Int J Data Sci Anal
DOI 10.1007/s41060-017-0045-2

REGULAR PAPER

Online detection of continuous changes in stochastic processes

Kohei Miyaguchi[1] · **Kenji Yamanishi**[1]

Received: 28 July 2016 / Accepted: 25 January 2017

Abstract We are concerned with detecting continuous changes in stochastic processes. In conventional studies on non-stationary stochastic processes, it is often assumed that changes occur abruptly. By contrast, we assume that they take place continuously. The proposed scheme consists of an efficient algorithm and rigorous theoretical analysis under the assumption of continuity. The contribution of this paper is as follows: We first propose a novel characterization of processes for continuous changes. We also present a time- and space-efficient online estimator of the characteristics. Then, employing the proposed estimate, we propose a method for detecting changes together with a criterion for tuning its hyper-parameter. Finally, the proposed methods are shown to be effective through experimentation involving real-life data from markets, servers, and industrial machines.

Keywords Continuous change · Local linear regression · Online change detection

1 Introduction

1.1 Motivation for and purposes of this paper

This paper addresses the issue of detecting changes in non-stationary stochastic processes. Specifically, we focus on the online setting. That is, when given a time series sequentially, we are concerned with detecting change points in a sequential manner. In conventional studies on change detection, researchers have sought to detect time points when the statistical models of data suddenly change [2,9]. In real situations, however, changes may not occur abruptly, but rather incrementally over some successive periods of time. We call such changes *continuous changes*. Actually, there exist many phenomena that may be characterized by continuous changes (e.g., seismic motion before and after an earthquake, and stock prices in markets).

In this paper, we consider the problem of detecting time points when continuous changes start. This problem is worthwhile studying from a practical point of view, because the starting point of continuous changes can be considered a symptom of big changes that will occur in future. Therefore, detecting them in early stages can lead to predictions of important events in the future. Despite the importance of continuous changes, it has not yet been explored how to detect them. It is certain that the detection of continuous changes has been covered in some previous studies in the context of concept drift [3,7]. Nevertheless, they have been thought of as succession of relatively small, 'abrupt' changes.

There are three purposes of this paper, and they are summarized as follows: The first is to introduce a framework for detecting continuous changes. In it, we define the magnitude of continuous changes and formalize the problem of measuring them. The second purpose is to propose an efficient algorithm for detecting continuous changes. The third pur-

This paper is an extended version of the DSAA'2015 Long Presentation paper [16].

✉ Kohei Miyaguchi
kohei_miyaguchi@mist.i.u-tokyo.ac.jp

Kenji Yamanishi
yamanishi@mist.i.u-tokyo.ac.jp

[1] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Published online: 21 February 2017

pose is to empirically demonstrate the effectiveness of our proposed algorithm in comparison with existing algorithms.

1.2 Novelty of this paper

The significance and novelty of this paper are summarized as follows:

(1) A *novel framework for continuous change detection*: We first define a measure of continuous changes for parametric models. It is designed on the basis of Kullback–Leibler divergence between the model before and after a change point, which has been known as a typical measure of change. In our framework, we assume that the parameter value changes smoothly over time. We employ a weighted linear regression to model it. We thereby derive a novel measure of change by plugging the localized maximum likelihood estimates of the parameter and its rate of change into the approximation of the Kullback–Leibler divergence. We justify this measure theoretically by proving that it is invariant with respect to parameterization. We show several examples of calculations of this measure for parametric models, such as the independent exponential family and auto-regression model.

(2) A *novel efficient algorithm for online detection of continuous changes*: Real-time change point detection is more favorable than batch detection for a variety of applications. We develop an efficient algorithm for detecting continuous changes in an online fashion. It is designed on the basis of the following two key ideas: (i) to efficiently calculate change scores by utilizing a recurrence relation for the weighted linear regression and (ii) to calculate a threshold for scores dynamically. We establish an alarm when the score exceeds the threshold, where the threshold may change over time. Combining (i) with (ii) above yields an online algorithm for detecting continuous changes in the computation time $O(N)$ for the total data size N.

(3) A *novel criterion for choosing hyper-parameters*: We also present a criterion that measures the fitness of the proposed model without knowing whether there are any changes in data. This enables us to automatically tune the hyper-parameters of the proposed method, such that one is no longer worried about the choice of hyper-parameters.

(4) *Empirical demonstration of the effectiveness of our method*: We used synthetic and real datasets to compare our method to existing online change point detection methods in terms of how accurately and how early they can detect continuous changes. Specifically, we applied our method to malware detection, economic event detection, and industrial incident detection. We therefore show that our proposed method, together with the hyper-parameter-choosing criterion, is able to detect symptoms of important events significantly earlier than other methods.

The following updates are made compared with the preliminary version: (a) more comprehensive theoretical treatments are supplied for the proposed method, especially for the thresholding algorithms (Sect. 4.2); (b) a criterion for choosing hyper-parameters is provided, and its validity is empirically shown (Sects. 5 and 6.2); and (c) applicability of the proposed method on multivariate time series is also verified in the experiments together with a novel change localization technique (Sect. 6.2.3).

1.3 Related work

Many methods have been proposed for detecting changes that happen abruptly in stochastic processes [2,5,8,10,15]. Online methods for detecting them have also been developed in [1,5,13,18–21]. Those methods somehow test whether the two sample sets clipped from neighboring two sliding windows are generated from an identical distribution. It is assumed that changes occur at some discrete time points and that the generated distribution is piecewise stationary. Therefore, it follows that they can unnecessarily degrade their performance in detecting continuous changes, which take place over some periods of time rather than a discrete point.

Change detection is related to the topic of *concept drift* (see, for example, [6,7,14,22]). Changes that occur gradually over time are called *incremental changes* in the context of concept drifts [7,22], but there are no studies on online detection algorithms tailored for incremental changes to the best of our knowledge. Recently, changes in the rate of change have been studied in the scenario of volatility shift change detection [12]. This implicitly assumes that changes can be continuous. Our work differs in that ours deals with the rate of change with continuously changing smooth models, while [12] deals with that with a piecewise stationary model.

The remainder of this paper is organized as follows: Sect. 2 introduces a measure for continuous changes. Section 3 gives a time- and space-efficient algorithm for computing the proposed estimates, as well as a number of examples for some statistical models. In Sect. 4, we present a method for detecting continuous changes employing them. Section 5 provides us with a criterion for choosing hyper-parameters of the proposed method. Section 6 shows experimental results on synthetic and real datasets. Section 7 gives concluding remarks.

2 Estimating the magnitude of continuous changes

2.1 Measures of magnitude of changes

In this section, we introduce a measure for changes in stochastic processes. Consider a parametric space of ergodic Markov models, $\mathcal{M} = \{p(\cdot|\cdot;\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$, and a stochastic process $\mathcal{P} = \{X_t\}_{t=0}^{n-1}$. Let x_t be an instance of X_t drawn from parameter $\theta_t \in \Theta$ given past instances $x_0^{t-1} = [x_0, x_1, \ldots, x_{t-1}]$, such that

$$\mathrm{d}P\left(x_t|x_0^{t-1}\right) = p\left(x_t|x_0^{t-1};\theta_t\right)\mathrm{d}x_t,$$

or, we may write,

$$X_t \sim \theta_t.$$

for $t = 0, 1, \ldots, t-1$. For simplicity, we here suppose that the model space $\mathcal{M}$ satisfies some smoothness and standard regularity conditions. Specifically, the negative logarithmic likelihood $-\log p(x_t|x_0^{t-1};\theta)$ is assumed to be differentiable continuous, bounded, and strongly convex with respect to parameter θ. It is also assumed that the Fisher information at each point of the parametric model Θ is (symmetric) positive definite,

$$I(\theta) \overset{\mathrm{def}}{=} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_\theta\left[-\frac{\partial^2}{\partial\theta^2}\log p\left(X_0^{n-1};\theta\right)\right] \succ 0, \tag{1}$$

where $\mathbb{E}_\theta[\cdot]$ denotes the expectation with respect to $X_k \sim \theta$ $(\forall k)$, and $p(x_0^{n-1};\theta) = \prod_{t=0}^{n-1} p(x_t|x_0^{t-1};\theta)$ denotes the joint probability mass (or density) function.

We stipulate that time t is a *change point in the process* $\mathcal{P}$ if and only if $\theta_t \neq \theta_{t-1}$. Then, it is reasonable to estimate whether t is a change point by estimating some divergence measure $d(\theta_t, \theta_{t-1})$ and by comparing it to a threshold β. The Kullback–Leibler (KL) divergence is one of the most common divergence measures for such purpose. Let y_t be the KL divergence between probability densities specified by θ_t and θ_{t-1}

$$\begin{aligned} y_t &\overset{\mathrm{def}}{=} D(\theta_t\|\theta_{t-1}) \\ &= \lim_{n\to\infty}\frac{1}{n}\mathbb{E}_{\theta_t}\left[\log\frac{p\left(X_0^{n-1};\theta_t\right)}{p\left(X_0^{n-1};\theta_{t-1}\right)}\right]. \end{aligned} \tag{2}$$

Henceforth, we refer to y_t as the *magnitude of change* at time t and consider how to estimate y_t.

We say that a change is *discrete* if no change occurs before and after the corresponding point. The magnitude of a discrete change can be estimated in the following manner: Assuming that the data distribution is stationary before and after the change point t, one can estimate θ_t and θ_{t-1}, respectively, with, for example, the technique of maximum likelihood estimation. Plugging those estimates into y_t gives a simple estimator of the magnitude of the change

$$\hat{y}_t = D(\hat{\theta}_t\|\hat{\theta}_{t-1}).$$

Then, one can detect discrete changes by finding time t such that $\hat{y}_t$ is sufficiently large. A substantial number of previous methods for detecting change points can be viewed as a special case of the above scheme.

By contrast, we say that a change is *continuous* if a change occurs over some successive period of time. In this case, direct estimations of θ_t and θ_{t-1} as described above do not make sense since every step of the change is surrounded with other small changes. In order to deal with continuous changes, we consider another characterization of the measure y_t, rather than parameters themselves. In this paper, we estimate y_t through the following proxy measure:

$$z_t \overset{\mathrm{def}}{=} \delta_t^\top I(\theta_t)\delta_t, \tag{3}$$

where $\delta_t \overset{\mathrm{def}}{=} \theta_t - \theta_{t-1}$ is the rate of the parameter change. In fact, this proxy works well where changes are relatively small; the proxy measure coincides with $2y_t$ in the limit of $\|\delta_t\| \to 0$. This is demonstrated by expanding y_t with respect to δ_t:

$$\begin{aligned} 2y_t &= 2D(\theta_t\|\theta_t - \delta_t) \\ &= 2\cdot\frac{1}{2}\delta_t^\top I(\theta_t)\delta_t + \mathrm{o}(\|\delta_t\|^2) \\ &= z_t(1+\mathrm{o}(1)). \end{aligned} \tag{4}$$

Therefore, it suffices for estimating whether t is a change point (i.e., $z_t \neq 0$). Note that the plug-in estimation of z_t requires estimating δ_t, which requires the smoothness of the parameter sequence θ_0^{n-1} in exchange of its piecewise stationarity. Hence, it fits better for detecting continuous changes.

