T$ be the streaming multidimensional random variable, where each feature takes values from a (small) finite set of categories, $x_i \in \mathcal{D}_i$, $i = 1, \ldots, n$. Thus, each $\mathbf{x}$ is an element of $\mathcal{D}$, where $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \ldots, \mathcal{D}_n$. Consider a pmf $P_1(\mathbf{x})$ over $\mathcal{D}$, from which the data in window W_1 has been sampled. Suppose that the data in $W_2 = \{\mathbf{z}_1, \ldots, \mathbf{z}_{M_2}\}$ is sampled from distribution $P_2(\mathbf{x})$. The likelihood of W_2 with respect to $P_1(\mathbf{x})$ is

$$L(W_2|P_1) = \prod_{j=1}^{M_2} P_1(\mathbf{z}_j) = \prod_{\mathbf{x}\in\mathcal{D}} P_1(\mathbf{x})^{K_\mathbf{x}},$$

where $K_\mathbf{x}$ is the number of elements of W_2 equal to $\mathbf{x}$, i.e.,

$$K_\mathbf{x} = |\{\mathbf{z} \mid \mathbf{z} \in W_2, \mathbf{z} = \mathbf{x}\}|,$$

- *The author is with the School of Computer Science, Bangor University, Dean Street, Bangor, Gwynedd LL57 1UT, United Kingdom. E-mail: l.i.kuncheva@bangor.ac.uk.*

Manuscript received 2 July 2010; revised 12 July 2011; accepted 11 Oct. 2011; published online 19 Oct. 2011.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-07-0368.
Digital Object Identifier no. 10.1109/TKDE.2011.226.

$|\cdot|$ denoting cardinality. Similarly, the likelihood of W_2 having come from distribution P_2 is

$$L(W_2|P_2) = \prod_{\mathbf{x}\in\mathcal{D}} P_2(\mathbf{x})^{K_\mathbf{x}}.$$

If W_2 came from P_2 it would have higher likelihood with respect to P_2 than with respect to P_1. Taking the logarithm of the likelihood ratio (assuming $P_1(\mathbf{x}) > 0$, $P_2(\mathbf{x}) > 0$ for any $\mathbf{x} \in \mathcal{D}$), we have

$$LLR = \log \prod_{\mathbf{x}\in\mathcal{D}} \left\{ \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} \right\}^{K_\mathbf{x}} = \sum_{\mathbf{x}\in\mathcal{D}} K_\mathbf{x} \log \left\{ \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} \right\}.$$

Dividing by the cardinality of W_2, M_2, and taking the limit at $M_2 \to \infty$, we arrive at the *asymptotic scaled log-likelihood ratio*

$$\begin{aligned}\lim_{M_2\to\infty} \left\{ \frac{LLR}{M_2} \right\} &= \sum_{\mathbf{x}\in\mathcal{D}} \lim_{M_2\to\infty} \left\{ \frac{K_\mathbf{x}}{M_2} \right\} \log \left\{ \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} \right\} \\ &= \sum_{\mathbf{x}\in\mathcal{D}} P_2(\mathbf{x}) \log \left\{ \frac{P_2(\mathbf{x})}{P_1(\mathbf{x})} \right\} \qquad (1) \\ &= KL(P_2 \| P_1),\end{aligned}$$

where $KL(P_2\|P_1)$ is the Kullback-Leibler distance between distributions P_1 and P_2, which is proposed by Dasu et al. [16] as a measure for change detection. If the two distributions are identical, the value of $KL(P_2\|P_1)$ is 0. The larger the value, the higher the likelihood that P_2 is different from P_1. In the real-life case, we do not have P_1 and P_2 but only approximations of these estimated from windows W_1 and W_2, respectively. The estimates can be plugged in (1) to obtain an estimate of the criterion, $KL(\hat{P}_2\|\hat{P}_1)$. The usefulness of the K-L criterion depends on the quality of the approximations and on finding a threshold λ such that change is declared if $KL > \lambda$. The approximations $\hat{P}_1$ and $\hat{P}_2$ will be better if the window sizes are large in relation to the cardinality of $\mathcal{D}$. Usually, W_1 is expanded until change is detected, giving a good basis for approximating P_1. On the other hand, P_2 has to be estimated from a short "recent" window, and the values may be noisy.

