Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation

Song Liu
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
song@sg.cs.titech.ac.jp

Makoto Yamada
NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Kyoto 619-0237, Japan.
yamada.makoto@lab.ntt.co.jp

Nigel Collier
National Institute of Informatics (NII)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.
European Bioinformatics Institute (EBI)
Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
collier@nii.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/˜sugi

Abstract

The objective of change-point detection is to discover abrupt property changes lying behind time-series data. In this paper, we present a novel statistical change-point detection algorithm based on non-parametric divergence estimation between time-series samples from two retrospective segments. Our method uses the relative Pearson divergence as a divergence measure, and it is accurately and efficiently estimated by a method of direct density-ratio estimation. Through experiments on artificial and real-world datasets including human-activity sensing, speech, and Twitter messages, we demonstrate the usefulness of the proposed method.

Keywords

change-point detection, distribution comparison, relative density-ratio estimation, kernel methods, time-series data

1 Introduction

Detecting abrupt changes in time-series data, called *change-point detection*, has attracted researchers in the statistics and data mining communities for decades (Basseville and Nikiforov, 1993; Gustafsson, 2000; Brodsky and Darkhovsky, 1993). Depending on the delay of detection, change-point detection methods can be classified into two categories: *Real-time detection* (Adams and MacKay, 2007; Garnett et al., 2009; Paquet, 2007) and *retrospective detection* (Basseville and Nikiforov, 1993; Takeuchi and Yamanishi, 2006; Moskvina and Zhigljavsky, 2003a).

Real-time change-point detection targets applications that require immediate responses such as robot control. On the other hand, although retrospective change-point detection requires longer reaction periods, it tends to give more robust and accurate detection. Retrospective change-point detection accommodates various applications that allow certain delays, for example, climate change detection (Reeves et al., 2007), genetic time-series analysis (Wang et al., 2011), signal segmentation (Basseville and Nikiforov, 1993), and intrusion detection in computer networks (Yamanishi et al., 2000). In this paper, we focus on the retrospective change-point detection scenario and propose a novel non-parametric method.

Having been studied for decades, some pioneer works demonstrated good change-point detection performance by comparing the probability distributions of time-series samples over past and present intervals (Basseville and Nikiforov, 1993). As both the intervals move forward, a typical strategy is to issue an alarm for a change point when the two distributions are becoming significantly different. Various change-point detection methods follow this strategy, for example, the *cumulative sum* (Basseville and Nikiforov, 1993), the *generalized likelihood-ratio method* (Gustafsson, 1996), and the *change finder* (Takeuchi and Yamanishi, 2006). Such a strategy has also been employed in novelty detection (Guralnik and Srivastava, 1999) and outlier detection (Hido et al., 2011).

Another group of methods that have attracted high popularity in recent years is the *subspace* methods (Moskvina and Zhigljavsky, 2003a,b; Ide and Tsuda, 2007; Kawahara et al., 2007). By using a pre-designed time-series model, a subspace is discovered by principal component analysis from trajectories in past and present intervals, and their dissimilarity is measured by the distance between the subspaces. One of the major approaches is called *subspace identification*, which compares the subspaces spanned by the columns of an *extended observability matrix* generated by a state-space model with system noise (Kawahara et al., 2007). Recent efforts along this line of research have led to a computationally efficient algorithm based on *Krylov subspace learning* (Ide and Tsuda, 2007) and a successful application of detecting climate change in south Kenya (Itoh and Kurths, 2010).

The methods explained above rely on pre-designed parametric models, such as underlying probability distributions (Basseville and Nikiforov, 1993; Gustafsson, 1996), autoregressive models (Takeuchi and Yamanishi, 2006), and state-space models (Moskvina and Zhigljavsky, 2003a,b; Ide and Tsuda, 2007; Kawahara et al., 2007), for tracking specific statistics such as the mean, the variance, and the spectrum. As alternatives,

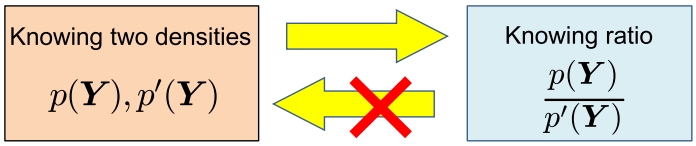


Figure 1: Rationale of direct density-ratio estimation.

non-parametric methods such as *kernel density estimation* (Csörgö and Horváth, 1988; Brodsky and Darkhovsky, 1993) are designed with no particular parametric assumption. However, they tend to be less accurate in high-dimensional problems because of the so-called *curse of dimensionality* (Bellman, 1961; Vapnik, 1998).

To overcome this difficulty, a new strategy was introduced recently, which estimates the *ratio* of probability densities directly without going through density estimation (Sugiyama et al., 2012b). The rationale of this density-ratio estimation idea is that knowing the two densities implies knowing the density ratio, but not vice versa; knowing the ratio does not necessarily imply knowing the two densities because such decomposition is not unique (Figure 1). Thus, direct density-ratio estimation is substantially easier than density estimation (Sugiyama et al., 2012b). Following this idea, methods of direct density-ratio estimation have been developed (Sugiyama et al., 2012a), e.g., *kernel mean matching* (Gretton et al., 2009), the *logistic-regression method* (Bickel et al., 2007), and the *Kullback-Leibler importance estimation procedure* (KLIEP) (Sugiyama et al., 2008). In the context of change-point detection, KLIEP was reported to outperform other approaches (Kawahara and Sugiyama, 2012) such as the *one-class support vector machine* (Schölkopf et al., 2001; Desobry et al., 2005) and *singular-spectrum analysis* (Moskvina and Zhigljavsky, 2003b). Thus, change-point detection based on direct density-ratio estimation is promising.

The goal of this paper is to further advance this line of research. More specifically, our contributions in this paper are two folds. The first contribution is to apply a recently-proposed density-ratio estimation method called the *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009) to change-point detection. The basic idea of uLSIF is to directly learn the density-ratio function in the least-squares fitting framework. Notable advantages of uLSIF are that its solution can be computed analytically (Kanamori et al., 2009), it achieves the optimal non-parametric convergence rate (Kanamori et al., 2012b), it has the optimal numerical stability (Kanamori et al., 2013), and it has higher robustness than KLIEP (Sugiyama et al., 2012a). Through experiments on a range of datasets, we demonstrate the superior detection accuracy of the uLSIF-based change-point detection method.

The second contribution of this paper is to further improve the uLSIF-based change-point detection method by employing a state-of-the-art extension of uLSIF called *relative uLSIF* (RuLSIF) (Yamada et al., 2013). A potential weakness of the density-ratio based approach is that density ratios can be unbounded (i.e., they can be infinity) if the denominator density is not well-defined. The basic idea of RuLSIF is to consider *relative density ratios*, which are smoother and always bounded from above. Theoretically, it was proved that RuLSIF possesses a superior non-parametric convergence property than plain uLSIF

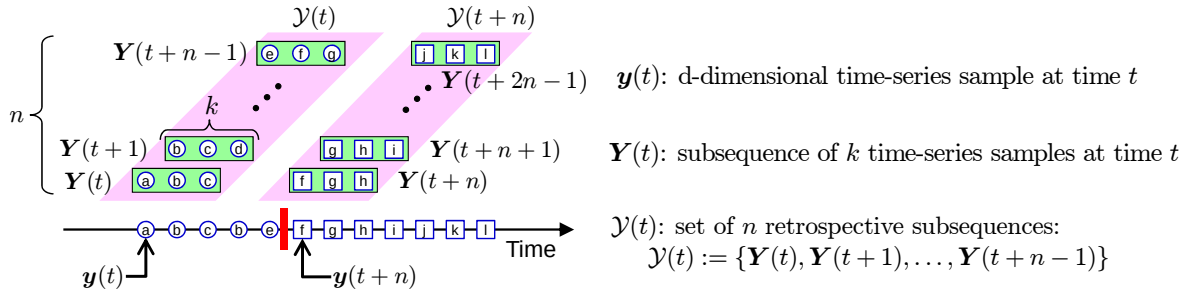


Figure 2: An illustrative example of notations on one-dimensional time-series data.

