NON-PARAMETRIC ONLINE CHANGE-POINT DETECTION WITH KERNEL LMS BY RELATIVE DENSITY RATIO ESTIMATION

Ikram Bouchikhi, André Ferrari, Cédric Richard

Université Côte d'Azur
OCA, CNRS, Lagrange, France

Anthony Bourrier, Marc Bernot

Thales Alenia Space
Cannes la Bocca, France

ABSTRACT

Change-points can be defined as the time instants at which the underlying properties of a time series change. Detecting such points can be very challenging, especially when no prior information is available on the data distribution and the nature of the change. This paper introduces an online nonparametric kernel-based change-point detection method built upon the direct density ratio estimation of two consecutive segments of the time series. Algorithms operating in reproducing kernel Hilbert spaces have demonstrated superiority over their linear counterparts, mainly because of their ability to deal with nonlinear problems with few prior information. However their major drawback lies in the linear growth of the models order with the number of input data, which dramatically increases computational cost and memory requirement. In addition to selecting a reproducing kernel and estimating the model parameters, designing a kernel-based model requires to determine a dictionary in order to get a finite-order model. This dictionary has a significant impact on performance, and requires careful consideration. As each new data point arrives, our algorithm updates the dictionary used to approximate the density ratio based on the coherence criterion. Then it updates the parameters of the model using the kernel least mean squares algorithm. Conditions for mean stability and asymptotic unbiasedness of our method are obtained under the null hypothesis for Gaussian input data. Mean-squared error is also studied. Finally, detection performances are evaluated by computer simulations.

***Index Terms*—** Change-point detection, nonparametric detection, reproducing kernel Hilbert space, kernel least-mean-square algorithm, online algorithm.

1. INTRODUCTION

Detecting abrupt changes in the intrinsic properties of time-series can provide valuable informations in a wide range of real world applications such as fraud, network intrusion and faults in operational control systems [1, 2, 4, 18, 22]. Part of the literature on change-point detection (CPD) focuses on parametric approaches, which assume that a model describing the data distributions before and after the change is available. Most of these algorithms rely on the likelihood ratio at two consecutive intervals of the time series. This is the case for the cumulative sum type algorithms [1]. In their simplest form, these algorithms assume not only that the parameters that undergoes the change are known, but also their pre- and post-change values, e.g., change in the mean or in the variance [8]. In case where the above mentioned parameters are unknown, the generalized likelihood ratio [7], which consists of replacing all the unknown parameters by their maximum likelihood estimates, can be used. In addition to a loss of performance, these approaches do not often allow online implementations. Another group of methods based on subspace identification has been considered. Their main idea is that if, at a certain time instant, the mechanism generating the time series changes, the subspace spanned by the signal trajectory also changes as shown, e.g., in [10]. These geometric approaches implicitly assume the a linear dynamic model describe the time-series data.

In many practical situations, stochastic models that properly describe the data are not available, and the aforementioned methods are susceptible to deviations of the signal from the assumed model. Within this context, nonparametric alternatives have been devised without parametric assumption and have gained wide interest. In [11] the authors compare three nonparametric CPD algorithms: the KLIEP (Kullback-Leibler Importance Estimation Procedure), the uLSIF (unconstrained Least Squares Importance Fitting) and the RuLSIF (Relative unconstrained Least Squares Importance Fitting). These batch algorithms consist of fitting the likelihood ratio in a Reproducing Kernel Hilbert Space (RKHS), referred to as relative density ratio estimation. The RuLSIF CPD algorithm was demonstrated to be more robust than the KLIEP algorithm. An online implementation of the KLIEP-based CPD algorithm was introduced in [9] but, to our knowledge, an online version of the RuLSIF has not been derived and analyzed yet. This paper proposes an online version of a RuLSIF-based CPD algorithm. It consists of approximating the density ratio on two consecutive intervals of the time series, by a weighted sum of Gaussian kernel functions in a RKHS, and estimating the weights with the Kernel Least Mean Squares algorithm (KLMS) [12, 16]. The major drawback of this method is the linear growth of the model order with the number of input data, which dramatically increases the computational cost and memory requirement, and prevents its online implementation. The coherence sparsification rule was introduced in [16] to address this issue. It consists of discarding the kernel functions whose removal is expected to have a negligible effect on the quality of the model.

This paper is organized as follows. Section 2 formulates the problem and presents the batch CPD algorithm based on relative density ratio estimation. Section 3 describes our online CPD algorithm. Section 4 analyzes the performance of the algorithm.