The K-L detector can be applied to continuous-valued streaming data too. In this case, a form of discretization is applied to split the space and keep the nonparametric character of the criterion. Dasu et al. suggest building *kdq* trees which can be updated with the streaming data. Each leaf is an instance $\mathbf{x} \in \mathcal{D}$. The authors point out, however, that the space can be partitioned in a different way, and call their approach "plug-and-play." The partitioning method will likely impact the robustness and sensitivity of the change detection method but will not invalidate the criterion. *kdq* trees lead to a nonparametric approach as the type of the pmf is not being guessed in advance.

The second problem with the K-L distance criterion is that it is not related to a straightforward statistical test that will give us a fixed threshold λ. Bootstrap Monte Carlo sampling and permutation tests have been suggested for estimating a suitable threshold [7], [10], [16].

2.2 Hotelling's t^2 Test for Quantitative Data

Suppose that the streaming observations are quantitative, $\mathbf{x} = [x_1, \dots, x_n]^T \in \Re^n$, and the distributions from which W_1 and W_2 are sampled have respective probability density functions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$. A straightforward test for equivalence of the means of the two distributions, given the window sizes, is the Hotelling's t^2 test [18]. The null hypothesis is that W_1 and W_2 are drawn independently from two multivariate normal distributions with the same mean and covariance matrices. Denote the sample means by $\hat{\mu}_1$ and $\hat{\mu}_2$, the pooled sample covariance matrix by $\hat{\Sigma}$, and the cardinalities of the two windows by $M_1 = |W_1|$ and $M_2 = |W_2|$. The T^2 statistic is calculated as

$$\begin{aligned}T^2 = {} & \frac{M_1 M_2 (M_1 + M_2 - n - 1)}{n(M_1 + M_2 - 2)(M_1 + M_2)} \\ & \times (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2).\end{aligned}$$

Under the null hypothesis, T^2 has F distribution with degrees of freedom n and $M_1 + M_2 - n + 1$. The T^2 statistic is the squared Mahalanobis distance between the two sample means multiplied by a constant. The statistic itself can be used as a basis for a change detection criterion. However, the p-value of the statistical test is better for this purpose because the threshold λ is instantly available as the desired significance level.

The following proposition demonstrates the relationship between the Mahalanobis distance (hence, the Hotelling's t^2 test) and the *asymptotic scaled log-likelihood ratio*.

Proposition 1. *Let W_1 and W_2 be drawn from two n-dimensional normal distributions, p_1 and p_2 with the same covariance matrix: $p_1 \sim N(\mu_1, \Sigma)$ and $p_2 \sim N(\mu_2, \Sigma)$, where $\mu_1, \mu_2 \in \Re^n$. Then, the* asymptotic scaled log-likelihood ratio *equals half of the squared Mahalanobis distance between the two distribution means.*

Form the log-likelihood ratio for window W_2

$$\begin{aligned}LLR &= \log \left\{ \prod_{\mathbf{x}\in W_2} \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} \right\} \\ &= \log \prod_{\mathbf{x}\in W_2} \left\{ \frac{\exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu_2)^T \Sigma^{-1} (\mathbf{x}-\mu_2)^T \right\}}{\exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu_1)^T \Sigma^{-1} (\mathbf{x}-\mu_1)^T \right\}} \right\} \\ &= -\frac{1}{2} \sum_{\mathbf{x}\in W_2} \left\{ (\mathbf{x}-\mu_2)^T \Sigma^{-1} (\mathbf{x}-\mu_2)^T - (\mathbf{x}-\mu_1)^T \Sigma^{-1} (\mathbf{x}-\mu_1)^T \right\} \\ &= -\frac{1}{2} \sum_{\mathbf{x}\in W_2} \left\{ -\mu_2^T \Sigma^{-1} \mu_2 + \mu_1^T \Sigma^{-1} \mu_1 - 2\mathbf{x}^T \Sigma^{-1} \mu_2 + 2\mathbf{x}^T \Sigma^{-1} \mu_1 \right\}.\end{aligned}$$

In addition to the constant terms, the LLR expression includes linear terms on $\mathbf{x}$. Noting that

$$\lim_{M_2\to\infty} \left\{ \frac{1}{M_2} \sum_{\mathbf{x}\in W_2} \mathbf{x} \right\} = \mu_2,$$

we obtain

$$\begin{aligned}\lim_{M_2\to\infty} \left\{ \frac{LLR}{M_2} \right\} &= -\frac{1}{2} \left(-\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1 + 2\mu_1^T \Sigma^{-1} \mu_2 \right) \\ &= \frac{1}{2} (\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1).\end{aligned}$$

2.3 Semiparametric Log-Likelihood Change Detector

The above change detection criteria are standard statistical measures of discrepancy between two distributions, which have been found useful in change detection from streaming data due to their computational ease and robustness. Here, we add a semiparametric criterion and demonstrate its merit through a subsequent experimental study.