(Yamada et al., 2013), implying that RuLSIF gives an even better estimate from a small number of samples. We experimentally demonstrate that our RuLSIF-based change-point detection method compares favorably with other approaches.

The rest of this paper is structured as follows: In Section 2, we formulate our change-point detection problem. In Section 3, we describe our proposed change-point detection algorithms based on uLSIF and RuLSIF, together with the review of the KLIEP-based method. In Section 4, we report experimental results on various artificial and real-world datasets including human-activity sensing, speech, and Twitter messages from February 2010 to October 2010. Finally, in Section 5, conclusions together with future perspectives are stated.

2 Problem Formulation

In this section, we formulate our change-point detection problem.

Let $\boldsymbol{y}(t) \in \mathbb{R}^d$ be a d-dimensional time-series sample at time t. Let

$$\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{dk}$$

be a "subsequence"[1] of time series at time t with length k, where $^\top$ represents the transpose. Following the previous work (Kawahara and Sugiyama, 2012), we treat the subsequence $\boldsymbol{Y}(t)$ as a sample, instead of a single d-dimensional time-series sample $\boldsymbol{y}(t)$, by which time-dependent information can be incorporated naturally (see Figure 2). Let $\mathcal{Y}(t)$ be a set of n retrospective subsequence samples starting at time t:

$$\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+n-1)\}.$$

Note that $[\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+n-1)] \in \mathbb{R}^{dk \times n}$ forms a *Hankel matrix* and plays a key role in change-point detection based on subspace learning (Moskvina and Zhigljavsky, 2003a; Kawahara et al., 2007).

For change-point detection, let us consider two consecutive segments $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$. Our strategy is to compute a certain dissimilarity measure between $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$, and

[1] In fact, only in the case of one-dimensional time-series, $\boldsymbol{Y}(t)$ is a subsequence. For higher-dimensional time-series, $\boldsymbol{Y}(t)$ concatenates the subsequences of all dimensions into a one-dimensional vector.

use it as the plausibility of change points. More specifically, the higher the dissimilarity measure is, the more likely the point is a change point[2].

Now the problems that need to be addressed are what kind of dissimilarity measure we should use and how we estimate it from data. We will discuss these issues in the next section.

3 Change-Point Detection via Density-Ratio Estimation

In this section, we first define our dissimilarity measure, and then show methods for estimating the dissimilarity measure.

3.1 Divergence-Based Dissimilarity Measure and Density-Ratio Estimation

In this paper, we use a dissimilarity measure of the following form:

$$D(P_t \| P_{t+n}) + D(P_{t+n} \| P_t), \tag{1}$$

where P_t and P_{t+n} are probability distributions of samples in $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$, respectively. $D(P\|P')$ denotes the *f-divergence* (Ali and Silvey, 1966; Csiszár, 1967):

$$D(P\|P') := \int p'(\boldsymbol{Y}) f\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) \mathrm{d}\boldsymbol{Y}, \tag{2}$$

where f is a convex function such that $f(1) = 0$, and $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are probability density functions of P and P', respectively. We assume that $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are strictly positive. Since the f-divergence is asymmetric (i.e., $D(P\|P') \neq D(P'\|P)$), we symmetrize it in our dissimilarity measure (1) for all divergence-based methods[3].

The f-divergence includes various popular divergences such as the *Kullback-Leibler (KL) divergence* by $f(t) = t \log t$ (Kullback and Leibler, 1951) and the *Pearson (PE)*

[2]Another possible formulation is to compare distributions of samples in $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$ in the framework of *hypothesis testing* (Henkel, 1976). Although this gives a useful threshold to determine whether a point is a change point, computing the p-value is often time consuming, particularly in a non-parametric setup (Efron and Tibshirani, 1993). For this reason, we do not take the hypothesis testing approach in this paper, although it is methodologically straightforward to extend the proposed approach to the hypothesis testing framework.

[3]In the previous work (Kawahara and Sugiyama, 2012), the asymmetric dissimilarity measure $D(P_t||P_{t+n})$ was used. As we numerically illustrate in Section 4, the use of the symmetrized divergence contributes highly to improving the performance. For this reason, we decided to use the symmetrized dissimilarity measure (1).

divergence by $f(t) = \frac{1}{2}(t-1)^2$ (Pearson, 1900):

$$\mathrm{KL}(P\|P') := \int p(\boldsymbol{Y}) \log\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) \mathrm{d}\boldsymbol{Y}, \tag{3}$$

$$\mathrm{PE}(P\|P') := \frac{1}{2}\int p'(\boldsymbol{Y}) \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - 1\right)^2 \mathrm{d}\boldsymbol{Y}. \tag{4}$$

Since the probability densities $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are unknown in practice, we cannot directly compute the f-divergence (and thus the dissimilarity measure). A naive way to cope with this problem is to perform density estimation and plug the estimated densities $\widehat{p}(\boldsymbol{Y})$ and $\widehat{p}'(\boldsymbol{Y})$ in the definition of the f-divergence. However, density estimation is known to be a hard problem (Vapnik, 1998), and thus such a plug-in approach is not reliable in practice.

Recently, a novel method of divergence approximation based on *direct density-ratio estimation* was explored (Sugiyama et al., 2008; Nguyen et al., 2010; Kanamori et al., 2009). The basic idea of direct density-ratio estimation is to learn the density-ratio function $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ without going through separate density estimation of $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$. An intuitive rationale of direct density-ratio estimation is that knowing the two densities $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ means knowing their ratio, but not vice versa; knowing the ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ does not necessarily mean knowing the two densities $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ because such decomposition is not unique (see Figure 1). This implies that estimating the density ratio is substantially easier than estimating the densities, and thus directly estimating the density ratio would be more promising[4] (Sugiyama et al., 2012b).

In the rest of this section, we review three methods of directly estimating the density ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ from samples $\{\boldsymbol{Y}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}'_j\}_{j=1}^n$ drawn from $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$: The *KL importance estimation procedure* (KLIEP) (Sugiyama et al., 2008) in Section 3.2, *unconstrained least-squares importance fitting* (uLSIF) (Kanamori et al., 2009) in Section 3.3, and *relative uLSIF (RuLSIF)* (Yamada et al., 2013) in Section 3.4.

3.2 KLIEP

KLIEP (Sugiyama et al., 2008) is a direct density-ratio estimation algorithm that is suitable for estimating the KL divergence.

[4]Vladimir Vapnik advocated in his seminal book (Vapnik, 1998) that one should avoid solving a more difficult problem as an intermediate step. The *support vector machine* (Cortes and Vapnik, 1995) is a representative example that demonstrates the usefulness of this principle: It avoids solving a more general problem of estimating data generating probability distributions, and only learns a decision boundary that is sufficient for pattern recognition. The idea of direct density-ratio estimation also follows Vapnik's principle.

3.2.1 Density-Ratio Model

Let us model the density ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ by the following kernel model:

$$g(\boldsymbol{Y};\boldsymbol{\theta}) := \sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}, \boldsymbol{Y}_\ell), \tag{5}$$

where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)^\top$ are parameters to be learned from data samples, and $K(\boldsymbol{Y}, \boldsymbol{Y}')$ is a kernel basis function. In practice, we use the Gaussian kernel:

$$K(\boldsymbol{Y}, \boldsymbol{Y}') = \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{Y}'\|^2}{2\sigma^2}\right),$$

where σ (> 0) is the kernel width. In all our experiments, the kernel width σ is determined based on cross-validation.