2.2 Estimating linearly changing parameters

Below we state the intuition of the proposed method for estimating the proxy magnitude of changes $z_t = \delta_t I(\theta_t)\delta_t$ in a very simple case. We assume that the process changes linearly over time. In other words, we assume here that the increments of the parameter, δ_k $(k = 1, \ldots, n-1)$, are identical to one another. Under this seemingly strong assumption, we have the maximum likelihood estimator of θ_t and δ_t,

$$(\hat{\theta}_t, \hat{\delta}_t) = \arg\min_{(\theta,\delta)} \frac{1}{n}\sum_{k=0}^{n-1} L_k(\theta + (k-t)\delta), \tag{5}$$

where $L_k(\theta) = -\log p(x_k|x_0^{k-1};\theta)$ denotes the logarithmic loss for θ relative to the instance x_k. It is immediately apparent that there exists the unique minimizer for (5).

Proposition 1 *$\sum_{k=0}^{n-1} L_k(\theta+(k-t)\delta)$ is strictly convex with respect to θ and δ. Therefore, it has a unique minimizer.*

Proof It immediately follows from that the objective function is a positive-weighted summation of strongly convex functions $L_k(\theta+(k-t)\delta)$ with respect to θ and δ. □

Owing to the uniqueness, the minimization is reduced to solving the following equations:

$$\sum_{k=0}^{n-1} \frac{\partial L_k}{\partial\theta}(\hat{\theta}_t + (k-t)\hat{\delta}_t) = 0,$$
$$\sum_{k=0}^{n-1} (t-k)\frac{\partial L_k}{\partial\theta}(\hat{\theta}_t + (k-t)\hat{\delta}_t) = 0. \quad (6)$$

2.3 Estimating continuously changing parameters

It is not reasonable to assume that the process changes linearly over time in most practical cases. We rather assume that the process changes linearly over a local range of time. Under this more realistic assumption, we present the corresponding general form of the proposed method. Consider the minimization (5) putting a weight on each term that appears in the sum. By introducing a weight sequence $\{w_i\}_{i\in\mathbb{R}}$ $(0 \le w_i < \infty)$, we are able to fit the linear model locally. We define the *localized maximum likelihood estimator* by

$$(\hat{\theta}_t, \hat{\delta}_t) = \arg\min_{\theta,\delta} \sum_{k\in\Lambda} w_{k-t} L_k(\theta + (k-t)\delta) + g(\theta,\delta), \quad (7)$$

where $\Lambda \subset [0,n) \subset \mathbb{R}$ is the finite set of extended indices with even intervals that denotes the points of observations on the real line. Here, we added a convex (and smooth) regularizing term $g(\theta,\delta)$ to stabilize the estimate for a small sample size $|\Lambda|$. The regularizer $g(\theta,\delta)$ is equivalent to the prior $\propto \exp\{-g(\theta,\delta)\}$. Therefore, $(\hat{\theta}_t, \hat{\delta}_t)$ is a MAP estimate. Since w_i is nonnegative and $g(\theta,\delta)$ is convex, the uniqueness of the minimizer can be shown in the same vein as in Proposition 1. Thus, it can be solved with

$$\sum_{k\in\Lambda} w_{k-t}\frac{\partial L_k}{\partial\theta}(\hat{\theta}_t + (k-t)\hat{\delta}_t) + g_\theta(\theta,\delta) = 0,$$
$$\sum_{k\in\Lambda} (t-k) w_{k-t}\frac{\partial L_k}{\partial\theta}(\hat{\theta}_t + (k-t)\hat{\delta}_t) + g_\delta(\theta,\delta) = 0, \quad (8)$$

where $g_x = \frac{\partial g}{\partial x}$ denotes the derivative of $g(\theta,\delta)$ with respect to x.

Here, we design the weights and regularizer properly in order to infer the parameter value and its time derivative at point t. In order to fit the linear model locally, we reduce w_i when $|i|$ becomes sufficiently large. In addition, to take the balance between both sides of time point t, we calibrate the first-order moment of the weights,

$$\sum_{k\in\Lambda} (k-t) w_{k-t} = 0. \quad (9)$$

Note that, conversely, t can be seen as the center of the weights given by

$$t = \frac{\sum_{k=0}^{n-1} k w_{k-t}}{\sum_{k=0}^{n-1} w_{k-t}}. \quad (10)$$

This localized maximum likelihood estimate is also able to detect discrete changes. For instance, assume that each sample x_k is independent of the others and that there exists a discrete change point t such that there exist two parameters θ_- and θ_+ $(\theta_- \ne \theta_+)$ and that $\theta_k = \theta_-$ if $k < t$, otherwise $\theta_k = \theta_+$. Then, as in the following proposition, the resulting estimate of the magnitude $\hat{z}_t$ is bounded away from zero in probability in the limit of large $|\Lambda|$. It implies that the proposed estimate works well even if changes are discrete.

Proposition 2 *Assume that the weight sequence $\{w_i\}_{i\in\mathbb{R}}$ uniformly converges to a Riemann integrable function $\{\bar{w}_i\}_{i\in\mathbb{R}}$, where $0 < \int_0^n \bar{w}_{k-t} dk < \infty$. Then, under the preceding assumption of discrete changes, $\hat{\delta}_t$ given in (7) is bounded away from zero—except with a small probability for large $|\Lambda|$. Therefore, the estimated magnitude $\hat{z}_t$ is also asymptotically bounded away from zero in probability.*

Proof See "Appendix 1."

3 Approximating estimates

Although the idea behind the proposed estimate is very simple, there is a difficulty specific to the problem of detecting changes: Since the solution of (8) is often analytically intractable, it becomes computationally intractable as the number of observations $|\Lambda|$ grows. Even if it is calculated, there is another undesirable property: the resulting value of the estimated measure, $\hat{\delta}_t^\top I(\hat{\theta}_t)\hat{\delta}_t$, depends on the parameterization of model space Θ.

Utilizing the smoothness of the loss function L_k, we now derive an approximation of the estimate. Applying the technique of the local linear approximation on (8) yields

$$L_n^{0,1}(\hat{\theta}_t) + \hat{\delta}_t^\top L_n^{1,2}(\hat{\theta}_t) + g_\theta(\hat{\theta}_t, \hat{\delta}_t) = 0,$$
$$L_n^{1,1}(\hat{\theta}_t) + \hat{\delta}_t^\top L_n^{2,2}(\hat{\theta}_t) + g_\delta(\hat{\theta}_t, \hat{\delta}_t) = 0, \quad (11)$$

where we define $L_n^{j,l}(\theta) \overset{\text{def}}{=} \sum_{k=0}^{n-1} w_{k-t}(k-t)^j \left(\frac{\partial}{\partial\theta}\right)^l L_k(\theta)$. Note that we rescale the time indices of observation such that the weights $\{w_i\}$ are defined on integers. The solution to the approximated equations is not only computationally tractable in many cases, but also has the following noteworthy property.

Proposition 3 *Let* $(\hat{\theta}_t, \hat{\delta}_t)$ *be the approximated estimate solving* (11) *and let* $\hat{z}_t = \hat{\delta}_t^\top I(\hat{\theta}_t)\hat{\delta}_t$. *Then,* $\hat{z}_t$ *is invariant with respect to the parameterization of model space* Θ *if condition* (1) *holds.*

Proof See "Appendix 2."

So far we have presented a general scheme for estimating the magnitude of continuous changes. In the rest of this section, we give two examples of how to calculate estimates for concrete statistical models: the independent exponential family and Gaussian autoregressive (AR) models.

3.1 Independent exponential family

Suppose Θ is a member of the exponential family. We have

$$p(x_t|x_0^{t-1}; \theta_t) = \exp\{\theta_t^\top T(x_t) - A(\theta_t) - B(x_t)\},$$

where $T(\cdot)$ denotes the sufficient statistics, and then

$$\frac{\partial L_k}{\partial\theta}(\theta) = -T(x_k)^\top + \tau(\theta)^\top, \quad \frac{\partial^2 L_k}{\partial\theta^2}(\theta) = \frac{\partial\tau}{\partial\theta}(\theta),$$

where $\tau(\theta) \overset{\text{def}}{=} \frac{\partial}{\partial\theta}A(\theta) = \mathbb{E}_\theta[T(X)]$ denotes the expectation parameter. The regularizer is chosen as follows:

$$g(\theta, \delta) = -\gamma_0(\theta^\top\tau_0 - A(\theta)) + \frac{\gamma_1}{2}\delta^\top K\delta \tag{12}$$

Therefore, Eq. (11) is reduced to

$$\begin{aligned} 0 &= \gamma_0(-\tau_0 + \tau(\hat{\theta}_t)) \\ &\quad + \sum_{k=0}^{n-1} w_{k-t}\left\{-T(x_k) + \tau(\hat{\theta}_t) + (k-t)\frac{\partial\tau}{\partial\theta}(\hat{\theta}_t)\hat{\delta}_t\right\}, \\ 0 &= \gamma_1 K\hat{\delta}_t \\ &\quad + \sum_{k=0}^{n-1} w_{k-t}(k-t)\left\{-T(x_k) + \tau(\hat{\theta}_t) + (k-t)\frac{\partial\tau}{\partial\theta}(\hat{\theta}_t)\hat{\delta}_t\right\}. \end{aligned}$$

Thus, by taking $\hat{\tau}_t = \tau(\hat{\theta}_t)$, $\hat{\xi}_t = \frac{\partial\tau}{\partial\theta}(\hat{\theta}_t)\hat{\delta}_t$ and $K = I(\hat{\theta}_t)$, we have

$$\begin{bmatrix} T_n^0 + \gamma_0\tau_0 \\ T_n^1 \end{bmatrix} = W(n)\begin{bmatrix} \hat{\tau}_t \\ \hat{\xi}_t \end{bmatrix}, \tag{13}$$

where

$$W(n) \overset{\text{def}}{=} \begin{bmatrix} W_n^0 + \gamma_0 & W_n^1 \\ W_n^1 & W_n^2 + \gamma_1 \end{bmatrix} \otimes I_d,$$

$$W_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t},$$

$$T_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t}T(x_k).$$

Because we calibrated the weights so that $W_n^1 = 0$, Eq. (13) can be further reduced to

$$\hat{\tau}_t = \frac{T_n^0 + \gamma_0\tau_0}{W_n^0 + \gamma_0}, \quad \hat{\xi}_t = \frac{T_n^1}{W_n^2 + \gamma_1}. \tag{14}$$

Remember that the magnitude of change can be calculated as $\hat{z}_t = \hat{\xi}_t^\top \tilde{I}(\hat{\tau}_t)\hat{\xi}_t$, where $\tilde{I}(\cdot)$ denotes the Fisher information with respect to τ, since it is invariant to parameter transformation as proved in Proposition 3.