Consider a Gaussian mixture with c components as the distribution $p_1(\mathbf{x})$

$$\begin{aligned}p_1(\mathbf{x}) &= \sum_{i=1}^{c} P(i)\, p_1(\mathbf{x}|i) \\ &= \sum_{i=1}^{c} \frac{P(i)}{(2\pi)^{n/2} \det(\Sigma_i)^{\frac{1}{2}}} \\ &\quad \exp\left\{ -\frac{1}{2} (\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i) \right\},\end{aligned}$$

where $P(i)$ are the mixing coefficients ($\sum P(i) = 1$), μ_i are the means of the components and Σ_i are the respective covariance matrices. The likelihood of W_2 will not factorize because of the sum. Therefore, we propose to replace the likelihood with an upper bound thereof and take a logarithm of that

$$L(W_2|p_1) = \prod_{\mathbf{x}\in W_2} p_1(\mathbf{x}) \leq \prod_{\mathbf{x}\in W_2} \max_{i=1}^{c}\left\{\frac{1}{(2\pi)^{n/2}\det(\Sigma_i)^{\frac{1}{2}}} \times \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\}\right\}. \tag{2}$$

The log-likelihood bound (2) is

$$\overline{LL} = -\frac{1}{2}\sum_{\mathbf{x}\in W_2}(\mathbf{x}-\mu_{i*})^T\Sigma_{i*}^{-1}(\mathbf{x}-\mu_{i*}) \times \log\left\{\frac{1}{(2\pi)^{n/2}\det(\Sigma_{i*})^{\frac{1}{2}}}\right\},$$

where $i*$ is the chosen component for the respective $\mathbf{x}\in W_2$, i.e., the component whose centroid μ_{i*} has the smallest Mahalanobis distance to $\mathbf{x}$. The distribution p_1 is estimated from data. If the size of W_1 is not very large, the estimates of the individual covariance matrices may be spurious. Hence, we propose to use the same covariance matrix for all density components, $\Sigma_i = \Sigma$, $i = 1,\ldots,c$. Dropping off the constants and scaling, we obtain

$$\frac{\overline{LL}}{M_2} \propto -\frac{1}{M_2}\sum_{\mathbf{x}\in W_2}(\mathbf{x}-\mu_{i*})^T\Sigma^{-1}(\mathbf{x}-\mu_{i*}).$$

Put in words, the log-likelihood bound is proportional to the sum of the negative squared Mahalanobis distances between each observation and its closest mean.

Consider again the log-likelihood ratio of the data in window W_2

$$LLR = \log\frac{L(W_2|p_2)}{L(W_2|p_1)}.$$

Assume that the distribution generating W_2 is exactly p_2, and $L(W_2|p_2) = 1$. In other words, we are interested to find out to what extent W_2 fits within p_1. Thus, the LLR criterion becomes

$$LLR = -\log L(W_2|p_1). \tag{3}$$

Using the scaled upper bound LL, the proposed expression for SPLL is

$$SPLL = -\frac{\overline{LL}}{M_2} \propto \frac{1}{M_2}\sum_{\mathbf{x}\in W_2}(\mathbf{x}-\mu_{i*})^T\Sigma^{-1}(\mathbf{x}-\mu_{i*}).$$

If W_2 comes from p_1, the squared Mahalanobis distances have a chi-square distribution with n degrees of freedom (where n is the dimensionality of the feature space) [19]. The expected value is n and the standard deviation is $\sqrt{2n}$. If W_2 does not come from the same distribution, then the mean of the distances will deviate from n.