3.2.2 Learning Algorithm

The parameters $\boldsymbol{\theta}$ in the model $g(\boldsymbol{Y};\boldsymbol{\theta})$ are determined so that the KL divergence from $p(\boldsymbol{Y})$ to $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ is minimized:

$$\begin{aligned}
\mathrm{KL} &= \int p(\boldsymbol{Y}) \log\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})g(\boldsymbol{Y};\boldsymbol{\theta})}\right) \mathrm{d}\boldsymbol{Y} \\
&= \int p(\boldsymbol{Y}) \log\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) \mathrm{d}\boldsymbol{Y} - \int p(\boldsymbol{Y}) \log\left(g(\boldsymbol{Y};\boldsymbol{\theta})\right) \mathrm{d}\boldsymbol{Y}
\end{aligned}$$

After ignoring the first term which is irrelevant to $g(\boldsymbol{Y};\boldsymbol{\theta})$ and approximating the second term with the empirical estimates, the KLIEP optimization problem is given as follows:

$$\begin{aligned}
&\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell)\right), \\
&\text{s.t. } \frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell) = 1 \text{ and } \theta_1, \ldots, \theta_n \geq 0.
\end{aligned}$$

The equality constraint is for the normalization purpose because $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ should be a probability density function. The inequality constraint comes from the non-negativity of the density-ratio function. Since this is a convex optimization problem, the unique global optimal solution $\widehat{\boldsymbol{\theta}}$ can be simply obtained, for example, by a gradient-projection iteration. Finally, a density-ratio estimator is given as

$$\widehat{g}(\boldsymbol{Y}) = \sum_{\ell=1}^{n} \widehat{\theta}_\ell K(\boldsymbol{Y}, \boldsymbol{Y}_\ell).$$

KLIEP was shown to achieve the optimal non-parametric convergence rate (Sugiyama et al., 2008; Nguyen et al., 2010).

3.2.3 Change-Point Detection by KLIEP

Given a density-ratio estimator $\widehat{g}(\boldsymbol{Y})$, an approximator of the KL divergence is given as

$$\widehat{\mathrm{KL}} := \frac{1}{n}\sum_{i=1}^{n} \log \widehat{g}(\boldsymbol{Y}_i).$$

In the previous work (Kawahara and Sugiyama, 2012), this KLIEP-based KL-divergence estimator was applied to change-point detection and demonstrated to be promising in experiments.

3.3 uLSIF

Recently, another direct density-ratio estimator called uLSIF was proposed (Kanamori et al., 2009, 2012b), which is suitable for estimating the PE divergence.

3.3.1 Learning Algorithm

In uLSIF, the same density-ratio model as KLIEP is used (see Section 3.2.1). However, its training criterion is different; the density-ratio model is fitted to the true density-ratio under the squared loss. More specifically, the parameter $\boldsymbol{\theta}$ in the model $g(\boldsymbol{Y};\boldsymbol{\theta})$ is determined so that the following squared loss $J(\boldsymbol{Y})$ is minimized:

$$\begin{aligned} J(\boldsymbol{Y}) &= \frac{1}{2}\int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - g(\boldsymbol{Y};\boldsymbol{\theta})\right)^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} \\ &= \frac{1}{2}\int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right)^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} - \int p(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{Y} + \frac{1}{2}\int g(\boldsymbol{Y};\boldsymbol{\theta})^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}. \end{aligned}$$

Since the first term is a constant, we focus on the last two terms. By substituting $g(\boldsymbol{Y};\boldsymbol{\theta})$ with our model stated in (5) and approximating the integrals by the empirical averages, the uLSIF optimization problem is given as follows:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\theta} + \frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{6}$$

where the penalty term $\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is included for a regularization purpose. λ (≥ 0) denotes the regularization parameter, which is chosen by cross-validation (Sugiyama et al., 2008). $\widehat{\boldsymbol{H}}$ is the $n \times n$ matrix with the (ℓ,ℓ')-th element given by

$$\widehat{H}_{\ell,\ell'} := \frac{1}{n}\sum_{j=1}^{n} K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell) K(\boldsymbol{Y}'_j, \boldsymbol{Y}_{\ell'}). \tag{7}$$

$\widehat{\boldsymbol{h}}$ is the n-dimensional vector with the ℓ-th element given by

$$\widehat{h}_\ell := \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell).$$

It is easy to confirm that the solution $\widehat{\boldsymbol{\theta}}$ of (6) can be analytically obtained as

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}, \tag{8}$$

where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given as

$$\widehat{g}(\boldsymbol{Y}) = \sum_{\ell=1}^{n} \widehat{\theta}_\ell K(\boldsymbol{Y}, \boldsymbol{Y}_\ell).$$

3.3.2 Change-Point Detection by uLSIF

Given a density-ratio estimator $\widehat{g}(\boldsymbol{Y})$, an approximator of the PE divergence can be constructed as

$$\widehat{\mathrm{PE}} := -\frac{1}{2n}\sum_{j=1}^{n} \widehat{g}(\boldsymbol{Y}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n} \widehat{g}(\boldsymbol{Y}_i) - \frac{1}{2}.$$

This approximator is derived from the following expression of the PE divergence (Sugiyama et al., 2010, 2011b):

$$\mathrm{PE}(P\|P') = -\frac{1}{2}\int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right)^2 p'(\boldsymbol{Y})\mathrm{d}\boldsymbol{Y} + \int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) p(\boldsymbol{Y})\mathrm{d}\boldsymbol{Y} - \frac{1}{2}. \tag{9}$$

The first two terms of (9) are actually the negative uLSIF optimization objective without regularization. This expression can also be obtained based on the fact that the f-divergence $D(P\|P')$ is lower-bounded via the *Legendre-Fenchel convex duality* (Rockafellar, 1970) as follows (Keziou, 2003; Nguyen et al., 2007):

$$D(P||P') = \sup_h \left(\int p(\boldsymbol{Y})h(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} - \int p'(\boldsymbol{Y})f^*(h(\boldsymbol{Y}))\,\mathrm{d}\boldsymbol{Y} \right), \tag{10}$$

where f^* is the convex conjugate of convex function f defined at (2). The PE divergence corresponds to $f(t) = \frac{1}{2}(t-1)^2$, for which convex conjugate is given by $f^*(t^*) = \frac{(t^*)^2}{2} + t^*$. For $f(t) = \frac{1}{2}(t-1)^2$, the supremum can be achieved when $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} = h(\boldsymbol{Y}) + 1$. Substituting $h(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - 1$ into (10), we can obtain (9).

uLSIF has some notable advantages: Its solution can be computed analytically (Kanamori et al., 2009) and it possesses the optimal non-parametric convergence rate (Kanamori et al., 2012b). Moreover, it has the optimal numerical stability (Kanamori et al., 2013), and it is more robust than KLIEP (Sugiyama et al., 2012a). In Section 4, we will experimentally demonstrate that uLSIF-based change-point detection compares favorably with the KLIEP-based method.

3.4 RuLSIF

Depending on the condition of the denominator density $p'(\boldsymbol{Y})$, the density-ratio value $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ can be unbounded (i.e., they can be infinity). This is actually problematic because the non-parametric convergence rate of uLSIF is governed by the "sup"-norm of the true density-ratio function: $\max_{\boldsymbol{Y}} \frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$. To overcome this problem, *relative density-ratio estimation* was introduced (Yamada et al., 2013).

3.4.1 Relative PE Divergence

Let us consider the *α-relative PE-divergence* for $0 \leq \alpha < 1$:

$$\begin{aligned}\mathrm{PE}_\alpha(P\|P') &:= \mathrm{PE}(P\|\alpha P + (1-\alpha)P') \\ &= \int p'_\alpha(\boldsymbol{Y}) \left(\frac{p(\boldsymbol{Y})}{p'_\alpha(\boldsymbol{Y})} - 1\right)^2 \mathrm{d}\boldsymbol{Y},\end{aligned}$$

where $p'_\alpha(\boldsymbol{Y}) = \alpha p(\boldsymbol{Y}) + (1-\alpha)p'(\boldsymbol{Y})$ is the *α-mixture density*. We refer to

$$r_\alpha(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y})}{\alpha p(\boldsymbol{Y}) + (1-\alpha)p'(\boldsymbol{Y})}$$

as the *α-relative density-ratio*. The α-relative density-ratio is reduced to the plain density-ratio if $\alpha = 0$, and it tends to be "smoother" as α gets larger. Indeed, one can confirm that the α-relative density-ratio is bounded above by $1/\alpha$ for $\alpha > 0$, even when the plain density-ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ is unbounded. This was proved to contribute to improving the estimation accuracy (Yamada et al., 2013).