Note that the resulting estimate is nothing other than the locally weighted least squares regression on sufficient statistics $\{T(x_t)\}$, such that

$$(\hat{\tau}_t, \hat{\xi}_t) = \arg\min \sum_{k=0}^{n-1} w_{k-t}\|T(x_k) - \tau - (k-t)\xi\|_2^2, \tag{15}$$

if $\gamma_0 = \gamma_1 = 0$. This implies that, even if we employ a misspecified statistical model Θ, the estimate measures changes by projecting data to the space of sufficient statistics. For example, we can employ a multivariate independent Gaussian model:

$$\Theta \overset{\text{def}}{=} \left\{(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d\times d} \mid \Sigma \succ 0, \Sigma = \Sigma^\top\right\},$$

such that

$$p(x_t|x_0^{t-1}; \mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2}(x_t-\mu)^\top\Sigma^{-1}(x_t-\mu)\right\}}{\sqrt{2\pi}^d|\Sigma|^{1/2}}.$$

Then, the sufficient statistics are given by $T(x) = x \oplus \mathrm{vec}(xx^\top)$, which are the first and second moments of the data. Therefore, it is able to detect changes regardless of the true distributions, when the data changes in terms of its lowest two moments x and $xx^\top$.

3.2 Gaussian autoregressive models

Let Θ be an autoregressive model with Gaussian noise of order p. The conditional density function is given by

$$p(x_t|x_0^{t-1}; \nu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x_t - \nu^\top u_t)^2\right\},$$

where $\nu \in \mathbb{R}^{p+1}$ and $u_t \overset{\text{def}}{=} \begin{bmatrix} 1 & x_{t-1} & x_{t-2} & \cdots & x_{t-p} \end{bmatrix}^\top$. Let φ denote the trivial parameterization, (ν, σ^2), and θ denote the natural parameterization, $(\nu\sigma^{-2}, -\sigma^{-2}/2)$. Then, by analogy with the exponential family, we have

$$\frac{\partial L_k}{\partial\theta}(\theta) = -T(x_l, u_k)^\top + \tau(\theta, u_k)^\top,$$
$$\frac{\partial^2 L_k}{\partial\theta^2}(\theta) = \tau'(\theta, u_k)\frac{\partial\varphi}{\partial\theta},$$

where

$$T(x_k, u_k) \overset{\text{def}}{=} \begin{bmatrix} x_k u_k \\ x_k^2 \end{bmatrix},$$
$$\tau(\theta, u_k) \overset{\text{def}}{=} \mathbb{E}_\theta[T(X_k, u_k)] = \begin{bmatrix} (u_k u_k^\top)\nu \\ \nu^\top(u_k u_k^\top)\nu + \sigma^2 \end{bmatrix},$$
$$\tau'(\theta, u_k) \overset{\text{def}}{=} \frac{\partial\tau}{\partial\varphi}(\theta, u_k) = \begin{bmatrix} u_k u_k^\top & \mathbf{0} \\ 2\nu^\top(u_k u_k^\top) & 1 \end{bmatrix}.$$

Now consider the following regularization over (θ, δ):

$$g(\theta, \delta) = \gamma_0(-\theta^\top\tau_0 + A_0(\theta)) + \frac{\gamma_1}{2}\delta^\top K\delta,$$
$$\tau_0 = \begin{bmatrix} \mathbf{0} \\ b_0 \end{bmatrix} \quad (b_0 \in \mathbb{R}),$$
$$A_0(\theta) = \frac{\nu^\top\nu}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2. \tag{16}$$

Replacing $(\partial L/\partial\theta)$ and $(\partial^2 L/\partial\theta^2)$ in (11) employing the above equations, we then have

$$\begin{bmatrix} C_n^0 \\ S_n^0 + \gamma_0 b_0 \end{bmatrix} = \begin{bmatrix} (U_n^0 + \gamma_0 I_{p+1})\hat{\nu}_t \\ \hat{\nu}_t^\top(U_n^0 + \gamma_0 I_{p+1})\hat{\nu}_t + \hat{\sigma}_t^2(W_n^0 + \gamma_0) \end{bmatrix} + \begin{bmatrix} U_n^1 & \mathbf{0} \\ 2\hat{\nu}_t^\top U_n^1 & W_n^1 \end{bmatrix}\begin{bmatrix} \hat{\delta}_{\nu,t} \\ \hat{\delta}_{\sigma^2,t} \end{bmatrix},$$
$$\begin{bmatrix} C_n^1 \\ S_n^1 \end{bmatrix} = \begin{bmatrix} U_n^1\hat{\nu}_t \\ \hat{\nu}_t^\top U_n^1\hat{\nu}_t + \hat{\sigma}_t^2 W_n^1 \end{bmatrix} + \begin{bmatrix} U_n^2 + \gamma_1 I_{p+1} & \mathbf{0} \\ 2\hat{\nu}_t^\top(U_n^2 + \gamma_1 I_{p+1}) & W_n^2 + \gamma_1 \end{bmatrix}\begin{bmatrix} \hat{\delta}_{\nu,t} \\ \hat{\delta}_{\sigma^2,t} \end{bmatrix}, \tag{17}$$

where

$$C_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t} x_k u_k,$$
$$S_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t} x_k^2,$$
$$U_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t} u_k u_k^\top,$$
$$W_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t},$$
$$K = \begin{bmatrix} I_{p+1} & 2\hat{\nu}_t \\ 2\hat{\nu}_t^\top & 4\hat{\nu}_t^\top\hat{\nu}_t + 2 \end{bmatrix},$$

for $j = 0, 1$. Noticing the upper row in Eq. (17), we have

$$\begin{bmatrix} \hat{\nu}_t \\ \hat{\delta}_{\nu,t} \end{bmatrix} = U(n)^{-1}\begin{bmatrix} C_n^0 \\ C_n^1 \end{bmatrix}, \tag{18}$$

where

$$U(n) \overset{\text{def}}{=} \begin{bmatrix} U_n^0 + \gamma_0 I_{p+1} & U_n^1 \\ U_n^1 & U_n^2 + \gamma_1 I_{p+1} \end{bmatrix}.$$

Note that $U(n)$ is invertible if $\gamma_0, \gamma_1 > 0$ since it is positive definite,

$$U(n) = \sum_{k=0}^{n-1} w_{k-t}\begin{bmatrix} 1 & k-t \\ k-t & (k-t)^2 \end{bmatrix} \otimes u_k u_k^\top + \begin{bmatrix} \gamma_0 & \\ & \gamma_1 \end{bmatrix} \otimes I_{p+1}.$$

The rest of the estimator, $(\hat{\sigma}_t^2, \hat{\delta}_{\sigma^2,t})$, can be calculated easily from (17) and (18).

4 Algorithm for detecting changes

In this section, we describe an efficient online algorithm for detecting continuous changes utilizing the proposed estimate $\hat{z}_t = \hat{\delta}_t^\top I(\hat{\theta}_t)\hat{\delta}_t$ given by (11). Specifically, in monitoring changes in a data stream, it is often desirable to detect changes as soon as possible after they happen. To this end, one may choose weights such that later points of observation are assigned with more weights. One example of such weights is exponentially discounting sequence. Now let $\{w_i\}$ be an exponentially discounting weight sequence given by

$$w_i \overset{\text{def}}{=} (1-r)^{-i}$$

where $0 < r \ll 1$ denotes the *discounting rate*. The rate r can be seen as the *resolution* parameter that controls the time constant $c = -1/\log(1-r) \approx r^{-1}$ in which the weight decays to $1/e$ times. This means that each observation remains to be effective on the value of $\hat{z}(n)$ during the period of length proportional to c.

We regard the estimate $\hat{z}_t$ as a change score of the stream given observations until point n, where t is automatically given by (10). To clarify the dependence of $\hat{z}_t$ to n, we refer $\hat{z}(n)$ as to $\hat{z}_t$ from now on. We also define $t(n)$, $\hat{\theta}(n)$, and $\hat{\delta}(n)$ in the same vein.

In the remainder of this section, we first give an efficient method for solving (11) with exponentially discounting weights. We then give a procedure for generating alarms to indicate changes.

4.1 Computing weighted summations

We here propose an efficient method for computing the estimator satisfying (11). Our focus is on computing a weighted sum that frequently appears in the estimation process:

$$f_n^m \overset{\text{def}}{=} \sum_{k=0}^{n-1} (k-t)^m w_{k-t} f(x_0^k),$$

for a function defined over the space of the data, f. Let Δ denote the lag in detection, $\Delta = n - t(n)$. We have the following recurrence relation:

$$\begin{aligned} f_n^m &= \sum_{k=\Delta}^{n+\Delta-1} (k-n)^m (1-r)^{n-k} f(x_0^{k-\Delta}) \\ &= C_m f(x_0^{n-1}) + \sum_{k=\Delta}^{n+\Delta-2} (k-n)^m (1-r)^{n-k} f(x_0^{k-\Delta}) \\ &= C_m f(x_0^{n-1}) + (1-r) \sum_{j=0}^{m} (-1)^j \binom{m}{j} f_{n-1}^j, \end{aligned} \tag{19}$$

where $C_m \overset{\text{def}}{=} (\Delta - 1)^m (1-r)^{1-\Delta}$. Because we only need $m \le 2$ for the proposed method, it follows that updating f_n^m given $\left\{f_{n-1}^m\right\}_{m=0,1,2}$ can be done within a constant time and that the computational complexity of entire sequence $\{f_n^m\}_{n=1}^N$ achieves the optimal rate $\mathrm{O}(N)$ if the evaluation of f can be done within a constant time.

4.2 Making alarms with threshold

We now connect the estimate $\hat{z}(n)$ with an algorithm for detecting continuous changes. We activate an alarm when $\hat{z}(n)$ exceeds a threshold where the desirable value of the threshold will change over time. This is because $\hat{z}(n)$ is biased in the positive direction, even if there is no change occurred and the bias could be vary as n increases.