Next, we draw a parallel with the density test by Song et al. [10]. The authors use the log-likelihood ratio with the following statistic:

$$\delta = \log\left\{\frac{L(W_2|p_1)}{L(W_1|p_1)^{\frac{M_2}{M_1}}}\right\} \tag{4}$$

$$= \log\{L(W_2|p_1)\} - \frac{M_2}{M_1}\log\{L(W_1|p_1)\}. \tag{5}$$

Compare (3) and (5). There is a second term in (5) which accounts for how well the data in window W_1 matches the estimated distribution. The authors argue that δ follows a normal distribution with mean zero, and propose a procedure for estimating the variance. Fort this to hold, the approximation of the density p_1 must be kernel based, with Gaussian kernels centred at each point. The number of these kernels will have to be large enough so that the Central Limit Theorem can be applied to ensure normality of δ. SPLL, on the other hand, is based on a semiparametric approximation of the distribution and number of components in the Gaussian mixture could be as small as 2. The fundamental statistical argument of whether parametric, semiparametric or nonparametric approximation of the data distribution is best, is out of the scope of this paper.

The distribution p_1 can be estimated using the Expectation Maximization (EM) family of algorithms. This is the route taken in many studies, including [10], where EM is suggested for determining the bandwidths of the kernels. In streaming data, however, EM may be too expensive computationally; therefore we propose to obtain p_1 through clustering and subsequent single evaluation of the component means and the common covariance matrix.

Why would SPLL work? The Hotelling's t^2 test detects changes in the *means*, and *assumes* equal covariance matrices of W_1 and W_2. Therefore, if the change of the distribution comes from change in the variances or covariances between the features, the test will be powerless. The K-L distance criterion is nonparametric, and, in principle, is able to detect any type of discrepancy, including changes in the variance, as demonstrated in Dasu et al.'s study [16]. However, it lacks some fidelity when the distributions are not naturally discrete. By discretizing the feature space, we lose information, especially when the number of features is large. Important data structures may get smoothed over in the process. The semiparametric criterion is intended as a "middle ground" that combats the deficiencies of both criteria.

3 Experiments with Simulated Change in Real Data

3.1 Data, Protocol and Results

Experiments were carried out with 30 sets from UCI [20] and from a private repository, listed in order of increasing number of features in Table 1. The number of objects and the number of features for each data set are also shown.

This experimental study shows how the proposed semiparametric log-likelihood criterion for change detection in multivariate data compares to the other two criteria: K-L distance and the Hotelling's t^2 test. Since we are not proposing or estimating an optimal threshold on the criterion value, the relative merit of the criteria will be evaluated by their Receiver Operating Characteristic curves [21]. An ROC curve plots the sensitivity of a method (proportion of true changes correctly detected) versus $(1 - \text{specificity})$, where specificity is the true negative rate of the method (proportion of correctly labeled windows with no change). The ideal ROC curve consists of points (0,0), (0,1), and (1,1), where point (0,1) is the desired outcome (100 percent accuracy of detection). The area under the ROC curve is an indication of the quality of the method; the larger it is the better. The ideal ROC curve has $\text{AUC} = 1$.

To build the ROC curve for method A and data set B, we apply the following procedure:

1. Sample 50 times disjoint W_1 and W_2 from B, where each window is sampled without replacement. Calculate the 50 criterion values according to A, and store them in an array $A_{\text{no change}}$.