As explained in Section 3.1, we use symmetrized divergence

$$\mathrm{PE}_\alpha(P\|P') + \mathrm{PE}_\alpha(P'\|P)$$

as a change-point score, where each term is estimated separately.

3.4.2 Learning Algorithm

For approximating the α-relative density ratio $r_\alpha(\boldsymbol{Y})$, we still use the same kernel model $g(\boldsymbol{Y};\boldsymbol{\theta})$ given by (5). In the same way as the uLSIF method, the parameter $\boldsymbol{\theta}$ is learned by minimizing the squared loss between true and estimated relative ratios:

$$\begin{aligned}J(\boldsymbol{Y}) &= \frac{1}{2}\int p'_\alpha(\boldsymbol{Y})\Big(r_\alpha(\boldsymbol{Y}) - g(\boldsymbol{Y};\boldsymbol{\theta})\Big)^2 \mathrm{d}\boldsymbol{Y} \\ &= \frac{1}{2}\int p'_\alpha(\boldsymbol{Y}) r_\alpha^2(\boldsymbol{Y}) \,\mathrm{d}\boldsymbol{Y} - \int p(\boldsymbol{Y}) r_\alpha(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{Y} \\ &\quad + \frac{\alpha}{2}\int p(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{Y} + \frac{1-\alpha}{2}\int p'(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{Y}.\end{aligned}$$

Again, by ignoring the constant and approximating the expectations by sample averages, the α-relative density-ratio can be learned in the same way as the plain density-ratio. Indeed, the optimization problem of a relative variant of uLSIF, called RuLSIF, is given as the same form as uLSIF; the only difference is the definition of the matrix $\widehat{\boldsymbol{H}}$:

$$\widehat{H}_{\ell,\ell'} := \frac{\alpha}{n}\sum_{i=1}^{n} K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell)K(\boldsymbol{Y}_i, \boldsymbol{Y}_{\ell'}) + \frac{(1-\alpha)}{n}\sum_{j=1}^{n} K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell)K(\boldsymbol{Y}'_j, \boldsymbol{Y}_{\ell'}).$$

Thus, the advantages of uLSIF regarding the analytic solution, numerical stability, and robustness are still maintained in RuLSIF. Furthermore, RuLSIF possesses an even better non-parametric convergence property than uLSIF (Yamada et al., 2013).

3.4.3 Change-Point Detection by RuLSIF

By using an estimator $\widehat{g}(\boldsymbol{Y})$ of the α-relative density-ratio, the α-relative PE divergence can be approximated as

$$\widehat{\mathrm{PE}}_\alpha := -\frac{\alpha}{2n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{Y}_i)^2 - \frac{1-\alpha}{2n}\sum_{j=1}^{n}\widehat{g}(\boldsymbol{Y}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{Y}_i) - \frac{1}{2}.$$

In Section 4, we will experimentally demonstrate that the RuLSIF-based change-point detection performs even better than the plain uLSIF-based method.

4 Experiments

In this section, we experimentally investigate the performance of the proposed and existing change-point detection methods on artificial and real-world datasets including human-activity sensing, speech, and Twitter messages. The MATLAB implementation of the proposed method is available at

"`http://sugiyama-www.cs.titech.ac.jp/~song/change_detection/`".

For all experiments, we fix the parameters at $n = 50$ and $k = 10$. α in the RuLSIF-based method is fixed to 0.1. Sensitivity to different parameter choices and more issues regarding algorithm-specific parameter tuning will be discussed below.

4.1 Artificial Datasets

As mentioned in Section 3.1, we use the symmetrized divergence for change-point detection. We first illustrate how symmetrization of the PE divergence affects the change-point detection performance.

The top graph in Figure 3 shows an artificial time-series signal that consists of three segments with equal length of 200. The samples are drawn from $\mathcal{N}(0, 2^2)$, $\mathcal{N}(0, 1^2)$, and $\mathcal{N}(0, 2^2)$, respectively, where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Thus, the variances change at time 200 and 400. In this experiment, we consider three types of divergence measures:

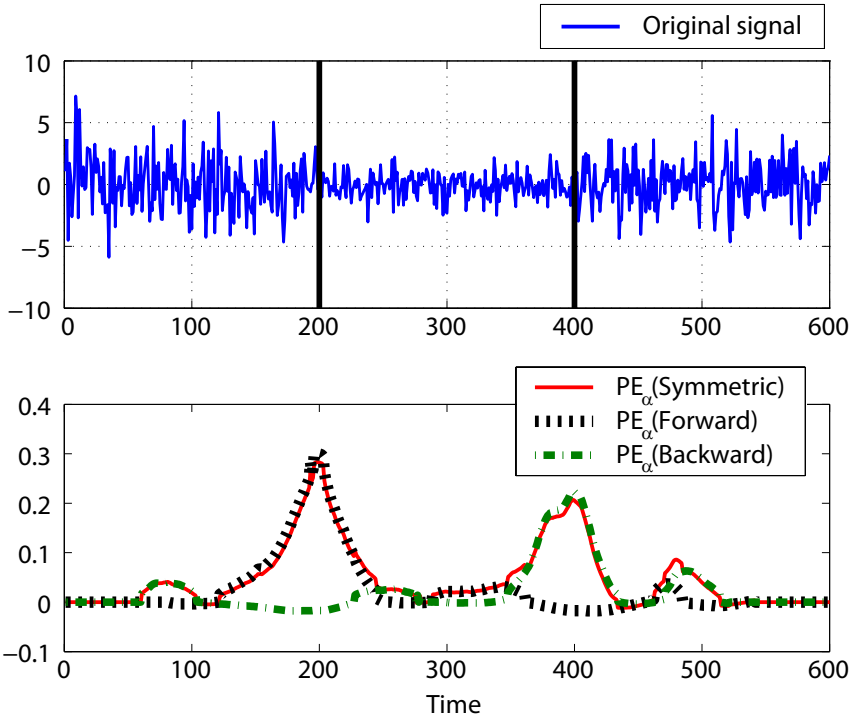


Figure 3: (Top) The original signal (blue) is segmented into 3 sections with equal length. Samples are drawn from the normal distributions $\mathcal{N}(0, 2^2)$, $\mathcal{N}(0, 1^2)$, and $\mathcal{N}(0, 2^2)$, respectively. (Bottom) Symmetric (red) and asymmetric (black and green) PE_α divergences.

- $\mathrm{PE}_\alpha(\mathrm{Symmetric}) : \mathrm{PE}_\alpha(P_t||P_{t+n}) + \mathrm{PE}_\alpha(P_{t+n}||P_t)$,
- $\mathrm{PE}_\alpha(\mathrm{Forward}) : \mathrm{PE}_\alpha(P_t||P_{t+n})$,
- $\mathrm{PE}_\alpha(\mathrm{Backward}) : \mathrm{PE}_\alpha(P_{t+n}||P_t)$.

Three divergences are compared in the bottom graph of Figure 3.

As we can see from the graphs, PE_α(Forward) detects the first change point successfully, but not the second one. On the other hand, PE_α(Backward) behaves oppositely. This implies that combining forward and backward divergences can improve the overall change-point detection performance. For this reason, we only use PE_α(Symmetric) as the change-point score of the proposed method from here on.

Next, we illustrate the behavior of our proposed RuLSIF-based method, and then compare its performance with the uLSIF-based and KLIEP-based methods. In our implementation, two sets of candidate parameters,

- $\sigma = 0.6d_{\mathrm{med}}$, $0.8d_{\mathrm{med}}$, d_{med}, $1.2d_{\mathrm{med}}$, and $1.4d_{\mathrm{med}}$,
- $\lambda = 10^{-3}$, 10^{-2}, 10^{-1}, 10^{0}, and 10^{1},

are provided to the cross-validation procedure, where d_{med} denotes the median distance between samples. The best combination of these parameters is chosen by grid search via cross-validation. We use 5-fold cross-validation for all experiments.