Let $\bar{z}(n)$ be an estimate for the expected value of $\hat{z}(n)$ in the case of no change—i.e., $X_k \sim \theta$ for all $0 \le k \le n-1$—

$$\bar{z}(n) \approx \mathbb{E}_\theta[\hat{z}(n)].$$

Then, we raise alarms if $\hat{z}(n)$ exceeds $\beta \bar{z}(n)$; the detection alarm for proposed method is given by

$$a_n \overset{\text{def}}{=} \begin{cases} 0 & (\hat{z}(n) \le \beta \bar{z}(n)) \\ 1 & (\hat{z}(n) > \beta \bar{z}(n)) \end{cases},$$

with a constant $\beta > 0$. In other words, we calculate a scale-corrected score of change s_n, such that

$$s_n \overset{\text{def}}{=} \frac{\hat{z}(n)}{\bar{z}(n)},$$

Algorithm 1 LLR(r, γ_0, γ_1)

```
Input: data stream $\{x_k\}_0^\infty$ and threshold $\beta$
for $n = 0, 1, \ldots$ do
  Solve regression (11) to get $(\hat{\theta}(n), \hat{\delta}(n))$
  Compute score $\hat{z}(n) = \hat{\delta}(n)^\top I(\hat{\theta}(n))\hat{\delta}(n)$
  Compute scale-corrected score $s_n = \hat{z}(n)/\bar{z}(n)$
  if $s_n > \beta$ then
    Raise alarm
  end if
end for
```

Algorithm 2 LLR(r, γ_0, γ_1) for exponential family

```
Input: data stream $\{x_k\}_0^\infty$ and threshold $\beta$
for $n = 0, 1, \ldots$ do
  Update $T_n^0$ and $T_n^1$ using $T(x_n)$
  Update $W_n^0, W_n^1, W_n^2, V_n^0, V_n^1$ and $V_n^2$
  Compute LLR estimates:
  $\hat{\tau}(n) = (T_n^0 + \gamma_0\tau_0)/(W_n^0 + \gamma_0)$, $\hat{\xi}(n) = T_n^1/(W_n^2 + \gamma_1)$
  Compute score $\hat{z}(n) = \hat{\xi}(n)^\top \tilde{I}(\hat{\tau}(n))\hat{\xi}(n)$
  Compute scale-corrected score $s_n = (W_n^2)^2\hat{z}(n)/V_n^2 d$
  if $s_n > \beta$ then
    Raise alarm
  end if
end for
```

and we raise an alarm if and only if s_n exceeds a constant β.

For the independent exponential family, estimate $\bar{z}(n)$ can be approximated by χ^2 statistics,

$$\begin{aligned} \bar{z}(n) &= \mathbb{E}_\theta\left[\mathrm{Tr}\left\{\tilde{I}(\hat{\tau}(n))\hat{\xi}(n)\hat{\xi}(n)^\top\right\}\right] \\ &\approx \mathrm{Tr}\left\{\tilde{I}(\tau)\mathbb{E}_\theta\left[\hat{\xi}(n)\hat{\xi}(n)^\top\right]\right\} \\ &= \frac{\mathrm{Tr}\left\{\tilde{I}(\tau)\mathbb{E}_\theta\left[T_n^1 T_n^{1\top}\right]\right\}}{(W_n^2)^2} = \frac{V_n^2}{(W_n^2)^2}\chi_d^2, \end{aligned} \tag{20}$$

whose mean is analytically given as $V_n^2 d/(W_n^2)^2$ where d is the dimensionality of the parameter space Θ, and $V_n^j \overset{\text{def}}{=} \sum_{k=0}^{n-1}(k-t)^j w_{k-t}^2$. Here, V_n^j is also computed employing a recurrence formula similar to (19). Note that this works well when each statistic $T(X_i)$ is uncorrelated with the others. If they are strongly correlated, one has to consider a further correction on s_n. If the correlation is time-invariant, however, the correction can be offset by the threshold β.

The entire algorithm for continuous change detection is shown in Algorithm 1. For reference, the same algorithm specialized for the independent exponential family is also shown in Algorithm 2.

5 Choosing hyper-parameters

Because change detection is a task of unsupervised learning, we cannot "train" hyper-parameters in an explicit manner. Therefore, we propose criteria for choosing those parameters.

The criteria are to be minimized in reference to “training period” of data.

The proposed method has two hyper-parameters, r and β. We focus on choosing the optimal discounting rate r that induces the best estimate of (θ_n, δ_n). In contrast, we believe that there is no *best* value of the threshold β, since it controls the trade-off relation between the false-positive and false-negative rate of the alarm $\{a_n\}$. The optimal balance of these false-alarm rates should be determined at a higher level (e.g., by users).

The hyper-parameter r controls the trade-off between the accuracy and delay (i.e., the variance and bias) of the alarms generated by the proposed method. The lag in the alarm in the proposed method, $n - t(n)$, coincides approximately with r^{-1} according to (10),

$$n - t(n) = \frac{\sum_{k=0}^{n-1}(n-k)(1-r)^{-k}}{\sum_{k=0}^{n-1}(1-r)^{-k}} \approx r^{-1}, \tag{21}$$

where the last approximation holds for large n and small r. Thus, small r biases the estimate $(\hat{\theta}(n), \hat{\delta}(n))$ and delays the detection. On the other hand, small r also reduces the variance of the estimate. For example, the variance of the estimate calculated with a model for the independent exponential family is evaluated as

$$\begin{aligned} \mathbb{V}\left[\hat{\theta}(n)\right] &= \frac{\sum_{k=0}^{n-1} w_{k-t}^2 \mathbb{V}[T(X_k)]}{(W_n^0)^2} \approx \frac{r}{2} C_0, \\ \mathbb{V}\left[\hat{\delta}(n)\right] &= \frac{\sum_{k=0}^{n-1}(k-t)^2 w_{k-t}^2 \mathbb{V}[T(X_k)]}{(W_n^2)^2} \approx \frac{r^3}{4} C_2, \end{aligned} \tag{22}$$

where C_0 and C_2 denote the averaged covariance of sufficient statistics,

$$\begin{aligned} C_0 &= \frac{\sum_{k=0}^{n-1} w_{k-t}^2 \mathbb{V}[T(X_k)]}{\sum_{l=0}^{n-1} w_{l-t}^2}, \\ C_2 &= \frac{\sum_{k=0}^{n-1}(k-t)^2 w_{k-t}^2 \mathbb{V}[T(X_k)]}{\sum_{l=0}^{n-1}(l-t)^2 w_{l-t}^2}. \end{aligned} \tag{23}$$

Therefore, the choice of r has a direct effect on the performance of the score s_n.

Let us now define our criterion for choosing the discounting rate r. In the first place, the weights are designed such that the resulting estimate $(\hat{\theta}(n), \hat{\delta}(n))$ approximates the current parameter and derivative (θ_n, δ_n) utilizing the past observations x_0^{n-1}. Hence, we are encouraged to evaluate the trade-off relationship in terms of a predictive error on an unseen sample x_n. We define the sequential predictive error,

$$\varepsilon(x_0^{n-1}) = -\sum_{k=0}^{n-1} \log p(x_k; \hat{\theta}(k) + (k - t(k))\hat{\delta}(k)), \tag{24}$$

and choose r to minimize it,

$$\hat{r}(x_0^{n-1}) \overset{\text{def}}{=} \arg\min_{r \in R} \varepsilon(x_0^{n-1}), \tag{25}$$

where R is a set of relevant values for the discounting rate.

For the independent exponential family, we predict the sufficient statistics $T(x_n)$ rather than raw data x_n itself. It is because of the aforementioned close relationship between proposed estimates and the least squares regression on $\{T(x_n)\}$. The predictive density of $T(x_n)$ according to the regression is given by the normal distribution with mean $\hat{\mu}_T(n) = \hat{\tau}(n) + (n - t(n))\hat{\xi}(n)$ and covariance $\hat{\Sigma}_T(n) = I(\hat{\theta}(n))$. Then, the cumulative predictive error is given by

$$\varepsilon(x_0^{n-1}) = -\sum_{k=0}^{n-1} \log \mathcal{N}\left[T(x_k); \hat{\mu}_T(k), \hat{\Sigma}_T(k)\right]. \tag{26}$$

Note that the above criterion can be applied not only in order to choose r but also to select the statistical model $\mathcal{M}$ itself.

6 Experiments

In this section, we show experimental results comparing the proposed method to conventional ones. First, the quantitative results on synthetic experiments are presented. Next, the qualitative results on experiments with three real-life data are exhibited.

6.1 Synthetic datasets

Now, we demonstrate the validity of the proposed method empirically using synthetic data. First, we explain how we generated the synthetic dataset. We generated three kinds of step-formed sequences, all 10,000 in length. For each of them, the underlying distribution changed continuously through nine periods, of length h starting from $n = 1000i$ $(i = 1, 2, \ldots, 9)$. Each sequence x_n $(n = 0, 1, \ldots, 9999)$ was independently drawn from the univariate Gaussian distribution with mean μ_n and variance 1, where

$$\mu_n = \textstyle\sum_{k=1}^{9}(10-k)S(n - 1000k + 1).$$

Here, $S(t)$ denotes a *slope function* with slope length h,

$$S(t) = \begin{cases} 0 & (t < 0) \\ t/h & (0 \le t < h) \\ 1 & (h \le t) \end{cases}.$$

Note that the changes occur abruptly when $h = 1$.

Next, we introduce the online change detection method employed in this experiment. For the proposed method, we

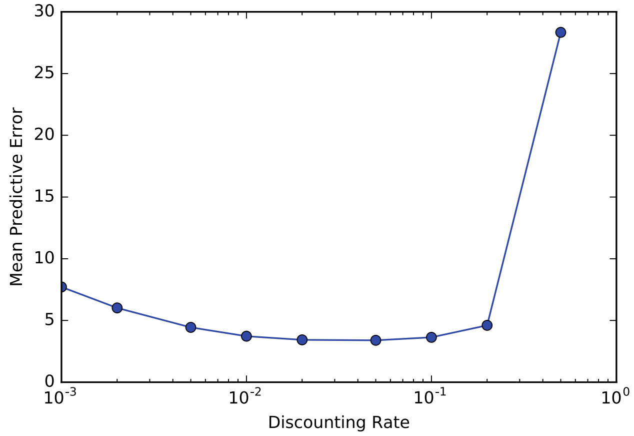


Fig. 1 Predictive errors of LLR on the training data with discounting rates $r \in \{.001, .002, .005, .01, .02, .05, .1, .2, .5\}$. The *horizontal axis* represents the discounting rate, and the *vertical axis* represents the samplewise predictive error $\varepsilon(x_0^{9999})/10{,}000$. The minimum is attained with $r = 0.05$

employed as the statistical model the univariate Gaussian distributions with unknown mean and unknown variance. We refer to this as *local linear regression* (LLR). In addition to LLR, we employed three other algorithms for comparison, both of which are designed to detect abrupt changes in an online fashion: (1) *Page–HinkleyTest* (PHT) [17], which is one of the most widely used methods of change monitoring, (2) *change finder* (CF) [18,19,21], as a the state of the art of abrupt change detection, and (3) the Bayesian method proposed by [1], which we refer to here as *Bayesian online change point detection* (BOCPD). To compare the performance of PHT with ours, we calculated the scores of change as the reciprocal of estimated run length given by PHT. Similarly, we compute the change scores of BOCPD as the posterior variance of parameters θ_t utilizing the posterior probability $P_n(l)$ of run length.

Although those four methods involve several parameters, we tuned them such that they perform their best with regard to the ROC-AUC score that we describe below. Specifically, for LLR, we have three hyper-parameters, r, γ_0, and γ_1. We chose $r = 0.05$ to minimize predictive errors over training data (see Fig. 1). We also set $\gamma_0 = \gamma_1 = 0$ (i.e., no regularization) because the univariate Gaussian model has only two parameters (μ, σ^2) and then not worth worried about its over-fitting.