TABLE 1
"Winning" Methods (Largest AUC) for the 12 Parameter Combinations

			Non-normalised data						Normalised data					
			$M=50$			$M=100$			$M=50$			$M=100$		
Data set	N	n	$K2$	$K3$	$K7$	$K2$	$K3$	$K7$	$K2$	$K3$	$K7$	$K2$	$K3$	$K7$
iris	150	4	■	■	○	■	■	○	■	■	■	■	■	■
thyroid	215	5	△	■	■	△	△	■	■	■	■	■	■	■
phoneme	5404	5	△	△	■	△	△	■	■	■	■	■	■	■
liver	345	6	■	△	■	△	■	△	■	■	■	■	■	○
pima	768	8	■	■	■	■	■	■	■	■	■	■	■	■
glass	214	9	■	■	△	■	△	■	■	■	■	■	■	■
breast	277	9	△	■	△	△	■	△	■	■	■	■	■	■
shuttle	58000	9	■	△	■	△	■	△	■	■	■	■	■	■
voice3[1]	238	10	■	△	■	■	■	△	■	■	■	■	■	■
voice9[1]	428	10	△	■	△	■	■	△	■	■	■	■	■	■
vowel	990	11	■	■	■	■	■	■	■	■	■	■	■	■
wine	178	13	■	■	■	■	■	■	■	■	■	■	■	■
scrapie[1]	3113	14	△	△	△	△	△	△	△	○	○	○	○	○
laryngeal1[1]	213	16	△	△	△	△	△	△	■	○	■	■	■	○
votes	232	16	■	△	△	■	△	■	■	■	■	■	■	■
laryngeal3[1]	353	16	■	■	△	△	■	■	■	■	■	■	■	■
laryngeal2[1]	692	16	△	■	■	■	△	■	■	○	■	■	■	■
pendigits	10992	16	■	■	■	■	■	■	■	■	■	■	■	■
letters	20000	16	■	■	■	■	■	■	■	■	■	■	■	■
heart	297	18	○	○	○	■	○	○	△	△	△	△	■	○
vehicle	846	18	■	■	■	■	■	■	■	■	■	■	■	■
image	2310	19	○	○	○	■	△	○	■	○	○	■	■	■
german	1000	24	△	△	△	△	△	△	△	△	○	■	△	○
wbc	569	30	■	■	△	■	■	△	■	■	■	■	■	■
palynomorphs[1]	609	31	△	△	△	△	△	△	○	○	○	○	○	○
ionosphere	351	34	○	○	○	△	■	△	■	△	○	■	■	○
satimage	6435	36	△	■	■	■	■	■	■	■	■	■	■	■
spectcontinuous	349	44	△	△	△	△	△	△	△	○	■	■	■	■
spam	4601	57	△	△	△	△	△	△	■	■	○	■	△	○
sonar	208	60	△	△	△	■	■	■	■	○	○	■	■	■
K-L distance		○	3	3	4	0	1	3	1	7	7	2	2	8
Hotelling's t^2		△	13	12	13	13	12	13	4	3	1	1	2	0
Semi-parametric LL		■	14	15	13	17	17	14	25	20	22	27	26	22

Notes:

If not indicated otherwise, the source for the data set is the UCI Machine Learning repository [20].

[1] *Private collection http://www.bangor.ac.uk/~mas00a/activities/real_data.htm.*

N *– number of instances,* n *– number of features.*

2. Sample 50 times disjoint W_1 and W_2 from B, where each window is sampled without replacement. Introduce change to W_2 by swapping two randomly selected features. Calculate the 50 criterion values according to A, and store them in an array A_{change}.
3. Pool $A_{\text{no change}}$ and A_{change} into a single array $A_{\text{criterion}}$. Set the threshold value at each element of $A_{\text{criterion}}$ and estimate the sensitivity (true positives change detections) and specificity (true negative nondetections), thereby producing 100 points on the ROC curve.

Note that we are comparing criteria rather than complete detection methods. To keep simplicity, speed and accessibility, we used K-L and SPLL criteria with the same semiparametric approximation of the density. K-means is a fast clustering algorithm which is widely available in statistical software. The outcome may vary depending on the random seeding of the K-means, so several runs are recommended, followed by selection of the clustering outcome with the best K-means criterion. Even with several runs, K-means clustering is orders of magnitude faster than EM or Monte Carlo procedures needed for other change detection algorithms. In result, the density approximation is cruder, and likely less accurate. However, time is of essence for the purposes of streaming data analysis.

We experimented with windows of equal size $M = |W_1| = |W_2| \in \{50, 100\}$. The number of clusters for K-L and SPLL was $K \in \{2, 3, 7\}$. Two sets of experiments were carried out: one with the original data and one where each feature was normalized to mean 0 and standard deviation 1. This gives $2 \times 3 \times 2 = 12$ experimental results for each data set and each method.