2. DENSITY RATIO ESTIMATION

2.1. Problem formulation

The CPD problem addressed in this paper is formulated as follows. Let $\{y_t\}_{t\in\mathbb{N}}$ be the time series in which we aim at detecting whether a change occurred and, if affirmative, where it occurred. Let:

$$\boldsymbol{y}_t = (y_t, y_{t+1}, \ldots, y_{t+k-1})^\top \in \mathbb{R}^k \tag{1}$$

be a subsequence of $\{y_t\}_{t\in\mathbb{N}}$. Successive values of time-series are generally not independent over time. To take this dependence into

account, nonparametric algorithms generally seek a change-point in the vectors $\boldsymbol{y}_t$, called the samples in the sequel. Let $\boldsymbol{Y}_{\text{ref}} \in \mathbb{R}^{k \times n_{\text{ref}}}$ be the following Hankel matrix of n_{ref} consecutive samples representing what will be considered as the reference interval:

$$\boldsymbol{Y}_t^{\text{ref}} = (\boldsymbol{y}_{t-n_{\text{ref}}}, \boldsymbol{y}_{t-n_{\text{ref}}+1}, \ldots, \boldsymbol{y}_{t-1})$$

To address the CDP problem, the adjacent interval $\boldsymbol{Y}_t^{\text{test}} \in \mathbb{R}^{k \times n_{\text{test}}}$ is also considered:

$$\boldsymbol{Y}_t^{\text{test}} = (\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_{t+n_{\text{test}}-1})$$

referred to as the test interval. We assume that the samples $\boldsymbol{y}_i$ in the reference and test intervals are i.i.d. and distributed respectively as:

$$\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t-1} \sim p(\boldsymbol{y}) \quad \text{and} \quad \{\boldsymbol{y}_i\}_{i=t}^{t+n_{\text{test}}-1} \sim p'(\boldsymbol{y})$$

2.2. Change point detection via direct density ratio estimation

Comparing the probability distributions of time series over past and present intervals was proved to be a suitable strategy for abrupt CPD in [1]. A possible approach is to estimate each probability distribution separately from $\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t-1}$ and $\{\boldsymbol{y}_i\}_{i=t}^{t+n_{\text{test}}-1}$ by any standard density estimation method [20]. However, only their ratio:

$$r(\boldsymbol{y}) = \frac{p(\boldsymbol{y})}{p'(\boldsymbol{y})} \tag{2}$$

is required to address the CPD problem. The aim of density ratio estimation approaches is to estimate $r(\boldsymbol{y})$ using $\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t+n_{\text{test}}-1}$.

In this paper, we focus on the nonparametric strategy introduced in [11, 20], referred to as RuLSIF, that consists of estimating (2) in a RKHS from the reference and test data. Note that [11] introduces a robust estimation of $r(\boldsymbol{y})$ as it is unbounded outside the support of $p'(\boldsymbol{y})$. Basically, it consists of substituting the denominator of $r(\boldsymbol{y})$ by $\alpha p(\boldsymbol{y}) + (1-\alpha)p'(\boldsymbol{y})$ with $0 \leq \alpha \leq 1$, resulting in a ratio upper bounded by $1/\alpha$. In the sequel, without loss of generality, we shall restrict ourselves to the RuLSIF algorithm ($\alpha = 0$) in order to simplify the presentation.

2.3. Density ratio approximation

Following [23], approximating $r(\boldsymbol{y})$ by any function $g(\boldsymbol{y})$ to be defined later can be performed by minimizing the mean square loss:

$$\mathcal{C}(g) = \frac{1}{2}\mathsf{E}_{p'(\boldsymbol{y})}\{(r(\boldsymbol{y}) - g(\boldsymbol{y}))^2\} \tag{3}$$

Expanding (3) and using $r(\boldsymbol{y})p'(\boldsymbol{y}) = p(\boldsymbol{y})$ leads to:

$$\mathcal{C}(g) = \frac{1}{2}\mathsf{E}_{p'(\boldsymbol{y})}\{g^2(\boldsymbol{y})\} - \mathsf{E}_{p(\boldsymbol{y})}\{g(\boldsymbol{y})\} + \text{Const.} \tag{4}$$