In evaluating these methods, we first fixed β to a constant and converted change point scores $\{s_n\}$ into binary alarms $\{a_n\}$ thresholding with β

$$a_n \overset{\text{def}}{=} \begin{cases} 1 & (s_n \geq \beta), \\ 0 & (s_n < \beta). \end{cases}$$

Then, we evaluated the change detection algorithms in terms of the benefit and false-alarm rate defined as follows: Let T be the maximum tolerant delay of the change detection. When a change occurred at point t^*, we define the *benefit* of an alarm at time t with respect to t^*,

$$b(t; t^*) = \begin{cases} 1 & (0 \leq t - t^* < T), \\ 0 & (\text{otherwise}), \end{cases}$$

as considered in [4]. The total benefit of alarm sequence a_0^{n-1} is calculated as

$$B(\{a_n\}) \overset{\text{def}}{=} \sum_{k=0}^{9999} a_k \max_{t^* \in S} b(k; t^*).$$

Here, S denotes the set of all the change points. The number of *false alarms* is calculated as

$$N(\{a_n\}) \overset{\text{def}}{=} \sum_{k=0}^{9999} a_k \mathbb{I}(\forall t^* \in S, b(k, t^*) = 0)$$
$$= \sum_{k=0}^{9999} a_k - B(\{a_n\}),$$

where $\mathbb{I}(t)$ denotes the binary function that takes 1 if and only if proposition t is true. Finally, we visualized the performance by plotting the true-positive rate (TPR), $B/\sup_\beta B$, against the false-positive rate (FPR), $N/\sup_\beta N$, with a varying threshold parameter β. Through these performance metrics, we regard the alarms raised by T step after true changes as correct detection and the others as false detection. Specifically, for $T = 0$, we can evaluate the usefulness of change score s_n to detect changes as they are occurring. This scheme of evaluation is adopted since early detection of ongoing changes and subsequent countermeasures are important especially in the scenarios where continuous changes are expected.

The threshold β strongly affects the benefit and the number of false alarms. In order to evaluate the performances independent of the choice of β, we employed the area under the receiver operator characteristics (ROC) curve (ROC-AUC). ROC-AUC integrates TPR and FPR over all the possible values of threshold β. ROC-AUC takes one if there exists an ideal threshold, such that TPR attains its maximum keeping FPR zero. By contrast, it takes about 0.5 for random i.i.d. scores $\{s_n\}$.

The results of the experiment are summarized in Figs. 2 and 4. The top part of Fig. 2 shows an example of the generated sequences, and the bottom three parts show examples of how the three different scores varied over time. One can see in Fig. 2 that LLR had less delay than BOCPD and that LLR suppressed noise constantly, while the score of CF included considerable noise. Figure 3 shows examples of the ROC curves with tolerance $T = 50$, and that LLR and

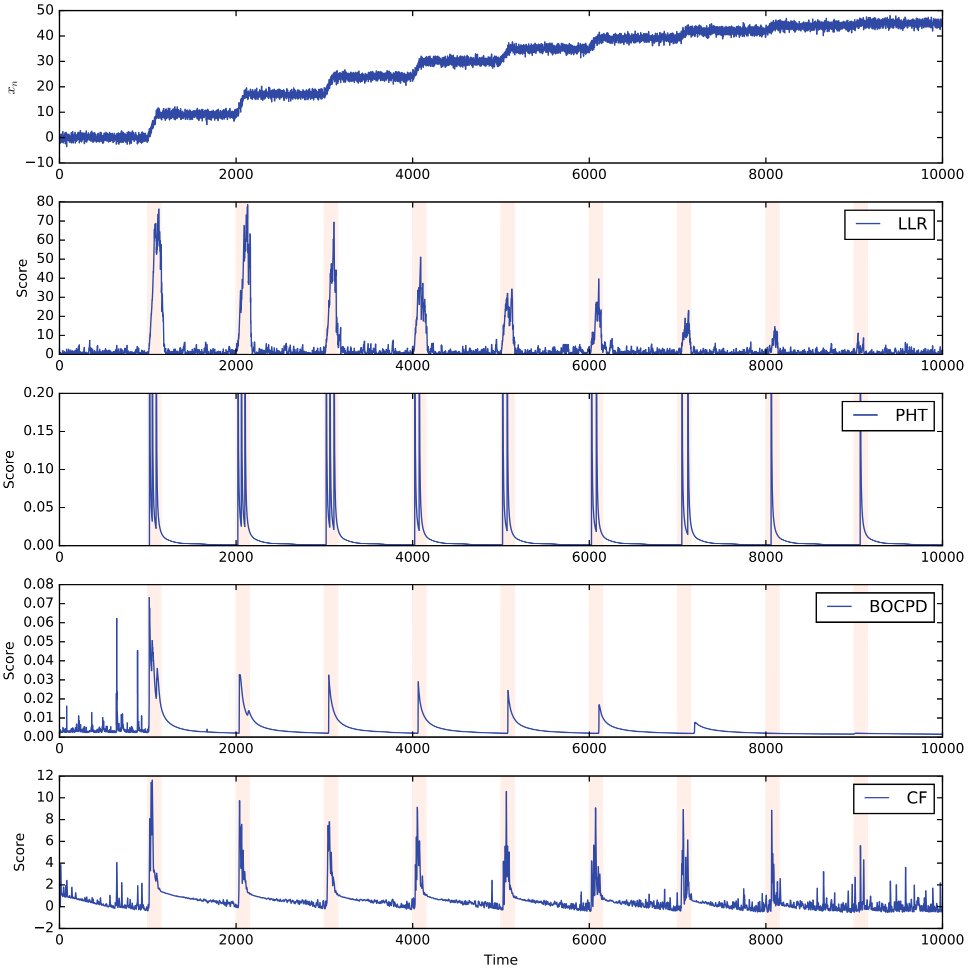


Fig. 2 Example of data and corresponding scores of change. The *horizontal axis of each plot* represents time. The *top plot* shows an example of synthetic data with $h = 100$. The *bottom three plots* show corresponding scores s_n given by different methods, LLR, PHT, BOCPD, and CF. The periods of change are indicated with *red shading* (color figure online)

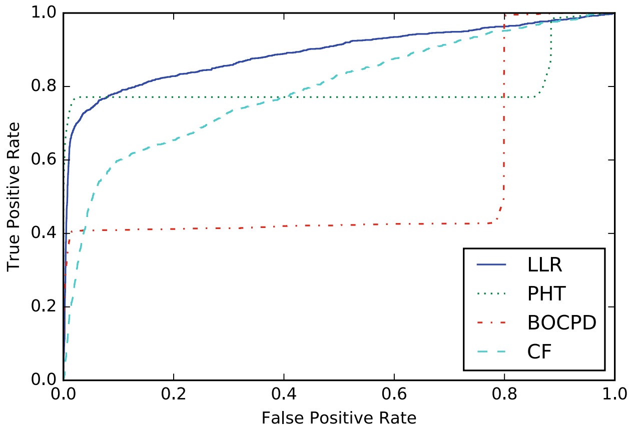
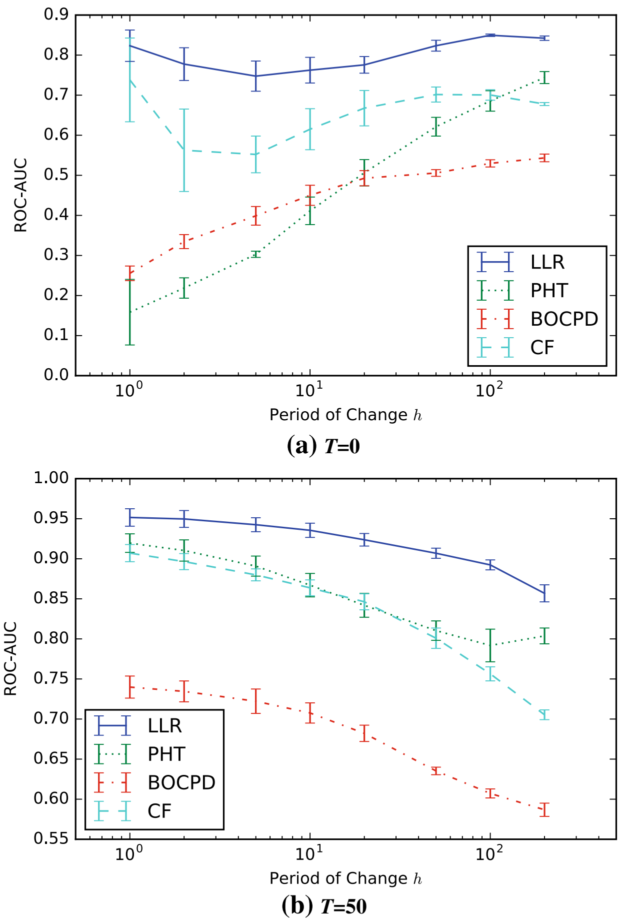


Fig. 3 Example of ROC curves with $h = 100$ and $T = 50$. The *horizontal axis* represents the false-positive rate ($FP/(FP + FN)$), and the *vertical one* represents the true-positive rate ($TP/(TP + TN)$). *Areas under the curves* indicate the performance of corresponding methods

PHT performs well in particular with the low-false-alarm setting. We evaluated the performance with respect to averages and standard deviations of ROC-AUC over five independent sequences with tolerant delay $T = 0$ and $T = 50$, respectively (Fig. 4). As seen in the figure, LLR improved ROC-AUC compared to the other methods with the all configuration of h and T. Noticing that the performance of LLR and CF is robust with respect to the change in tolerance T and that those of PHT and BOCPD drastically degrade when $T = 0$ and h is small, one can see that LLR and CF are good at early detection of changes. Specifically, it is remarkable that LLR, whose hyper-parameters are selected automatically and independently to the ROC-AUC metric outperformed the other methods, whose hyper-parameters are tuned in order to maximize ROC-AUC. Also note that LLR outperformed the others even in the discrete case ($h = 1$). This is because the former merely detects whether there is a trend in the change, whereas the latter detects individual changes and ignores their trends.

6.2 Real datasets

With three distinct real-world datasets, we qualitatively compared change detection methods, including LLR. We used (1) malware attack data, (2) economic time series data, and (3) industrial boiler data.



(a) ***T*=0**

(b) ***T*=50**

Fig. 4 Areas under the ROC curves with various $T \in \{0, 50\}$ and $h \in \{1, 2, 5, 10, 20, 50, 100, 200\}$. The *horizontal axes* represent the size of the periods of changes, h, and the *vertical ones* represent the area under the ROC curve (ROC-AUC). The *vertical error bars* show the standard deviations of ROC-AUC

6.2.1 Malware detection

First, we used eighteen days of transaction records logged on a server system when a `backdoor` was planted on it. This dataset was provided by LAC Corporation (http://laccorp.com/). It is known that some types of malware, such as a `backdoor` reveal symptoms (e.g., scanning) in the transactions before the attack actually starts. Such symptoms can be discovered by detecting the starting point of continuous changes in the transaction data.