We chose the following ways to display a summary of the results:

- Instead of numerical values, Table 1 shows symbolically the "winner" from the three competing change detection criteria for each data set and for each of the 12 combinations of parameters. The bottom three rows of the table give the sum of the data sets for which the respective method has won.
- Figs. 1 and 2 plot the ROC curves for the experiments with nonnormalized and normalized data, respectively. The ROC curves for the individual data sets are plotted with thin lines in order to show the variability of the mean curve, also plotted in the figure with a thick line. Each of the three subfigures contains 180 thin lines: 30 data sets ×6 combination of parameter values (K and M). Fig. 3 overlays the average ROC curves in the same plot, separately for nonnormalized and the normalized data.
- To find out which data sets have proven hard for the different methods, we use glyph plots (spider plots) of the AUC for the three methods, separately for the normalized and the nonnormalized data (Figs. 4 and 5). The spikes in a glyph plot correspond to the data sets. Next, to the tip of each spike in Fig. 4, we plotted the number of features for the respective data set. The spikes revolve counterclockwise, from smaller to larger feature sets, matching the arrangement of the data sets in Table 1. If all changes in all data sets were detected correctly, the AUCs will all equal 1, and the spikes should stretch to form a circle. A circle with radius 1 is plotted on each glyph plot for reference. A good criterion will spread more toward the circle. Nonshaded glyph plots are shown together in Fig. 6 for the nonnormalized and the normalized data.

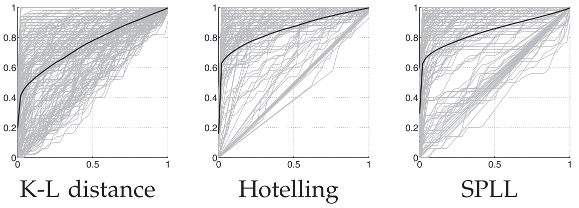

Fig. 1. ROC curves for the three methods for the nonnormalized data.

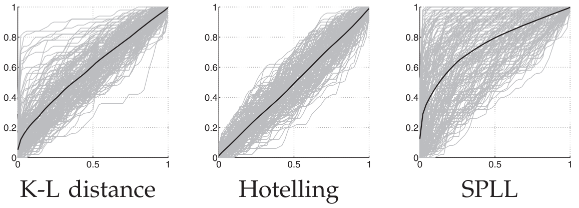

Fig. 2. ROC curves for the three methods for the normalized data.

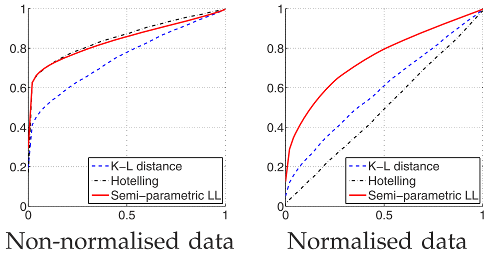

Fig. 3. Joint plots of the ROC curves for the three methods.

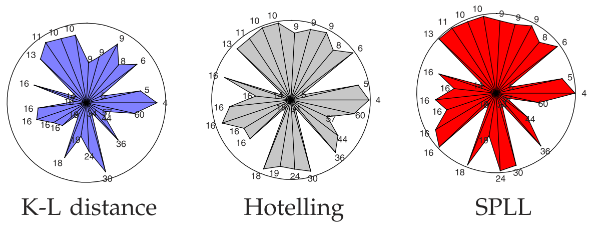

Fig. 4. Glyph plots of the AUC for the three methods for the nonnormalized data.

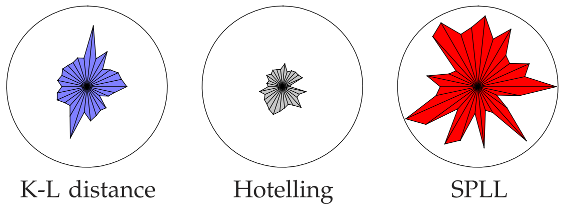

Fig. 5. Glyph plots of the AUC for the three methods for the normalized data.

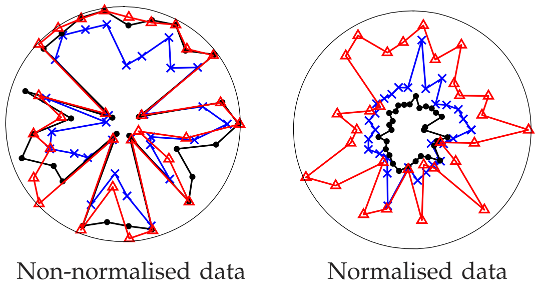

Fig. 6. Joint glyph plots of the AUC for the three methods: —●— K-L distance, —×— Hotelling, —△— SPLL.

- Finally, Table 2 gives the numerical values of the AUC, averaged across the data sets, for the 12 parameter setups.