We use the following 4 artificial time-series datasets that contain manually inserted change-points:

- **Dataset 1 (Jumping mean):** The following 1-dimensional auto-regressive model borrowed from Takeuchi and Yamanishi (2006) is used to generate 5000 samples (i.e., $t = 1, \ldots, 5000$):

  $$y(t) = 0.6y(t-1) - 0.5y(t-2) + \epsilon_t,$$

  where ϵ_t is a Gaussian noise with mean μ and standard deviation 1.5. The initial values are set as $y(1) = y(2) = 0$. A change point is inserted at every 100 time steps by setting the noise mean μ at time t as

  $$\mu_N = \begin{cases} 0 & N = 1, \\ \mu_{N-1} + \frac{N}{16} & N = 2, \ldots, 49, \end{cases}$$

  where N is a natural number such that $100(N-1) + 1 \leq t \leq 100N$.

- **Dataset 2 (Scaling variance):** The same auto-regressive model as Dataset 1 is used, but a change point is inserted at every 100 time steps by setting the noise standard deviation σ at time t as

  $$\sigma = \begin{cases} 1 & N = 1, 3, \ldots, 49, \\ \ln(e + \frac{N}{4}) & N = 2, 4, \ldots, 48. \end{cases}$$

- **Dataset 3 (Switching covariance):** 2-dimensional samples of size 5000 are drawn from the origin-centered normal distribution, and a change point is inserted at every 100 time steps by setting the covariance matrix $\boldsymbol{\Sigma}$ at time t as

  $$\boldsymbol{\Sigma} = \begin{cases} \begin{pmatrix} 1 & -\frac{4}{5} - \frac{N-2}{500} \\ -\frac{4}{5} - \frac{N-2}{500} & 1 \end{pmatrix} & N = 1, 3, \ldots, 49, \\ \begin{pmatrix} 1 & \frac{4}{5} + \frac{N-2}{500} \\ \frac{4}{5} + \frac{N-2}{500} & 1 \end{pmatrix} & N = 2, 4, \ldots, 48. \end{cases}$$

- **Dataset 4 (Changing frequency):** 1-dimensional samples of size 5000 are generated as

  $$y(t) = \sin(\omega x) + \epsilon_t,$$

  where ϵ_t is a origin-centered Gaussian noise with standard deviation 0.8. A change point is inserted at every 100 points by changing the frequency ω at time t as

  $$\omega_N = \begin{cases} 1 & N = 1, \\ \omega_{N-1} \ln(e + \frac{N}{2}) & N = 2, \ldots, 49. \end{cases}$$

Note that, to explore the ability of detecting change points with different significance, we purposely made latter change-points more significant than earlier ones in the above datasets.

Figure 4 shows examples of these datasets for the last 10 change points and corresponding change-point score obtained by the proposed RuLSIF-based method. Although the last 10 change points are the most significant, we can see from the graphs that, for Dataset 3 and Dataset 4, these change points can be even hardly identified by human. Nevertheless, the change-point score obtained by the proposed RuLSIF-based method increases rapidly after changes occur.

Next, we compare the performance of RuLSIF-based, uLSIF-based, and KLIEP-based methods in terms of the *receiver operating characteristic (ROC) curves* and the area under the ROC curve (AUC) values. We define the *true positive rate* and *false positive rate* in the following way (Kawahara and Sugiyama, 2012):

- True positive rate (TPR): $n_{\mathrm{cr}}/n_{\mathrm{cp}}$,
- False positive rate (FPR): $(n_{\mathrm{al}} - n_{\mathrm{cr}})/n_{\mathrm{al}}$,

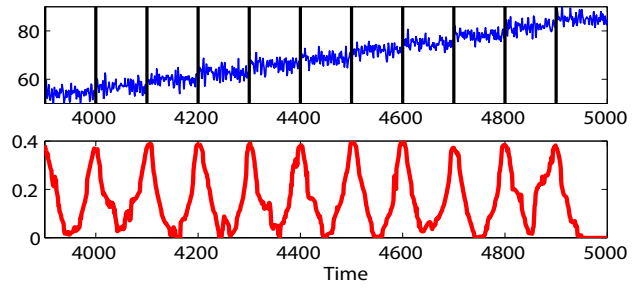
where n_{cr} denotes the number of times change points are correctly detected, n_{cp} denotes the number of all change points, and n_{al} is the number of all detection alarms.

Following the strategy of the previous researches (Desobry et al., 2005; Harchaoui et al., 2009), peaks of a change-point score are regarded as detection alarms. More specifically, a detection alarm at step t is regarded as correct if there exists a true alarm at step t^* such that $t \in [t^* - 10, t^* + 10]$. To avoid duplication, we remove the kth alarm at step t_k if $t_k - t_{k-1} < 20$.

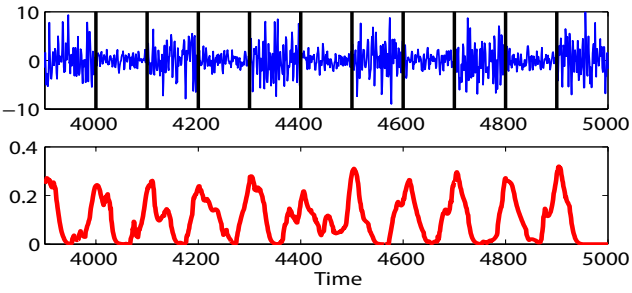
We set up a threshold η for filtering out all alarms whose change-point scores are lower than or equal to η. Initially, we set η to be equal to the score of the highest peak. Then, by lowering η gradually, both TPR and FPR become non-decreasing. For each η, we plot TPR and FPR on the graph, and thus a monotone curve can be drawn.

Figure 5 illustrates ROC curves averaged over 50 runs with different random seeds for each dataset. Table 1 describes the mean and standard deviation of the AUC values over 50 runs. The best and comparable methods by the t-test with significance level 5% are described in boldface. The experimental results show that the uLSIF-based method tends to outperform the KLIEP-based method, and the RuLSIF-based method even performs better than the uLSIF-based method.

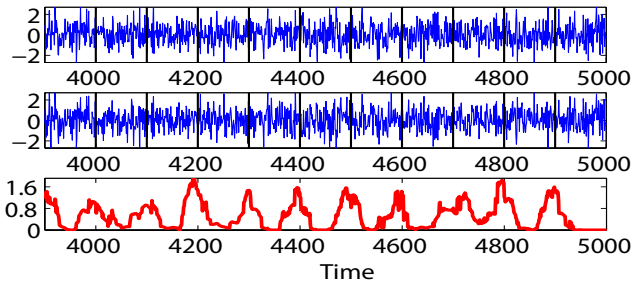
Finally, we investigate the sensitivity of the performance on different choices of n and k in terms of AUC values. In Figure 6, the AUC values of RuLSIF ($\alpha = 0.1$ and 0.2), uLSIF (which corresponds to RuLSIF with $\alpha = 0$), and KLIEP were plotted for $k = 5$, 10, and 15 under a specific choice of n in each graph. We generate such graphs for all 4 datasets with $n = 25$, 50, and 75. The result shows that the proposed method consistently performs better than the other methods, and the order of the methods according to the performance is kept unchanged over various choices of n and k. Moreover, the RuLSIF methods with $\alpha = 0.1$ and 0.2 perform rather similarly. For this reason, we keep using the medium parameter values among the candidates in the following experiments: $n = 50$, $k = 10$, and $\alpha = 0.1$.