We counted the maximum number of transactions having the identical IP address to an identical URL each second. The total length of the data was 1,551,498, and it was very sparse. We refer to this statistic as MNT and employed them as the input sequence for this experiment. Meanwhile, we counted the number of transactions in which the server returned the message `500ServerError`. We refer to this statistic as 500SE. `500ServerError` is known as a sign of an attack through `backdoors`. We applied to 500SE the Kleinberg's burst-detection algorithm [15] [henceforth, *burst detection* (BD)] with base 2 and transition cost 1 in order to detect bursts of `500ServerError` messages. The detected bursts can be thought of as the time points for the emergence of the attack. Thus, we utilized it to validate change scores. In summary, we attempted to find the appearance of attacks observing only MNT, without 500SE, and compared the resulting score of change with 500SE.

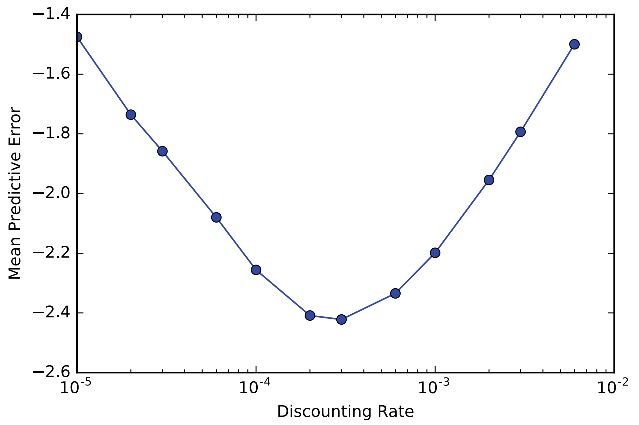


Fig. 5 Predictive errors of LLR during the first week of malware detection data. The *horizontal axis* represents the discounting rate, and the *vertical one* represents the samplewise predictive error. The twelve candidates are shown as *blue dots* in the figure. The minimum was attained with $r = 0.0003$ (color figure online)

As shown in top of Fig. 6, two groups of attacks can be distinguished in the transaction data. The starting points of those two attacks were at `13:01:45 on July 18` and at `01:18:49 on July 28`. We applied three methods: burst detection, change finder and local linear regression. Because MNT is too long for the naive algorithm of BOCPD in terms of its computational complexity, we did not employ it in this experiment. We employed the Poisson distribution for our method to model counting processes. The optimal value of the discounting rate is given as $r = 0.0003$ (see Fig. 5), where $\gamma_0 = \gamma_1 = \tau_0 = 1$. The bottom three plots in Fig. 6 show that LLR and BD detected bursts of `500ServerError`, whereas CFs score was very noisy. Note that the performance of LLR with the automatically chosen hyper-parameter r is comparable to that of BD with a manually chosen hyper-parameters.

6.2.2 Dow Jones returns

To investigate LLR's capability of detecting changes in autocorrelated sequences, we used part of the economic time series[1] that was originally used in [11] and in [1]. It is a sequence of daily *returns* of the Dow Jones Industrial Aver-

[1] http://hips.seas.harvard.edu/content/bayesian-online-changepoint-detection.

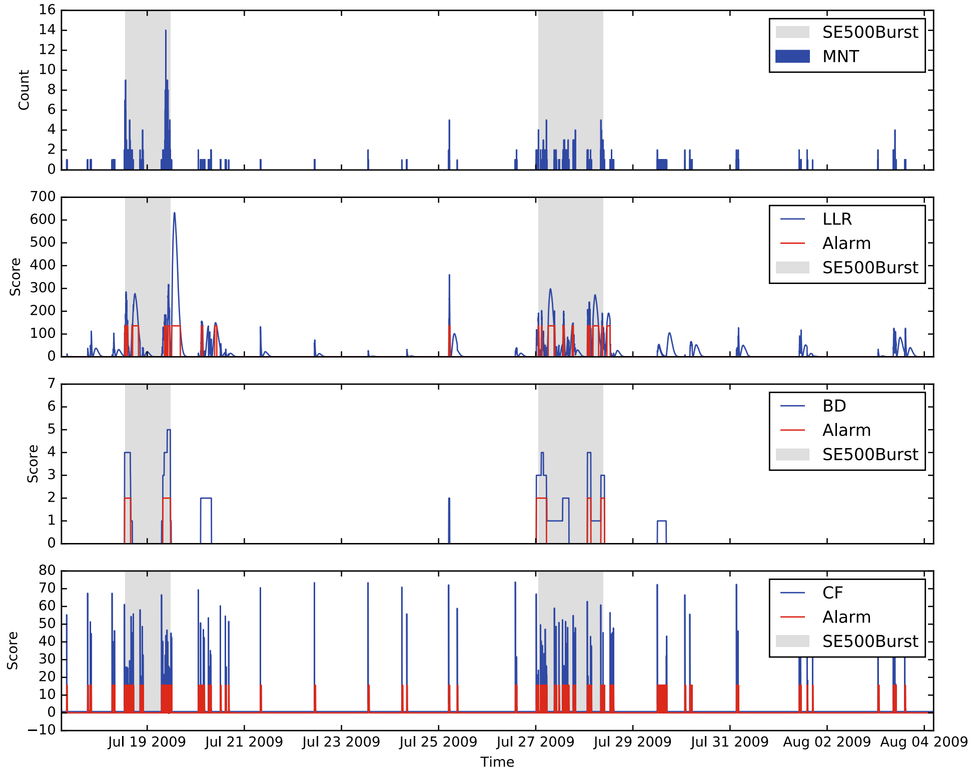


Fig. 6 Malware detection data and corresponding scores of change. The *top plot* shows the burst level of `500ServerError` message calculated with burst detection with the *bold orange line*. The *gray line* shows MNT with the *bold black line* of 500SE overlaid. The *bottom three plots* show change scores of MNT given with burst detection, change finder and our method (in this order). *Bold red lines* show the top 5 percent of the respective scores (color figure online)

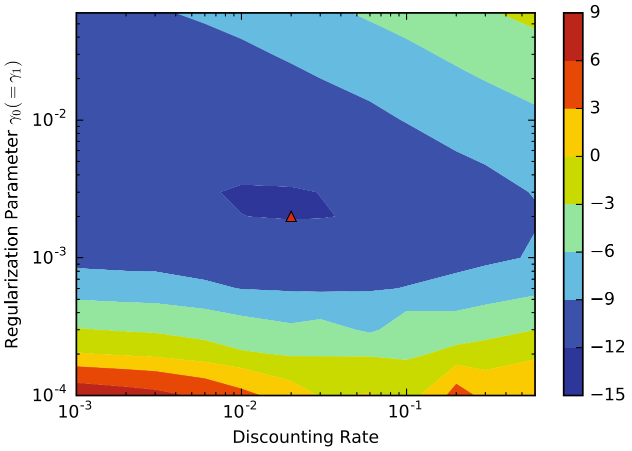


Fig. 7 Predictive errors of LLR during the first year of the market data. The *horizontal axis* represents the discounting rate, and the *vertical one* represents the regularization parameter. The minimum is *dotted with a red triangle* at $r = 0.02$ and $\gamma_0 = \gamma_1 = 0.002$ (color figure online)

age from July 5, 1972 to June 3, 1975 (top of Fig. 8). The returns are calculated as $R_n = p_n/p_{n-1} - 1$, where p_n denotes the closing price of day n. In the sequence, the variance of daily returns tended to change continuously or suddenly in association with various world events.

We applied LLR, BOCPD, and CF to this sequence and investigated how their scores were related to the three events: (1) the conviction of G. Gordon Liddy and James W. McCord, Jr. in the Watergate incident; (2) the declaration of the oil embargo by the Organization of Petroleum Exporting Countries (OPEC) against the USA; and (3) the resignation of then President Richard Nixon. Here we took $b_0 = 1$. The rest of the regularization parameters of LLR, γ_0 and γ_1, are tuned to minimize the mean predictive error as well as the discounting rate r (Fig. 7).

Figure 8 shows change scores of the respective methods versus time. The bottom three plot in Fig. 8 shows that the two conventional methods clearly captured the latter two events but gave vague or delayed scores for the first one. On the other hand, LLR raised its scores not only for the latter two, but also for the first one. This is supposed to be because a continuous change occurred around the first period.

6.2.3 *Tube failure of industrial boiler*

Finally, we examined the time series of forty sensors on an industrial boiler, as a typical multivariate time series. This data was provided by Toray Industries, Inc. The duration of the data was about three weeks, and the sampling rate was 1/30 Hz (the total length is $N = 59{,}041$). The most important fact is that a tube failure in the boiler likely due to its deterioration was logged at the very end of the data. Moreover, it oscillates during a period of eight hours by the normal operation. Thus, the data are highly non-stationary in shorter ranges, but the non-stationarity of data is not essential with respect to the incident.

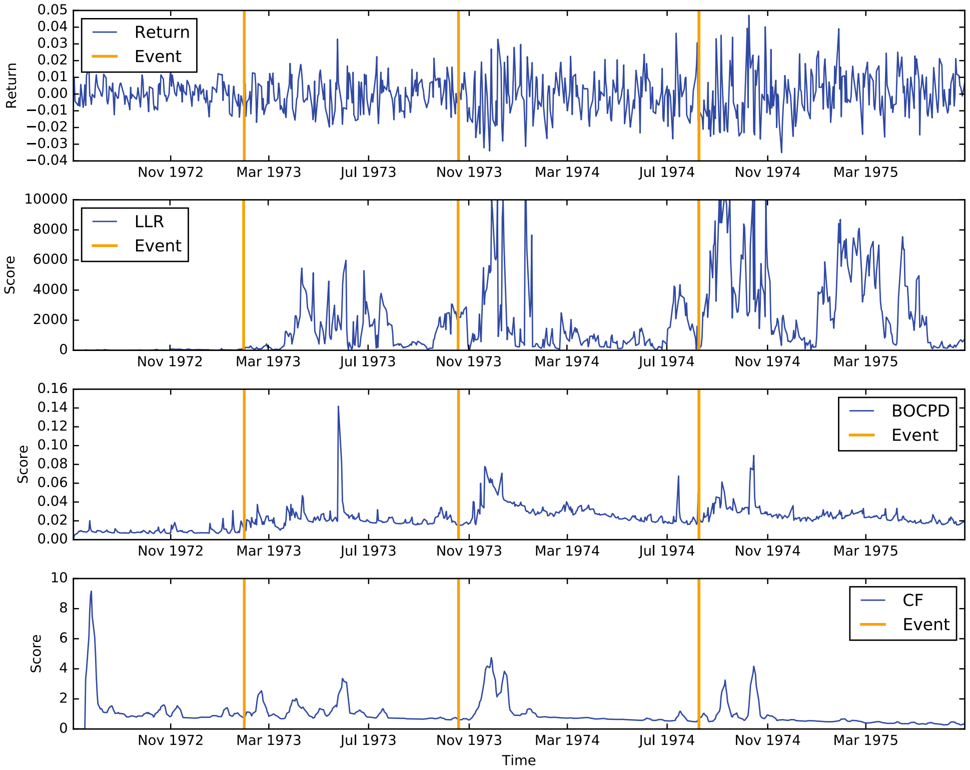


Fig. 8 The *top plot* shows the daily returns of the Dow Jones Industrial Average from July 5, 1972 to June 30, 1975. The *bottom three* show change scores given by three different methods with *blue lines*. Each plot has *three vertical orange lines* in it, which respectively show major historical events that may have affected the market index: (1) the conviction of G. Gordon Liddy and James W. McCord, Jr., former Nixon aids, on January 30th, 1973; (2) the beginning of the oil embargo by OPEC against the USA on October 19th, 1973; and (3) the resignation of then President Richar Nixon on August 9th, 1974 (color figure online)