The code for the experiment was written in Matlab, Version 7.6. The p-values for the Hotelling t^2 test were obtained from the Statistic Toolbox for Matlab.

3.2 Discussion

The semiparametric log-likelihood criterion is better than K-L and comparable to the Hotelling's t^2 test for the nonnormalized data, and better than both other criteria for the normalized data (Fig. 3). The dominance of SPLL is most visible in the right plot of Fig. 6, where the SPLL curve contains within, almost completely, the other two curves.

The accuracy of the Hotelling's t^2 test collapses dramatically for the normalized data because swapping features would shift the means of W_1 and W_2 by very little. This test's performance is inferior to that of both K-L and SPLL for the normalized data, as seen by the drastic decline of the number of wins from nonnormalized to normalized data (Table 1), and by the shrunk glyph plot in the middle in Fig. 5. By design, Hotelling's t^2 test does not depend upon the number of clusters K. Curiously, the larger sample size led to worse AUC for the normalized data.

The K-L distance criterion appeared to be the worst in our experiment. Here, we split the feature space according to a cluster structure derived from W_1, and would identify Voronoi cells even when a clear cluster structure is not present. Arguably, a different strategy of splitting the feature space for the purpose of approximating p_1 and p_2 could lead to better accuracy. Viewed on its own, this criterion behaves as expected (Table 2): larger sample M and larger number of clusters K lead to better approximation of the pmfs and give better AUC. K-L method outperforms the Hotelling's t^2 test for the normalized data bit is inferior to SPLL.

The proposed SPLL performs well on data sets with fewer features, which can be seen from the glyph plots (Figs. 4, 5, and 6)—the inward wedges that cause the star-like appearance of the plot for the nonnormalized data come from larger feature sets. The Hotelling's method suffers from a similar problem but to a lesser extent. The problem with SPLL is that some data sets manage to

TABLE 2
Average AUC across the 30 Data Sets for the 12 Parameter Combinations

Normalised	M	K	K-L distance	Hotelling's t^2	Semi-parametric LL
No	50	2	0.7065	0.8389	0.8102
No	50	3	0.7632	0.8309	0.8206
No	50	7	0.7888	0.8473	0.8191
No	100	2	0.7122	0.8625	0.8697
No	100	3	0.7595	0.8727	0.8674
No	100	7	0.8076	0.8697	0.8778
Yes	50	2	0.5279	0.5032	0.7085
Yes	50	3	0.5907	0.5193	0.7204
Yes	50	7	0.6240	0.5080	0.7028
Yes	100	2	0.5306	0.4810	0.7710
Yes	100	3	0.5911	0.4842	0.7892
Yes	100	7	0.6792	0.4788	0.7874

"fool" the clustering completely, spoil the approximation of the pdfs, and lead to results little better than chance. This can be seen from the bulky set of thin lines running along the diagonal in the right plot in Fig. 1, and even below the diagonal, corresponding to predictions that are worse than chance. In other words, when SPLL gets it wrong, it is worse than both other methods. An example of a data set where this happens is "scrapie." This set contains binary features and does not have a clear cluster structure; hence, the semiparametric model does not fit well the data.

K-L is affected by the number of the cells in the distribution K (number of components on the mixture). Larger K improves K-L notably, which indicates sensitivity to the parameter choice (Table 2). SPLL, on the other hand, is much less affected by varying K, being at the same time better than K-L. This much desired robustness comes from the fact that SPLL is based upon *an upper bound* of the log likelihood, which involves only one of the K components of the mixture (one cluster) for each data point.

Finally, working with nonnormalized data is typically avoided because features that happen to have a larger span will have also a dominant role in determining the outcome of the analyses. For normalized data, swapping two features will lead to subtle changes, mostly in the relationship between the features, which renders Hotelling's t^2 test powerless and favors SPLL.

Subtle as it may be, the change which we experimented with is abrupt as it occurs at once rather than progressively. Different issues will arise for gradual changes, the most crucial of which are the window sizes. Using a semiparametric approximation of the densities, SPLL criterion may not be as sensitive to small changes in the distributions as nonparametric criteria would be, e.g., the density test, for equal window sizes. For gradual changes, window sizes and running time complexity should be considered together for choosing a change detection method.