Approximating the expected values of $g^2(\boldsymbol{y})$ and $g(\boldsymbol{y})$ in (4) by empirical averages over $\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t-1}$ and $\{\boldsymbol{y}_i\}_{i=t}^{t+n_{\text{test}}-1}$, respectively, leads to the following optimization problem with empirical costs:

$$\min_{g \in \mathcal{G}} \left(\frac{1}{2n_{\text{ref}}} \sum_{i=t-n_{\text{ref}}}^{t-1} g^2(\boldsymbol{y}_i) - \frac{1}{n_{\text{test}}} \sum_{i=t}^{t+n_{\text{test}}-1} g(\boldsymbol{y}_i) + \frac{\lambda}{2}\|g\|_{\mathcal{G}}^2 \right) \tag{5}$$

where $\mathcal{G}$ denotes an arbitrary Hilbert space of real-valued functions on $\mathbb{R}$, and $\frac{\lambda}{2}\|g\|_{\mathcal{G}}^2$ a regularization term with $\lambda \geq 0$ to promote smoothness of g. Let $\mathcal{G}$ be a reproducing kernel Hilbert space, and let $\kappa(\cdot\,,\cdot)$ be the reproducing kernel of $\mathcal{G}$. By virtue of the Representer Theorem in [17], the function $g(\cdot)$ of $\mathcal{G}$ minimizing (5) can be expressed as a kernel expansion in terms of available data, namely,

$$g(\cdot, \boldsymbol{\theta}) = \sum_{i=t-n_{\text{ref}}}^{t+n_{\text{test}}-1} \theta_i \kappa(\cdot\,, \boldsymbol{y}_i) \tag{6}$$

where the θ_i are parameters to estimate from data $\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t+n_{\text{test}}-1}$. A common strategy to reduce the computational cost consists of using a reduced fixed dictionary $\boldsymbol{\omega} = \{\boldsymbol{y}_{\omega_i}\}_{i=1}^{L}$ designed from available training data. This leads to:

$$g(\cdot, \boldsymbol{\theta}) = \sum_{i=1}^{L} \theta_i \kappa(\cdot\,, \boldsymbol{y}_{\omega_i}) \tag{7}$$

where L is the cardinality of the dictionary and the model order. Designing the dictionary $\boldsymbol{\omega}$ is a critical point. As an example, in [21] the authors propose to choose randomly the $\boldsymbol{y}_{\omega_i}$ among the samples of the test interval. In order to capitalize on all the information available since the initial time instant $t = 0$, we recommend to design the dictionary based on L data in $\{\boldsymbol{y}_i\}_{i=0}^{t+n_{\text{test}}-1}$. An extensive literature addressing this issue in batch and online modes exists, see, e.g. [19] and references therein. The coherence-based sparsification rule introduced in the sequel has demonstrated its efficiency and is widely used for online selection of kernel-based dictionaries.

Substituting (6) into (5), and minimizing (5) w.r.t. $\boldsymbol{\theta}$, we find that $\hat{\boldsymbol{\theta}}_t$ is the solution of the strictly convex optimization problem:

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_t = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^L} J_t(\theta) \\ &\text{with } \; J_t(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{H}_t \boldsymbol{\theta} - \boldsymbol{h}_t^\top \boldsymbol{\theta} + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \end{aligned} \tag{8}$$

where $\boldsymbol{h}_t$ is the $L \times 1$ vector, and $\boldsymbol{H}_t$ is the $L \times L$ matrix, given by:

$$\boldsymbol{h}_t = \frac{1}{n_{\text{test}}} \sum_{i=t}^{t+n_{\text{test}}-1} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}}) \tag{9}$$

$$\boldsymbol{H}_t = \frac{1}{n_{\text{ref}}} \sum_{i=t-n_{\text{ref}}}^{t-1} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}}) \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}})^\top \tag{10}$$

with $\boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}}) = \left[\kappa(\boldsymbol{y}_i, \boldsymbol{y}_{\omega_1}), \ldots, \kappa(\boldsymbol{y}_i, \boldsymbol{y}_{\omega_L})\right]^\top$.

2.4. The test statistic

Let $g(\boldsymbol{y}; \hat{\boldsymbol{\theta}}_t)$ be the density-ratio estimator at time t, with $\hat{\boldsymbol{\theta}}_t$ the solution of the optimization problem (8). In [11] the authors suggest to detect the change point using the divergence between the two densities $p(\boldsymbol{y})$ and $p'(\boldsymbol{y})$. The Kullback-Leibler and the Pearson divergences can be expressed as the expectation of functions of $g(\boldsymbol{y}; \boldsymbol{\theta})$ w.r.t. $p(\boldsymbol{y})$ or $p'(\boldsymbol{y})$. A test statistic can then be obtained by approximating, as in Sec. 2.3, these expectations by empirical averages over data $\{\boldsymbol{y}_i\}_{i=t-n_{\text{ref}}}^{t-1}$ and $\{\boldsymbol{y}_i\}_{i=t}^{t+n_{\text{test}}-1}$.