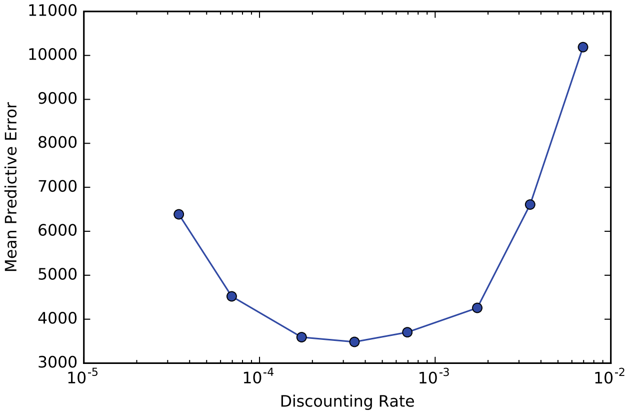


Fig. 9 Predictive error of LLR during the first week of the industrial boiler data. The *horizontal axis* represents the discounting rate, and the *vertical one* represents the mean predictive error. The minimum was attained with $r = 1/2880$

If the failure is due to deterioration, there should be symptoms in the data. Therefore, we applied the proposed method in order to determine whether one could detect them in advance. To make the data stationary, we took the difference in the data as shown in the top and middle plot in Fig. 10. Then, we employed the proposed method with a forty-variate normal distribution. In order to prevent overfitting, we set the regularization parameters as $\gamma_0 = \gamma_1 = 1$ and $\tau_0 = \tau(\mu = 0, \Sigma = I_{40})$. The optimal discounting rates was selected at $1/2880$ from nine candidates

$$\left\{\frac{.1}{2880}, \frac{.2}{2880}, \frac{.5}{2880}, \frac{1}{2880}, \frac{2}{2880}, \frac{5}{2880}, \frac{10}{2880}, \frac{20}{2880}, \frac{50}{2880}\right\},$$

such that it minimizes predictive errors during the first week (see Fig. 9). Since the corresponding time constant is $c = 2880$, which amounts to a day, the resulting score sequence is supposed to be immune to the normal oscillation, whose period is eight hours (i.e., 960 observations).

The score of the proposed method is shown in the bottom of Fig. 10. As seen in the figure, the proposed method did spot changes about one week before the tube failure. We further investigated the cause of the peak of the score by decomposing $\hat{z}(n)$,

$$\hat{z}(n) = \sum_{i=1}^{d} |v_i|^2, \tag{27}$$

where $v \overset{\text{def}}{=} I(\hat{\theta}(n))^{1/2}\hat{\delta}(n) \in \mathbb{R}^d$. Note that $d = 860$ denotes the dimensionality of the 40-variate Gaussian distribution. By looking into some largest components in the sum, we found that only three largest components out of the 860 components contribute up to 70% of the score at the peak moment as shown in Fig. 11a. Further, as v_i can be regarded as the degree of change in θ_i measured by the Fisher metric, the indices corresponding to large $|v_i|^2$ can be thought of as major causes of the detected change. We found that the three largest components are corresponding to the variance of the measured value of the 34–36th sensors as shown in Fig. 11b. In fact, there was substantial decrease in the variance of those

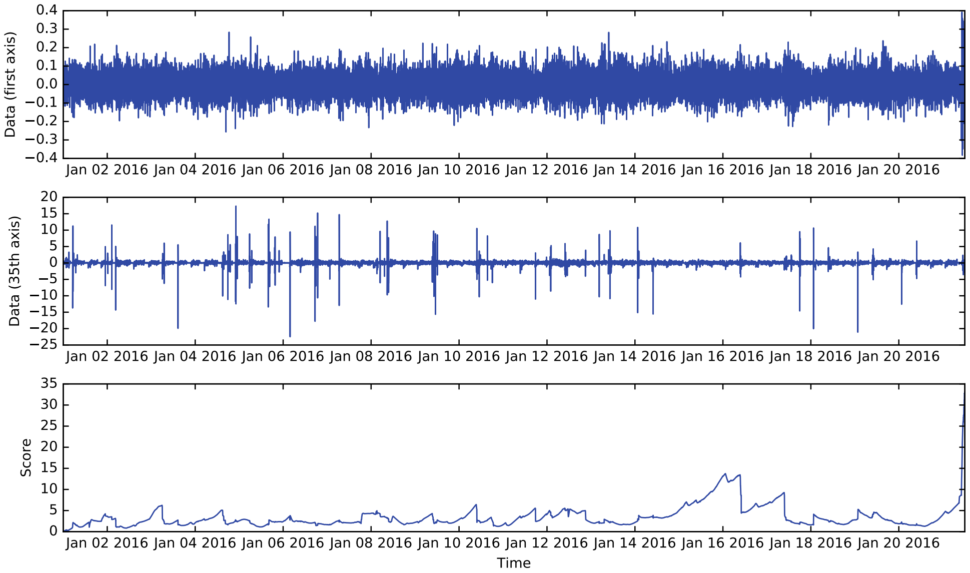


Fig. 10 Selected axes of the industrial boiler data and the corresponding scores of change. *Top and middle* the 2nd and 36th axes of the industrial boiler data. Anomalous values due to the tube failure of the boiler can be observed at the *right end of both plots*. In the *middle plot*, the variance of the sensor value decreased considerably around January 16. *Bottom* change scores computed with LLR ($r = 1/2880$). The tube failure is located by the score at the *right end of the data*. In addition, a symptom of the failure was detected around on January 16th

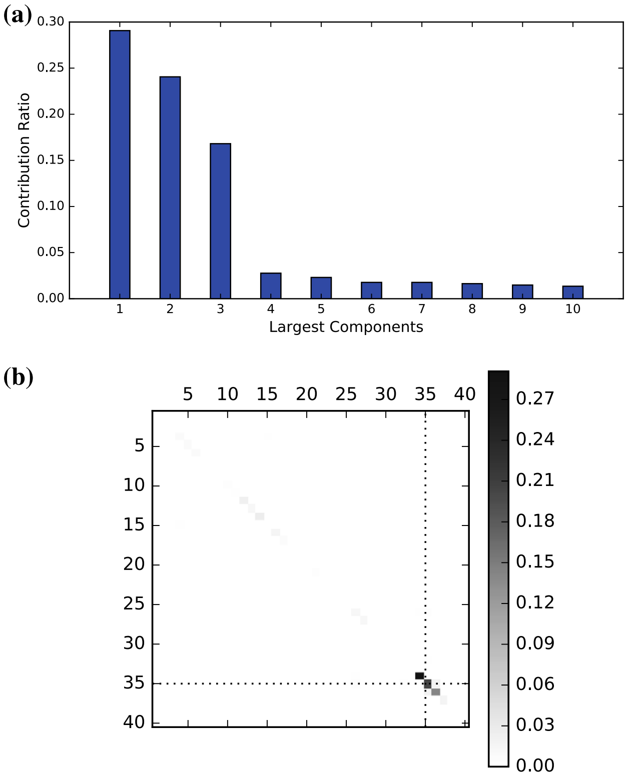


Fig. 11 Analysis on most contributing components at the peak moment of January 16, 2016. **a** The contribution ratios of the largest 10 components are shown. The *vertical axis* represents relative contribution factors $|v_i|^2/\|v\|^2$, while the *horizontal one* represents the ranks of contribution. **b** The ratios of contribution corresponding to the covariance parameter Σ at the peak is presented. The ratios are indicated with *gray scale* where the *vertical* and *horizontal axes* are corresponding to the *row* and *column* indices of the covariance matrix. There exist three remarkably large coefficients in it, which correspond to the variances of the 34–36th sensors

measurements at January 15, 2016, as shown in the middle of Fig. 10. Moreover, in response to the result of this analysis, the company confirmed that there was a foreign object in the pipe located upstream of those sensors after an inspection of the boiler and that this might be the cause of the tube failure occurred at January 21, 2016.

7 Concluding remarks

We proposed a novel model for continuously changing stochastic processes. We also described an online algorithm for estimating their characteristics, by employing a technique for localized linear regression. The estimate is invariant with respect to the parameterization and is computed with an $O(1)$-space and $O(1)$-time updating procedure. We then examined the statistical properties of the estimates and combined them into the novel algorithm for change detection. A criterion for choosing a hyper-parameter r of the algorithm was also proposed. In experiments with synthetic datasets, our method outperformed conventional methods in the trade-off between the true-positive rate and the false-positive rate on some synthetic datasets. Specifically, we demonstrated that our method is better at detecting continuous changes, and more robust even in detecting discrete changes. In experiments with real datasets, on the other hand, we saw that there likely exist continuous changes in real-life data and that our method is able to capture them well as expected.

From practical point of view, we recommend practitioners to employ the exponential family of distributions (Sect. 3.1), e.g., the (multivariate) Gaussian, Poisson, exponential, gamma, multinomial distribution. It is ensured to be computationally effective and highly flexible to model the statistical nature of data. If one is willing to explicitly model temporal dependence of data, then autoregressive models with Gaussian noise are available (Sect. 3.2).

One may wonder when the proposed approach can be applied for real-life data. Basically, it is designed to detect continuous and locally linear changes, but, as we have shown in Proposition 2, it can be used for detecting abrupt changes. Moreover, the core theoretical analysis like the parameter independence holds independent of the actual nature of data. On the other hand, asymptotic validation of the method is based on the assumption that data are distributed according to the employed statistical model and that changes can be captured by linear regression. We consider that these assumptions is not very restrictive but unfortunately difficult to verify in practice. Then, we recommend to test several statistical models on training datasets and choose the one with the smallest predictive error (given in Sect. 5).

The scalability is also of great interest in practice. For the length scalability, we have mentioned in Sect. 4 that our method runs in optimal time rate (e.g., linear with respect to the length of the data). For dimensional scalability, it depends on statistical models and individual analyses are needed. For instance, at most square time with respect to the dimensionality is required for multivariate Gaussian distributions, which is likely to be irreducible since just updating the likelihood costs square time.

The following problems remain for future study:

(1) *Further analysis of the statistical properties of the estimates.* The statistical distribution of the estimate $(\hat{\theta}(n), \hat{\delta}(n))$ plays an important role in our methodology. We have shown that it is approximated with χ^2 distribution. However, we feel the need of further analysis in cases of strongly correlated processes, specifically on the tail probability of $\hat{z}(n)$. It induces the desirable value of threshold β_α given the permissible rate of false alarms α.