4 Conclusion

We view change detection from a log-likelihood perspective and show that this framework accommodates the two most commonly used criteria: Kullback-Leibler distance and Hotelling's t^2 test for equal means. Drawing upon the existing log-likelihood change detection methods (specifically the density test by Song et al.), we propose a semiparametric log-likelihood detector whose idea is to overcome the weaknesses of both criteria. Hotelling's test is not designed to, and will not be able to detect a change in variance or covariance between the variables if the means are the same. K-L criterion, on the other hand, is nonparametric and is based upon a partition of the feature space. This criterion may be too crude for picking up changes if the window size is small, and the number of cells in the partition is kept small. Our proposed criterion performs well on most data sets, and fails badly on a few.

Future research should focus on establishing and refining a change detection threshold for SPLL. There are many further questions, for example, how are the two windows determined and updated with time? While these questions can be answered separately and independently, better results can be expected if the answers are tied up with the criterion.

References

[1] C.C. Aggarwal, *Data Streams: Models and Algorithms.* Springer, 2007.

[2] M. Basseville and I.V. Nikiforov, *Detection of Abrupt Changes - Theory and Application.* Prentice-Hall, Inc., 1993.

[3] E.S. Page, "Continuous Inspection Schemes," *Biometrika,* vol. 41, pp. 100-114, 1954.

[4] M.R. Reynolds Jr. and Z.G. Stoumbos, "The SPRT Chart for Monitoring a Proportion," *IIE Trans.,* vol. 30, pp. 545-561, 1998.

[5] M.R. Reynolds Jr. and Z.G. Stoumbos, "A General Approach to Modeling CUSUM Charts for a Proportion," *IIE Trans.,* vol. 32, pp. 515-535, 2000.

[6] R.P. Adams and D.J.C. MacKay, "Bayesian Online Changepoint Detection," technical report, Univ. of Cambridge, Cambridge, UK, 2007.

[7] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting Change in Data Streams," *Proc. 30th Int'l Conf. Very Large Data Bases (VLDB '04),* 2004.

[8] S.-S. Ho, "A Martingale Framework for Concept Change Detection in Time-Varying Data Streams," *Proc. 22nd Int'l Conf. Machine Learning (ICML),* pp. 321-327, 2005.

[9] A. Bifet and R. Gavaldà, "Learning from Time-Changing Data with Adaptive Windowing," *Proc. Seventh SIAM Int'l Conf. Data Mining,* pp. 443-448, 2007.

[10] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Statistical Change Detection for Multi-Dimensional Data," *KDD '07: Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 667-676, 2007.

[11] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with Drift Detection," *Proc. 17th Brazilian Symp. Artificial Intelligence Advances in Artificial Intelligence (SBIA '04),* pp. 286-295, 2004.

[12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys,* vol. 41, no. 3, article 15, 2009.

[13] M. Ye, X. Li, and M.E. Orlowska, "Projected Outlier Detection in High-Dimensional Mixed-Attributes Data Set," *Expert Systems with Applications,* vol. 36, no. 3, pp. 7104-7113, 2009.

[14] G. Widmer and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," *Machine Learning,* vol. 23, pp. 69-101, 1996.

[15] I. Koychev and R. Lothian, "Tracking Drifting Concepts by Time Window Optimisation," *Proc. 25th SGAI Int'l Conf. Innovative Techniques and Applications of Artificial Intelligence (AI '05),* pp. 46-59, 2005.

[16] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, "An Information-Theoretic Approach to Detecting Changes in Multi-Dimensional Data Streams," *Proc. 38th Symp. Interface of Statistics, Computing Science, and Applications (Interface '06),* 2006.

[17] M. Severo and J. Gama, "Change Detection with Kalman Filter and CUSUM," *Proc. Int'l Conf. Discovery Science,* pp. 243-254, 2006.

[18] H. Hotelling, "The Generalization of Student's Ratio," *Annals of Math. Statistics,* vol. 2, no. 3, pp. 360-378, 1931.

[19] B.S. Everitt, *A Handbook of Statistical Analyses Using S-Plus,* second ed. CRC Press, 2001.

[20] A. Asuncion and D. Newman, "UCI Machine Learning Repository," http://www.ics.uci.edu/mlearn/MLRepository.html, 2007.

[21] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," Technical Report HPL-2003-4, HP Labs, Palo Alto, http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf, 2003.