In this paper, we propose an alternative test statistic consisting of the approximate log-likelihood ratio for the samples of the test interval, namely,

$$d_t = -\sum_{i=t}^{t+n_{\text{test}}-1} \ln \frac{p(\boldsymbol{y}_i)}{p'(\boldsymbol{y}_i)} \approx -\sum_{i=t}^{t+n_{\text{test}}-1} \ln g(\boldsymbol{y}_i; \hat{\boldsymbol{\theta}}_t) \tag{11}$$

3. ONLINE CHANGE POINT DETECTION

3.1. Update strategy

Let $\boldsymbol{\theta}_t$ be an estimate of the parameters of the density ratio model obtained at instant t. When $t \rightarrow t+1$, [11] updates $\boldsymbol{\theta}_t$ by solving the new optimization problem (8) at $t+1$. We propose in this paper to use the KLMS algorithm to update the estimation of $\boldsymbol{\theta}_t$. The convergence behavior of the KLMS algorithm was studied in [3], and an extended analysis of the stochastic behavior of the KLMS algorithm with Gaussian kernel was proposed in [15]. It is important to note that, at each instant t, updating $g(\boldsymbol{y}, \boldsymbol{\theta}_t)$ is a two-step process that consists of updating the dictionary $\boldsymbol{\omega}$ (and the model order L) and the parameter vector $\boldsymbol{\theta}_t$. We will adopt the following two-step strategy inspired by [6].

Dictionnary update

Various strategies have been introduced in the online kernel filtering literature to update the dictionary $\boldsymbol{\omega}$. For example, Approximate Linear Dependency (ALD) [5] checks whether, in the feature space $\mathcal{G}$, the new candidate $\kappa(\cdot\,, \boldsymbol{y}_{t+1})$ can be well approximated by a linear combination of the current dictionary elements $\kappa(\cdot\,, \boldsymbol{y}_{\omega_i})$. If not, it is added to the dictionary. The coherence rule was introduced to avoid the computational complexity inherent to ALD. It is now widely used and considered as a state-of-the-art strategy. Coherence, defined by [16]:

$$\eta = \max_{i \neq j} |\kappa(\boldsymbol{y}_{\omega_i}, \boldsymbol{y}_{\omega_j})|, \tag{12}$$

reflects the largest correlation in the dictionary between dictionary elements. The coherence rule [16] for kernel-based dictionary selection consists of inserting $\boldsymbol{y}_{t+1}$ in the current dictionary $\boldsymbol{\omega}$ provided that its coherence remains below a given threshold η_0 in $[0, 1[$:

$$\max_{\boldsymbol{y}_{\omega_i} \in \boldsymbol{\omega}} |\kappa(\boldsymbol{y}_{t+1}, \boldsymbol{y}_{\omega_i})| \leq \eta_0 \tag{13}$$

It was proven in [16], that the dimension of dictionaries determined with rule (13) is finite due to the compactness of the input space.

Parameters update

Depending if the new sample $\boldsymbol{y}_{t+1}$ has been inserted into the dictionary, the parameter vector $\boldsymbol{\theta}_t$ is updated according to [16]:

- If $\max_{\boldsymbol{y}_{\omega_i} \in \boldsymbol{\omega}} |\kappa(\boldsymbol{y}_{t+1}, \boldsymbol{y}_{\omega_i})| > \eta_0$, the dictionary is unchanged and $\boldsymbol{\theta}_t$ is updated using a single step of the KLMS algorithm:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \mu \hat{\nabla} J_{t+1}(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \mu \big[(\boldsymbol{H}_{t+1} + \lambda \boldsymbol{I}) \boldsymbol{\theta}_t - \boldsymbol{h}_{t+1} \big] \end{aligned} \tag{14}$$

  where $\hat{\nabla} J_{t+1}(\boldsymbol{\theta}_t)$ denotes an instantaneous estimate of the gradient of $J_{t+1}(\cdot)$ evaluated at $\boldsymbol{\theta}_t$.