(2) *Extension toward a theory of predicting changes.* The starting point for continuous changes can be thought of as a symptom of a big change. In future research, we shall extend our framework to cover other various types of symptoms of changes.

(3) *Methodology of detecting anomalous changes.* For multidimensional statistical models, changes should be localized to a specific dimension (or tuple of dimensions) of the parameter in order to understand the cause of the change. We demonstrated such a localization technique in an ad hoc manner in Sect. 6.2.3, but more comprehensive research on this topic is a future task. Moreover, it is necessary to discriminate whether such localized changes are anomalous, since some kinds of changes are sometimes not anomalous and out of interest in practical situations.

Acknowledgements This work was partially supported by JST, CREST. We express our sincere gratitude to LAC Corporation and Toray Industries, Inc. for providing datasets.

Appendix 1: Proof of Proposition 2

Since the loss function L_k is bounded and strongly convex, the estimate $(\hat{\theta}_t, \hat{\delta}_t)$ given by (7) converges. Let us now presume that $(\bar{\theta}, 0)$ is the limit value of the estimate and derive a contradiction. From asymptotic evaluation of the gradient, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\left[\frac{1}{|\Lambda|}\sum_{k\in\Lambda} w_{k-t}L_k(\theta)\right]_{\theta=\bar{\theta}} = \frac{1}{|\Lambda|}\sum_{k\in\Lambda} w_{k-t}\frac{\partial}{\partial\theta}L_k(\bar{\theta})\\
&\to c_-\mathbb{E}_{\theta_-}\left[\frac{\partial}{\partial\theta}L_0(\bar{\theta})\right] + c_+\mathbb{E}_{\theta_+}\left[\frac{\partial}{\partial\theta}L_t(\bar{\theta})\right] = 0, \qquad (28)
\end{aligned}$$

and, employing (9),

$$\begin{aligned}
&\frac{\partial}{\partial\delta}\left[\frac{1}{|\Lambda|}\sum_{k\in\Lambda} w_{k-t}L_k(\bar{\theta}+(k-t)\delta)\right]_{\delta=0}\\
&= \frac{1}{|\Lambda|}\sum_{k\in\Lambda}(k-t)w_{k-t}\frac{\partial}{\partial\theta}L_k(\bar{\theta})\\
&= \frac{1}{|\Lambda|}\sum_{k\in\Lambda_{<t}}(k-t)w_{k-t}\frac{\partial}{\partial\theta}L_k(\bar{\theta})\\
&\quad+\frac{1}{|\Lambda|}\sum_{k\in\Lambda_{\geq t}}(k-t)w_{k-t}\frac{\partial}{\partial\theta}L_k(\bar{\theta})\\
&\quad\to c_1\left\{\mathbb{E}_{\theta_+}\left[\frac{\partial}{\partial\theta}L_t(\bar{\theta})\right] - \mathbb{E}_{\theta_-}\left[\frac{\partial}{\partial\theta}L_0(\bar{\theta})\right]\right\} = 0,
\end{aligned} \qquad (29)$$

where $c_- \overset{\text{def}}{=} \int_0^t \bar{w}_{k-t}dk$, $c_+ \overset{\text{def}}{=} \int_t^n \bar{w}_{k-t}dk$ and $c_1 \overset{\text{def}}{=} \int_t^n (k-t)\bar{w}_{k-t}dk$. Note that we can ignore the regularizer since $|\Lambda|$ is large. Then, we have

$$\begin{bmatrix} c_- & c_+ \\ -c_1 & c_1 \end{bmatrix}\begin{bmatrix} \mathbb{E}_{\theta_-}\frac{\partial}{\partial\theta}L_0(\bar{\theta}) \\ \mathbb{E}_{\theta_+}\frac{\partial}{\partial\theta}L_t(\bar{\theta}) \end{bmatrix} = 0. \qquad (30)$$

Since c_-, c_+, and c_1 are all positive constants, it follows that $\mathbb{E}_{\theta_-}\frac{\partial}{\partial\theta}L_0(\bar{\theta}) = \mathbb{E}_{\theta_+}\frac{\partial}{\partial\theta}L_t(\bar{\theta}) = 0$. Therefore, we have

$\theta_- = \bar\theta = \theta_+$ since $L(\theta)$ is strictly convex. However, this is in contradiction with t being a change point. □

Appendix 2: Proof of Proposition 3

Considering another parameterization τ with a non-singular bijection $U : \theta \mapsto \tau$, we have

$$\begin{aligned}
&\frac{\partial L_k}{\partial\theta}(\hat\theta_t) + (k-t)\hat\delta_t^\top \frac{\partial^2 L_k}{\partial\theta^2}(\hat\theta_t)\\
&= \frac{\partial \tilde L_k}{\partial\tau}(\hat\tau_t)\frac{\partial\tau}{\partial\theta} + (k-t)\hat\delta_t^\top \frac{\partial\tau}{\partial\theta}^\top \frac{\partial^2 \tilde L_k}{\partial\tau^2}(\hat\tau_t)\frac{\partial\tau}{\partial\theta}\\
&= \left\{\frac{\partial \tilde L_k}{\partial\tau}(\hat\tau_t) + (k-t)\hat\xi_t^\top \frac{\partial^2 \tilde L_k}{\partial\tau^2}(\hat\tau_t)\right\}\frac{\partial\tau}{\partial\theta},
\end{aligned} \qquad (31)$$

where $\tilde L_k(\tau) \overset{\text{def}}{=} L_k(U^{-1}(\tau))$. Here $\hat\tau_t$ and $\hat\xi_t$ denote the estimates of the parameter and its time derivative with respect to the new parameterization τ,

$$(\hat\tau_t,\ \hat\xi_t) \overset{\text{def}}{=} \left(U(\hat\theta_t),\ \tfrac{\partial\tau}{\partial\theta}\hat\delta_t\right). \qquad (32)$$

Ensuring that bijection $U : \mathbb{R}^d \to \mathbb{R}^d$ is non-singular, the Jacobian $\partial\tau/\partial\theta$ is full rank. Therefore, multiplying the inverse Jacobian $\partial\theta/\partial\tau$ to both sides of (11) and (31) respectively from the right yields

$$\begin{aligned}
&\tilde L_n^{0,1}(\hat\tau_t) + \hat\xi_t^\top \tilde L_n^{1,2}(\hat\tau_t) + \tilde g_\theta(\hat\tau_t, \hat\xi_t) = 0,\\
&\tilde L_n^{1,1}(\hat\tau_t) + \hat\xi_t^\top \tilde L_n^{2,2}(\hat\tau_t) + \tilde g_\delta(\hat\tau_t, \hat\xi_t) = 0,
\end{aligned}$$

where $\tilde g(\hat\tau_t, \hat\xi_t) \overset{\text{def}}{=} g(U^{-1}(\hat\tau_t), \frac{\partial\theta}{\partial\tau}\hat\xi_t)$ denotes the regularizer on (τ, ξ) induced with mapping U. Note that bijection U is non-singular in general if the two parameterizations satisfy the condition (1). Consequently, it has been proved that the estimator $(\hat\tau_t, \hat\xi_t)$ solves the Eq. (11) with parameterization τ. It follows that the estimate of magnitude z_t with respect to alternative parameterization τ coincides with the original one,

$$\begin{aligned}
\hat z_t &= \hat\delta_t^\top I(\hat\theta_t)\hat\delta_t\\
&= \hat\delta_t^\top \frac{\partial\tau}{\partial\theta}^\top \tilde I(\hat\tau_t)\frac{\partial\tau}{\partial\theta}\hat\delta_t = \hat\xi_t^\top \tilde I(\hat\tau_t)\hat\xi_t.
\end{aligned}$$

Here $\tilde I(\tau) \overset{\text{def}}{=} \lim_{n\to\infty}\frac{1}{n}E_\tau\left[-\frac{\partial^2}{\partial\tau^2}\log p(X_0^{n-1}; U^{-1}(\tau))\right]$ is the Fisher information with respect to parameterization τ. □

References

1. Adams, R.P., MacKay, D.J.C.: Bayesian online changepoint detection (2007). arXiv:0710.3742
2. Basseville, M., Nikiforov, I.V.: Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Englewood Cliffs (1993)
3. Bifet, A., Gavaldà, R.: Learning from time-changing data with adaptive windowing. In: Proceedings of SIAM International Conference on Data Mining, pp. 443–448 (2007)
4. Fawcett, T., Provost, F.: Activity monitoring: noticing interesting changes in behavior. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 53–62 (1999)
5. Fearnhead, P., Liu, Z.: On-line inference for multiple change point problem. J. R. Stat. Soc. Ser. B **69**((Part 4)), 589–605 (2007)
6. Gama, J., Medas, P., Castillo, G., Rodrigues, P.: Learning with drift detection. In: Proceedings of SBIA Brazillian Symposium on Artificial Intelligence, pp. 285–295 (2004)
7. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A.: A survey on concept drift adaptation. ACM Computing Surveys (CSUR) **46**(4), 44 (2014)
8. Guralnik, V., Srivastava, J.: Event detection from time series data. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 33–42 (1999)
9. Gustafsson, F.: The marginalized likelihood ratio test for detecting abrupt changes. In: IEEE Transactions on Automatic Control, vol. 41, pp. 66–78. IEEE (1996)
10. Hinkley, D.V.: Inference about the change-point in a sequence of random variables. Biometrika **57**(1), 1–17 (1970)
11. Hsu, D.A.: Tests for variance shift at an unknown time point. Appl. Stat. **26**, 279–284 (1977)
12. Huang, D.T.J., Koh, Y.S., Dobbie, G., Pears, R.: Detecting volatility shift in data streams. In: IEEE International Conference on Data Mining (2014)
13. Ide, T., Kashima, H.: Eigenspace-based anomaly detection in computer system. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 440–449 (2004)
14. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Proceedings of the Thirtieth International Conference on VLDB, pp. 180–191 (2004)
15. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Discov. **7**(4), 373–397 (2003)
16. Miyaguchi, K., Yamanishi, K.: On-line detection of continuous changes in stochastic processes. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–9. IEEE (2015)
17. Page, E.: Continuous inspection schemes. Biometrika **41**(1/2), 100–115 (1954)
18. Takahashi, T., Tomioka, R., Yamanishi, K.: Discovering emerging topics in social streams via link anomaly detection. IEEE Trans. Knowl. Data Eng. **26**(1), 120–130 (2014)
19. Takeuchi, J., Yamanishi, K.: A unifying framework for detecting outliers and change points from time series. IEEE Trans. Knowl. Data Eng. **18**(18), 676–681 (2006)

20. Urabe, Y., Yamanishi, K., Tomioka, R., Iwai, H.: Real-time change-point detection using sequentially discounting normalized maximum likelihood coding. In: Advances in Knowledge Discovery and Data Mining, pp. 185–197. Springer (2011)
21. Yamanishi, K., Takeuchi, J.: A unifying framework for detecting outliers and change points from non-stationary time series data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 676–681 (2002)
22. Žliobaitė, I.: Learning under concept drift: an overview. Technical Report, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania (2009)