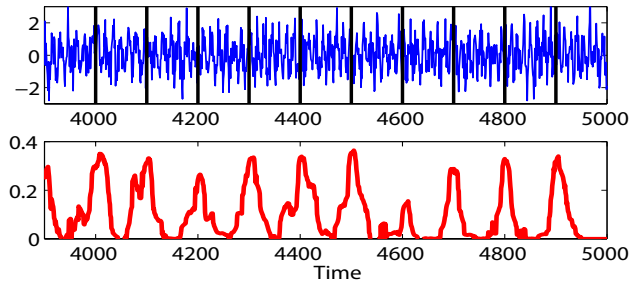
(a) Dataset1



(b) Dataset2

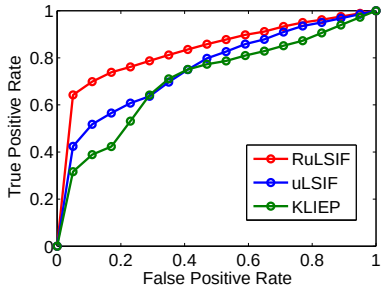


(c) Dataset3

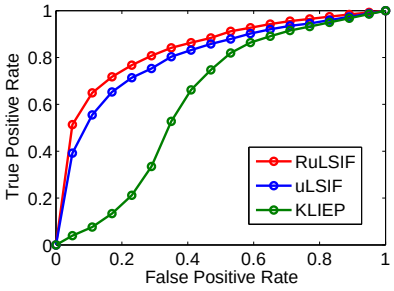


(d) Dataset4

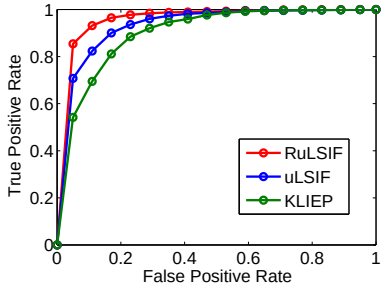
Figure 4: Illustrative time-series samples (upper) and the change-point score obtained by the RuLSIF-based method (lower). The true change-points are marked by black vertical lines in the upper graphs.



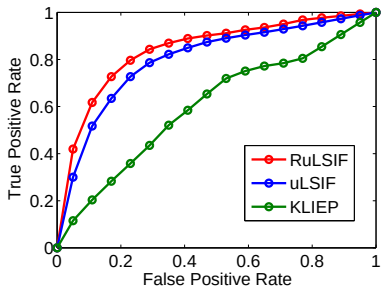
(a) Dataset1



(b) Dataset2

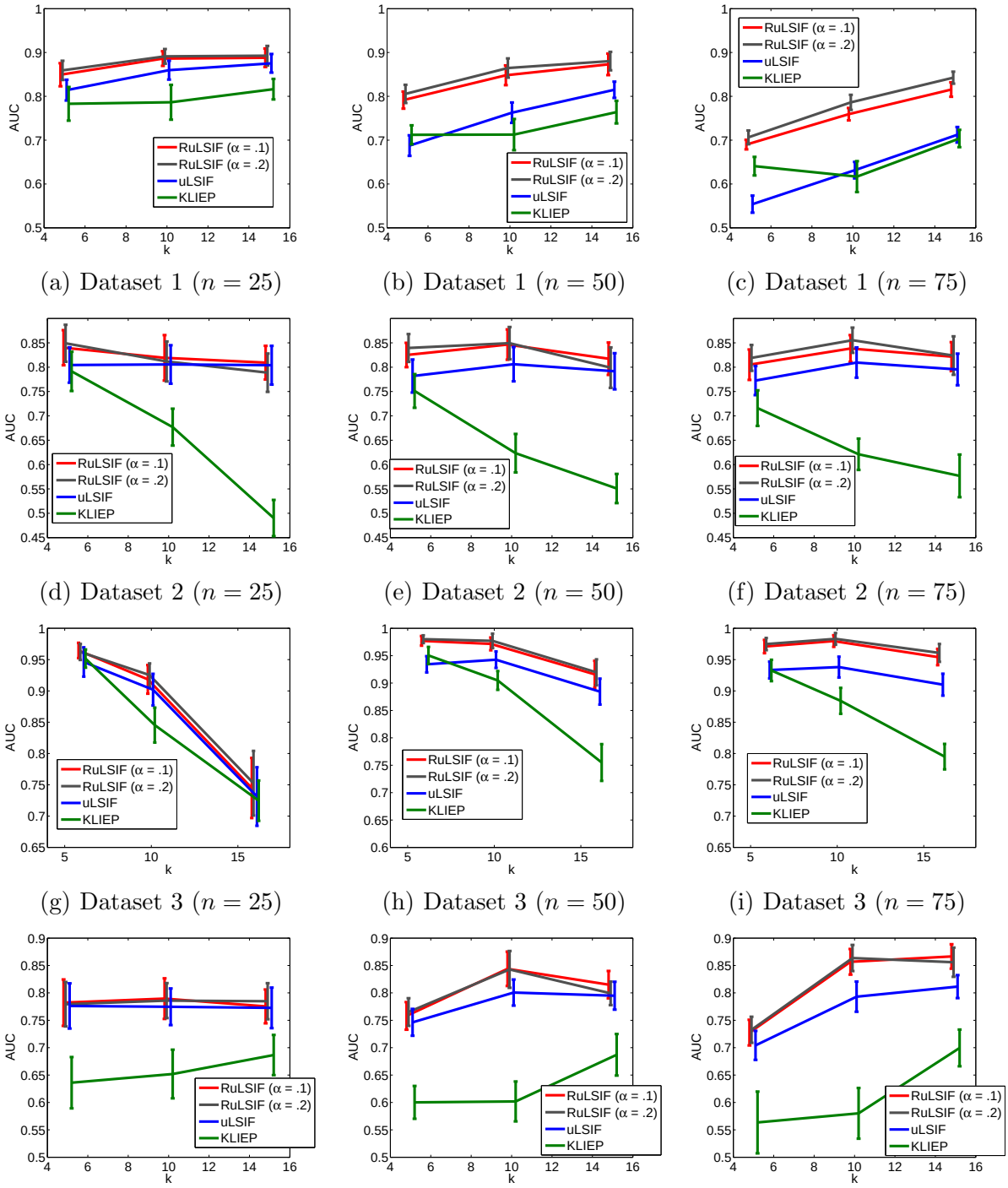


(c) Dataset3



(d) Dataset4

Figure 5: Average ROC curves of RuLSIF-based, uLSIF-based, and KLIEP-based methods.



(a) Dataset 1 ($n = 25$) (b) Dataset 1 ($n = 50$) (c) Dataset 1 ($n = 75$)

(d) Dataset 2 ($n = 25$) (e) Dataset 2 ($n = 50$) (f) Dataset 2 ($n = 75$)

(g) Dataset 3 ($n = 25$) (h) Dataset 3 ($n = 50$) (i) Dataset 3 ($n = 75$)

(j) Dataset 4 ($n = 25$) (k) Dataset 4 ($n = 50$) (l) Dataset 4 ($n = 75$)

Figure 6: AUC plots for $n = 25, 50, 75$ and $k = 5, 10, 15$. The horizontal axes denote k, while the vertical axes denote AUC values.

Table 1: The AUC values of RuLSIF-based, uLSIF-based, and KLIEP-based methods. The best and comparable methods by the t-test with significance level 5% are described in boldface.

	RuLSIF	uLSIF	KLIEP
Dataset 1	**.848(.023)**	.763(.023)	.713(.036)
Dataset 2	**.846(.031)**	.806(.035)	.623(.040)
Dataset 3	**.972(.012)**	.943(.015)	.904(.017)
Dataset 4	**.844(.031)**	.801(.024)	.602(.036)

4.2 Real-World Datasets

Next, we evaluate the performance of the density-ratio estimation based methods and other existing change-point detection methods using two real-world datasets: Human-activity sensing and speech.

We include the following methods in our comparison.

- **Singular spectrum transformation (SST) (Moskvina and Zhigljavsky, 2003a; Ide and Tsuda, 2007; Itoh and Kurths, 2010):** Change-point scores are evaluated on two consecutive trajectory matrices using the distance-based singular spectrum analysis. This corresponds to a state-space model with no system noise. For this method, we use the first 4 eigenvectors to compare the difference between two subspaces, which was confirmed to be reasonable choice in our preliminary experiments.