- If $\max_{\boldsymbol{y}_{\omega_i} \in \boldsymbol{\omega}} |\kappa(\boldsymbol{y}_{t+1}, \boldsymbol{y}_{\omega_i})| \leq \eta_0$, $\boldsymbol{y}_{t+1}$ is added to the dictionary (and $L \leftarrow L+1$), and $\boldsymbol{\theta}_t$ is updated according to:

$$\boldsymbol{\theta}_{t+1} = \begin{pmatrix} \boldsymbol{\theta}_t \\ 0 \end{pmatrix} - \mu \big[(\boldsymbol{H}_{t+1} + \lambda \boldsymbol{I}) \begin{pmatrix} \boldsymbol{\theta}_t \\ 0 \end{pmatrix} - \boldsymbol{h}_{t+1} \big] \tag{15}$$

The pseudo-code of the corresponding algorithm is given in Alg.1. The recursive computation of $\boldsymbol{h}_{t+1}$ and $\boldsymbol{H}_{t+1}$ from $\boldsymbol{h}_t$ and $\boldsymbol{H}_t$ whether the new sample $\boldsymbol{y}_{t+1}$ is inserted in the dictionary or not is not detailed due to lack of space.

Algorithm 1: Pseudo code for online update of the dictionary and the parameter vector $\boldsymbol{\theta}_t$.

Require: $\boldsymbol{y}_{t+n_{\text{test}}}, \boldsymbol{\theta}_t$
  if $\max_{\boldsymbol{y}_{\omega_i} \in \boldsymbol{\omega}} |\kappa(\boldsymbol{y}_{t+1}, \boldsymbol{y}_{\omega_i})| > \eta_0$ **then**
    $\boldsymbol{h}_{t+1} = \frac{1}{n_{\text{test}}} \sum_{i=t+1}^{t+n_{\text{test}}} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}})$
    $\boldsymbol{H}_{t+1} = \frac{1}{n_{\text{ref}}} \sum_{i=t-n_{\text{ref}}+1}^{t} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}}) \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}})^\top$
    $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mu \big[(\boldsymbol{H}_{t+1} + \lambda \boldsymbol{I}) \boldsymbol{\theta}_t - \boldsymbol{h}_{t+1} \big]$
  else
    $\boldsymbol{\omega} = \{\boldsymbol{\omega}, \boldsymbol{y}_{t+1}\}$
    $\boldsymbol{h}_{t+1} = \frac{1}{n_{\text{test}}} \sum_{i=t+1}^{t+n_{\text{test}}} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}})$
    $\boldsymbol{H}_{t+1} = \frac{1}{n_{\text{ref}}} \sum_{i=t-n_{\text{ref}}+1}^{t} \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}}) \boldsymbol{\kappa}(\boldsymbol{y}_i, \boldsymbol{y}_{\boldsymbol{\omega}})^\top$
    $\boldsymbol{\theta}_{t+1} = (\boldsymbol{\theta}_t^\top, 0)^\top - \mu \big[(\boldsymbol{H}_{t+1} + \lambda \boldsymbol{I}) (\boldsymbol{\theta}_t^\top, 0)^\top - \boldsymbol{h}_{t+1} \big]$
  end if
  $d_{t+1} = \sum_{i=t+1}^{t+n_{\text{test}}} \ln \Big(\sum_{\boldsymbol{y}_{\omega_i} \in \boldsymbol{\omega}} [\hat{\theta}_{t+1}]_i \kappa(., \boldsymbol{y}_{\omega_i}) \Big)$
  Test: $d_{t+1} \underset{H_1}{\overset{H_0}{\lessgtr}} \xi$

4. PERFORMANCE ANALYSIS

4.1. Convergence of the parameters

The performance of the algorithm depends on the model order, the kernel used, and the properties of the signal operating environment. In this section, we analyse the convergence of $\boldsymbol{\theta}_t$ under the null hypothesis, i.e., no change-point is present, and for a predefined dictionary $\boldsymbol{\omega}$. We assume that, under the null hypothesis, $\boldsymbol{y}$ is Gaussian distributed, namely, $\boldsymbol{y} \sim \mathcal{N}_k(\boldsymbol{m}, \boldsymbol{R})$. The analysis below is conducted with the Gaussian reproducing kernel defined as follows:

$$\kappa(\boldsymbol{y}, \boldsymbol{y}') = e^{-\frac{\|\boldsymbol{y}-\boldsymbol{y}'\|^2}{2\sigma^2}} \tag{16}$$

where σ denotes the kernel bandwidth.