- **Subspace identification (SI) (Kawahara et al., 2007):** SI identifies a subspace in which time-series data is constrained, and evaluates the distance of target sequences from the subspace. The subspace spanned by the columns of an observability matrix is used for estimating the distance from the subspace spanned by subsequences of time-series data. For this method, we use the top 4 significant singular values according to our preliminary experiment results.

- **Auto regressive (AR) (Takeuchi and Yamanishi, 2006):** AR first fits an AR model to time-series data, and then auxiliary time-series is generated from the AR model. With an extra AR model-fitting, the change-point score is given by the log-likelihood. The order of the AR model is chosen by Schwarz's Bayesian information criterion (Schwarz, 1978).

- **One-class support vector machine (OSVM) (Desobry et al., 2005):** Change-point scores are calculated by OSVM using two sets of descriptors of signals. The kernel width σ is set to the median value of the distances between samples, which is a popular heuristic in kernel methods (Schölkopf and Smola, 2002). Another parameter ν is set to 0.2, which indicates the proportion of outliers.

First, we use a human activity dataset. This is a subset of the *Human Activity Sensing Consortium (HASC) challenge 2011*[5], which provides human activity information collected by portable three-axis accelerometers. The task of change-point detection is to segment the time-series data according to the 6 behaviors: "*stay*", "*walk*", "*jog*", "*skip*", "*stair up*", and "*stair down*". The starting time of each behavior is arbitrarily decided by each user. Because the orientation of accelerometers is not necessarily fixed, we take the ℓ_2-norm of the 3-dimensional (i.e., x-, y-, and z-axes) data.

In Figure 7(a), examples of original time-series, true change points, and change-point scores obtained by the RuLSIF-based method are plotted. This shows that the change-point score clearly captures trends of changing behaviors, except the changes around time 1200 and 1500. However, because these changes are difficult to be recognized even by human, we do not regard them as critical flaws. Figure 7(b) illustrates ROC curves averaged over 10 datasets, and Figure 7(c) describes AUC values for each of the 10 datasets. The experimental results show that the proposed RuLSIF-based method tends to perform better than other methods.

Next, we use the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset provided by National Institute of Informatics (NII)[6], which records human voice in a noisy environment. The task is to extract speech sections from recorded signals. This dataset offers several voice recordings with different background noises (e.g., noise of highway and restaurant). Segmentation of the beginning and ending of human voice is manually annotated. Note that we only use the annotations as the ground truth for the final performance evaluation, not for change-point detection (i.e., this experiment is still completely unsupervised).

Figure 8(a) illustrates an example of the original signals, true change-points, and change-point scores obtained by the proposed RuLSIF-based method. This shows that the proposed method still gives clear indications for speech segments. Figure 8(b) and Figure 8(c) show average ROC curves over 10 datasets and AUC values for each of the 10 datasets. The results show that the proposed method significantly outperforms other methods.

4.3 Twitter Dataset

Finally, we apply the proposed change-point detection method to the *CMU Twitter dataset*[7], which is an archive of Twitter messages collected from February 2010 to October 2010 via the Twitter application programming interface.

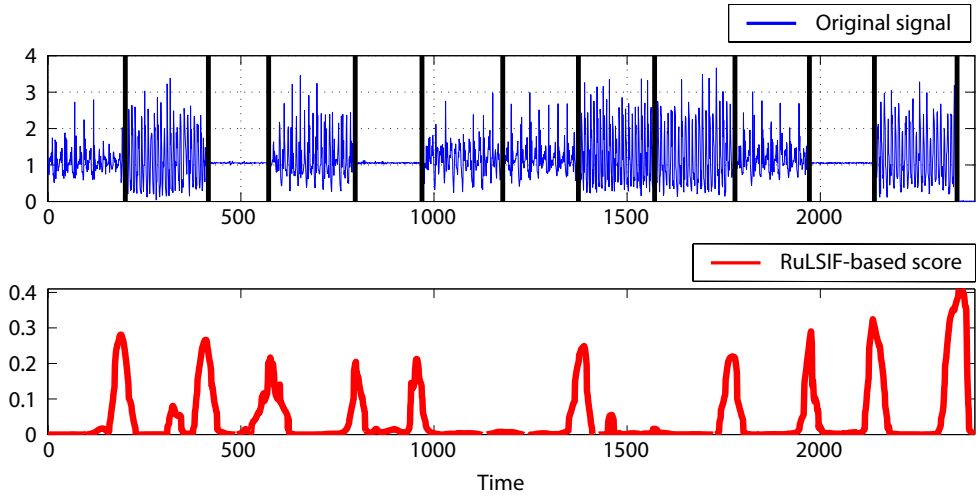
Here we track the degree of popularity of a given topic by monitoring the frequency of selected keywords. More specifically, we focus on events related to "*Deepwater Horizon oil spill in the Gulf of Mexico*" which occurred on April 20, 2010[8], and was widely broadcast among the Twitter community. We use the frequencies of 10 keywords: "*gulf*", "*spill*",

[5] `http://hasc.jp/hc2011/`

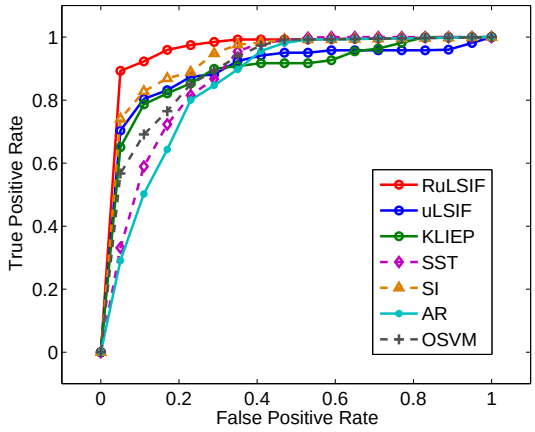
[6] `http://research.nii.ac.jp/src/eng/list/index.html`

[7] `http://www.ark.cs.cmu.edu/tweets/`

[8] `http://en.wikipedia.org/wiki/Deepwater_Horizon_oil_spill`



(a) One of the original signals and change-point scores obtained by the RuLSIF-based method