The first step consists of evaluating the optimal ("true") parameter vector $\boldsymbol{\theta}^*$. Replacing $r(\boldsymbol{y})$ in (3) by 1, and $g(\boldsymbol{y})$ by (7), we obtain the minimizer $\boldsymbol{\theta}^*$ of $J_t(\theta)$ by substituting $\boldsymbol{H}_t$ and $\boldsymbol{h}_t$, respectively, by $\boldsymbol{H}$ and $\boldsymbol{h}$, defined as:

$$\begin{aligned} \boldsymbol{h} &= \mathsf{E}_{p'(\boldsymbol{y})} \{ \boldsymbol{\kappa}(\boldsymbol{y}, \boldsymbol{y}_{\boldsymbol{\omega}}) \} \\ \boldsymbol{H} &= \mathsf{E}_{p'(\boldsymbol{y})} \{ \boldsymbol{\kappa}(\boldsymbol{y}, \boldsymbol{y}_{\boldsymbol{\omega}}) \boldsymbol{\kappa}(\boldsymbol{y}, \boldsymbol{y}_{\boldsymbol{\omega}})^\top \} \end{aligned}$$

Considering the kernel κ in (16), it can be shown that the entries of $\boldsymbol{h}$ and $\boldsymbol{H}$ are given by:

$$[\boldsymbol{h}]_\ell = e^{-\frac{\|\boldsymbol{y}_{\omega_\ell}\|^2}{2\sigma^2}} \; \mathsf{E}_{p'} \left\{ e^{-\frac{\|\boldsymbol{y}\|^2 - 2\boldsymbol{y}_{\omega_\ell}^\top \boldsymbol{y}}{2\sigma^2}} \right\} \tag{17}$$

$$[\boldsymbol{H}]_{\ell,q} = e^{-\frac{\|\boldsymbol{y}_{\omega_\ell}\|^2 + \|\boldsymbol{y}_{\omega_q}\|^2}{2\sigma^2}} \; \mathsf{E}_{p'} \left\{ e^{-\frac{\|\boldsymbol{y}\|^2 - (\boldsymbol{y}_{\omega_\ell} + \boldsymbol{y}_{\omega_q})^\top \boldsymbol{y}}{\sigma^2}} \right\} \tag{18}$$

with $\ell,\, q \in \{1, \ldots, L\}$. By considering that $\boldsymbol{y}$ is Gaussian distributed, closed-form expressions can be obtained for both expectations in (17), (18) by using the moment generating function of a quadratic form of a Gaussian vector. See, e.g., [14]. We denote the error at time instant t by $\boldsymbol{e}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}^*$. With (14) we have:

$$\begin{aligned} \boldsymbol{e}_{t+1} = & \big[\boldsymbol{I} - \mu (\boldsymbol{H}_{t+1} + \lambda \boldsymbol{I}) \big] \boldsymbol{e}_t \\ & + \mu \big[\boldsymbol{h}_{t+1} - \boldsymbol{h} - (\boldsymbol{H}_{t+1} - \boldsymbol{H}) \boldsymbol{\theta}^* \big] \end{aligned} \tag{19}$$

Assuming that $\boldsymbol{\kappa}(\boldsymbol{y}, \boldsymbol{y}_{\boldsymbol{\omega}})\, \boldsymbol{\kappa}(\boldsymbol{y}, \boldsymbol{y}_{\boldsymbol{\omega}})^\top$ and $\boldsymbol{e}_t$ are independent, which is a usual hypothesis in the analysis of adaptive filters [13],

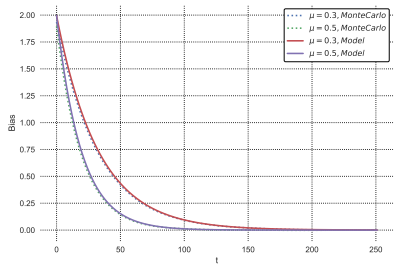

Fig. 1. Bias of $[\theta_t]_1$ as a function of t ($n_{\text{ref}} = n_{\text{test}} = 22$).

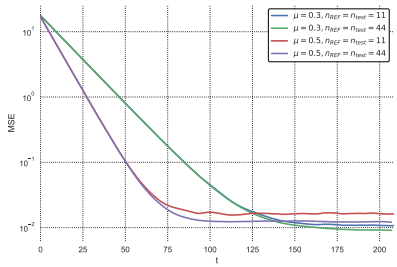

Fig. 2. Mean-squared error of θ_t as a function of t.

taking the expectation of (19), and using that $\boldsymbol{H}_t$ and $\boldsymbol{h}_t$ are unbiased estimates of $\boldsymbol{h}$ and $\boldsymbol{H}$, yields the following recursion for the bias $\mathrm{b}(\boldsymbol{\theta}_t) = \mathsf{E}\{\boldsymbol{e}_t\}$:

$$\mathrm{b}(\boldsymbol{\theta}_{t+1}) = \big[\boldsymbol{I} - \mu(\boldsymbol{H} + \lambda\boldsymbol{I})\big]\mathrm{b}(\boldsymbol{\theta}_t) \tag{20}$$

This proves that under the null hypothesis, for a given dictionary $\boldsymbol{\omega}$ and under the Gaussian assumption, the online Alg. 1 is mean stable and asymptotically unbiased if the matrix $\boldsymbol{I} - \mu(\boldsymbol{H} + \lambda\boldsymbol{I})$ is stable, that is,

$$\mu < \frac{2}{\zeta_{\max}\{\boldsymbol{H} + \lambda\boldsymbol{I}\}}. \tag{21}$$

where $\zeta_{\max}\{\cdot\}$ is the maximal eigenvalue of its matrix argument.

To test the model (20) accuracy, we assumed that y_t is a second-order autoregressive process ($a_1 = -2\cos(0.44\pi), a_2 = 0.95^2$) with Gaussian input. We set the algorithm parameters as follows: $k = 5$, $n_{\text{ref}} = n_{\text{test}} = 22$, $L = 5$, and $\sigma = 1$. The 5 elements of the dictionary $\boldsymbol{\omega}$ were selected among the $\boldsymbol{y}_i$ without considering the coherence rule in this experiment. We ran algorithm (14) for two different values of μ. The results were averaged over 200 Monte-Carlo runs. Figure 1 illustrates the behavior of the first entry of $\mathrm{b}(\boldsymbol{\theta}_t)$. We observe that the theoretical curves provided by (20) match well the actual performance provided by Monte-Carlo simulations. Figure 2 shows the behavior of the mean-squared error $\mathsf{E}\{\|\theta_t - \theta^*\|^2\}$ as a function of t. The theoretical analysis of the MSE is beyond the scope of this communication due to lack of space. The results in Figure 2 are presented for different values of μ and n_{test}. We observe the classical behavior w.r.t μ of stochastic gradient descent algorithms: a large value for μ accelerates the convergence but increases the steady-state error, and vice versa. It is also interesting to note the effect of n_{ref} and n_{test}. It can be noticed from (9)–(10), that under mild assumptions $(\boldsymbol{H}_t, \boldsymbol{h}_t)$ converge to $(\boldsymbol{H}, \boldsymbol{h})$ for $n_{\text{ref}} \to \infty$ and $n_{\text{test}} \to \infty$. Hence, a larger value for n_{test} reduces the steady state error.

4.2. Detection performance

This section experimentally evaluates the detection performance of Alg. 1. We considered a noisy sinusoid y_t of 300 samples with an abrupt change of its frequency at $t = 150$. The SNR was set to 3dB. The parameters of the algorithm were set to: $k = 5$, $n_{\text{ref}} = n_{\text{test}} = 22$, $\sigma = 0.7$, $\mu = 0.5$ and $\lambda = 0.01$. The threshold η_0 of the coherence rule used to update the dictionary was set to $\eta_0 = 0.01$. Figure 3 illustrates a realization of the signal y_t and the associated test statistic d_t. This figure shows the ability of the proposed algorithm to detect the abrupt change in this case. Figure 4 provides the ROC curves associated to the test statistic d_t obtained using both the Rulsif algorithm and our online algorithm, computed for 5.10^4 runs of y_t with different random seeds. The probabilities of detection and false alarm were estimated by thresholding d_t on the whole interval, e.g., $P_{\text{detec}} = P(\exists t \in (0, 300),\ d_t > \xi)$. This curve can be used to calibrate the threshold ξ for the online detection. We should mention that, besides the difference between the two methods in updating the model parameters $\boldsymbol{\theta}_i$, our methods uses a reduced, well chosen dictionary, whereas the Rulsif with a sliding window uses the entire set of test samples as dictionary elements. This may explain why our online method performed better than the batch Rulsif.

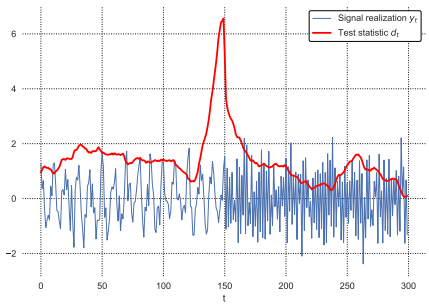

Fig. 3. Signal realization and associated test statistic computed for $n_{ref} = n_{test} = 22$ and $k = 5$.

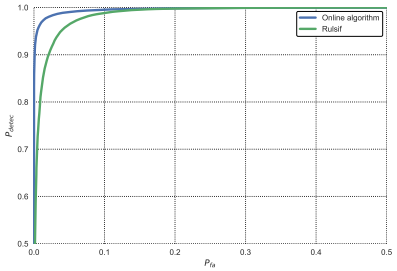

Fig. 4. ROC curves computed for the test statistics associated to 5.10^4 realizations of y_t using $n_{ref} = n_{test} = 22$ and $k = 5$.

5. CONCLUSION

Our contribution in this paper was to extend the existing RuLSIF change-point detection method to an online method where, in order to optimize the computational cost and the performance, we used the coherence-based sparsification rule to update the dictionary, and the kernel least mean square algorithm to update the weights of the density ratio model as new data arrives. The proposed method showed promising results. It was tested with different types of changes —change in the mean, in the frequency, in the amplitude. Conditions of stability and unbiasedness under the null hypothesis were also studied.

6. REFERENCES

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

[2] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Institute of Mathematical Statistics*, 17(3):235 – 249, 2002.

[3] J. Chen, W. Gao, C. Richard, and J.-C. M.Bermudez. Convergence analysis of kernel LMS algorithm with pre-tuned dictionary. *Proc. IEEE ICASSP'14, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[4] M. Csrgo and L. Horvth. *Limit Theorems in Change-Point Analysis*. John Wiley and sons, 1997.

[5] Y. Engel, S. Mannor, and R. Meir. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*, 52(8):2275 – 2285, 2004.

[6] W. Gao, J. Chen, C. Richard, and J. Huang. Online dictionary learning for kernel LMS. *IEEE Transactions on Signal Processing*, 62(11):2765 – 2777, 2014.

[7] F. Gustafsson. The marginalized likelihood ratio test for detecting abrupt changes. *IEEE Transactions on Automatic Control*, pages 66 – 78, 1996.

[8] C. Inclan and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913 – 923, 1994.

[9] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114 – 127, 2012.

[10] Y. Kawahara, T. Yairi, and K. Machida. Change-point detection in time-series data based on subspace identification. In *Proc. IEEE ICDM'07, IEEE International Conference on Data Mining*, March 2007.

[11] S. Liu, M. Yamada, N. Collier, and M. Sugiyamaa. Change-point detection in time series data by relative density-ratio estimation. *Neural Networks*, pages 72 – 83, 2013.

[12] W. Liu, P. P. Pokharel, and J. C. Principe. The kernel least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 56(2):543 – 554, 2008.

[13] J. Minkoff. Comment on the "Unnecessary assumption of statistical independence between reference signal and filter weights in feedforward adaptive systems". *IEEE Transactions on Signal Processing*, 49(5):1109, 2001.

[14] J. Omura and T. Kailath. Some useful probability distributions. Technical Report 7050 - 6, Stanford Electronics Laboratories, Stanford University, 1965.

[15] W. D. Parreira, J.-C. M.Bermudez, C. Richard, and J.-Y. Tourneret. Stochastic behavior analysis of the gaussian kernel least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 60(5):2208–2222, 2012.

[16] C. Richard, J. C. M. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *IEEE Trans. Signal Process.*, 57(3):1058 – 1067, 2009.

[17] B. Schölkopf, R. Herbrich, and R. Williamson. A generalized representer theorem. Technical Report NC2-TR-2000-81, NeuroCOLT, Royal Holloway College, University of London, UK, 2000.

[18] N. Sheng and H. Wang. Fault detection and diagnosis for operational control systems. *Automation and Computing (ICAC)*, 2015.

[19] K. Slavakis, P. Bouboulis, and S. Theodoridis. Online learning in reproducing kernel Hilbert spaces. *E-Reference, Signal Processing, Elsevier*, 2013.

[20] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[21] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bnau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699 – 746, 2008.

[22] A. Tartakovsky, B. Rozovskii, R. Blazek, and H. Kim. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372 – 3382, 2006.

[23] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Proc. NIPS'11, Neural Information Processing Systems*, 2011.