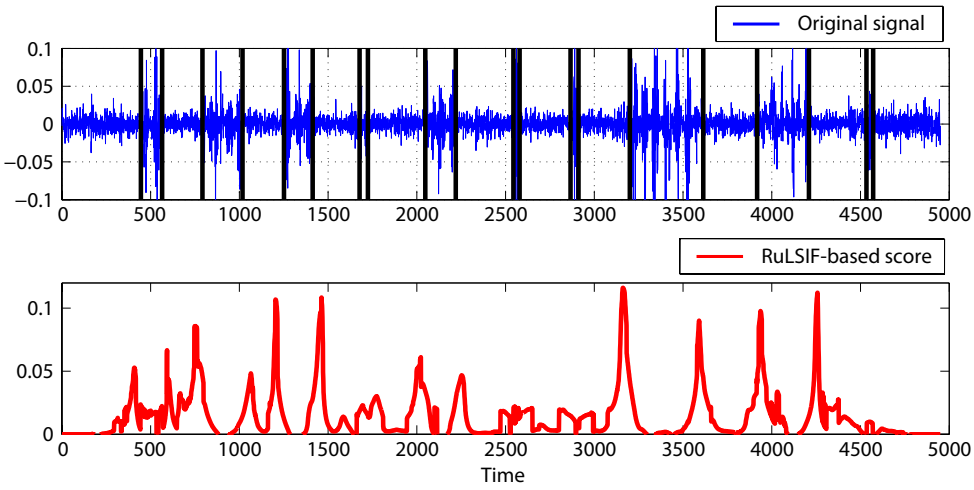


(b) Average ROC curves

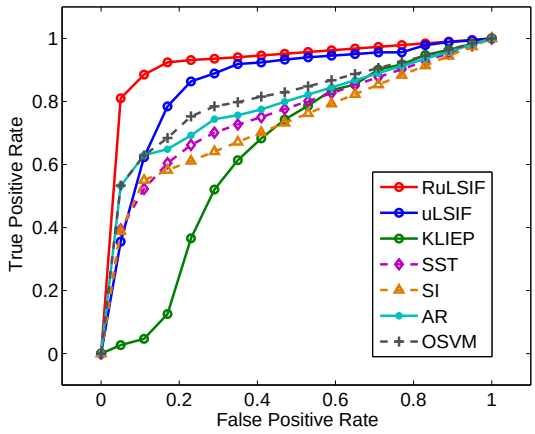
ID	RuLSIF	uLSIF	KLIEP	AR	SI	SST	OSVM
1001	.974	.853	.838	.899	.958	.903	.900
1002	.996	.963	.909	.872	.969	.880	.905
1003	.989	.854	.929	.869	.895	.851	.937
1004	.996	.868	.890	.881	.941	.886	.891
1005	.938	.952	.972	.849	.972	.915	.943
1006	.933	.918	.889	.778	.890	.925	.842
1007	.972	.857	.834	.850	.941	.817	.891
1008	.995	.922	.930	.892	.981	.860	.907
1009	.987	.880	.907	.833	.979	.842	.951
1010	.991	.952	.889	.821	.915	.867	.903
Ave.	**.977**	.902	.900	.854	.944	.875	.907
Std.	**.024**	.044	.042	.037	.034	.034	.032

(c) AUC values. The best and comparable methods by the t-test with significance level 5% are described in boldface.

Figure 7: HASC human-activity dataset.



(a) One of the original signals and change-point scores obtained by the RuLSIF-based method

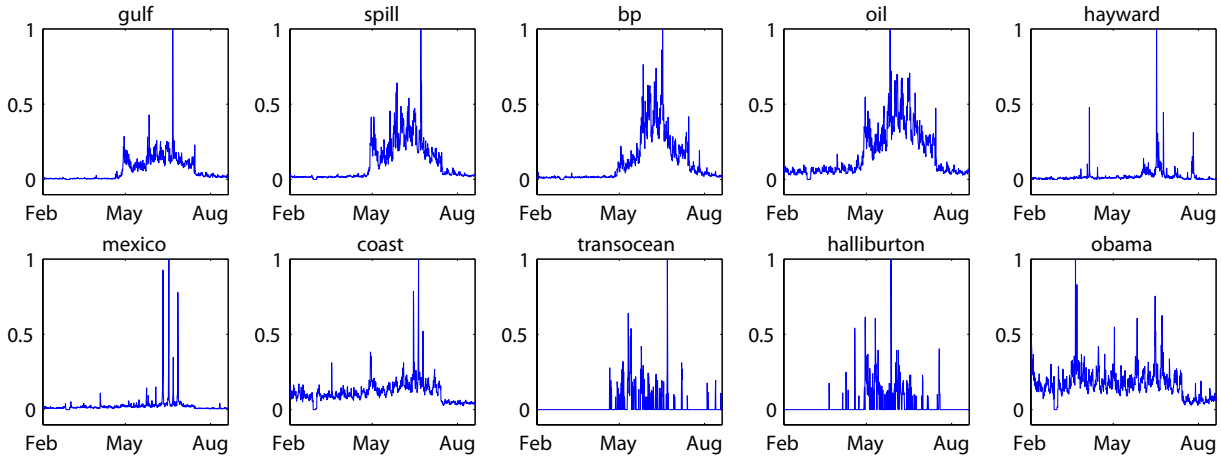


(b) Average ROC curves

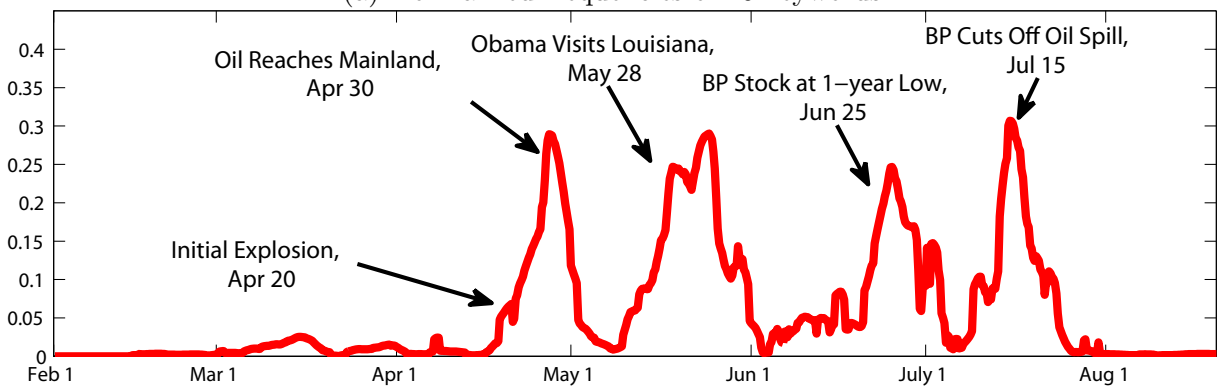
ID	RuLSIF	uLSIF	KLIEP	AR	SI	SST	OSVM
01	1.00	.902	.650	.860	.690	.806	.800
02	.911	.845	.712	.733	.800	.745	.725
03	.963	.931	.708	.910	.899	.807	.932
04	.903	.813	.587	.816	.735	.685	.751
05	.927	.907	.565	.831	.823	.809	.840
06	.857	.913	.676	.868	.740	.736	.838
07	.987	.797	.657	.807	.759	.797	.829
08	.962	.757	.581	.629	.704	.682	.800
09	.924	.913	.693	.738	.744	.781	.790
10	.966	.856	.554	.796	.725	.790	.850
Ave.	**.940**	.863	.638	.798	.762	.764	.815
Std.	**.044**	.059	.061	.081	.063	.049	.057

(c) AUC values. The best and comparable methods by the t-test with significance level 5% are described in boldface.

Figure 8: CENSREC speech dataset.



(a) Normalized frequencies of 10 keywords



(b) Change-point score obtained by the RuLSIF-based method and exemplary real-world events

Figure 9: Twitter dataset.

"*bp*", "*oil*", "*hayward*", "*mexico*", "*coast*", "*transocean*", "*halliburton*", and "*obama*" (see Figure 9(a)). We perform change-point detection directly on the 10-dimensional data, with the hope that we can capture correlation changes between multiple keywords, in addition to changes in the frequency of each keyword.

For quantitative evaluation, we referred to the Wikipedia entry "Timeline of the Deepwater Horizon oil spill"[9] as a real-world event source. The change-point score obtained by the proposed RuLSIF-based method is plotted in Figure 9(b), where four occurrences of important real-world events show the development of this news story.

As we can see from Figure 9(b), the change-point score increases immediately after the initial explosion of the deepwater horizon oil platform and soon reaches the first peak when oil was found on the sea shore of Louisiana on April 30. Shortly after BP announced its preliminary estimation on the amount of leaking oil, the change-point score rises quickly again and reaches its second peak at the end of May, at which time President Obama visited Louisiana to assure local residents of the federal government's support. On June 25, the BP stock was at its one year's lowest price, while the change-point score spikes at the third time. Finally, BP cut off the spill on July 15, as the score reaches its last peak.

[9] `http://en.wikipedia.org/wiki/Timeline_of_the_Deepwater_Horizon_oil_spill`

5 Conclusion and Future Perspectives

In this paper, we first formulated the problem of retrospective change-point detection as the problem of comparing two probability distributions over two consecutive time segments. We then provided a comprehensive review of state-of-the-art density-ratio and divergence estimation methods, which are key building blocks of our change-point detection methods. Our contributions in this paper were to extend the existing KLIEP-based change-point detection method (Kawahara and Sugiyama, 2012), and to propose to use uLSIF as a building block. uLSIF has various theoretical and practical advantages, for example, the uLSIF solution can be computed analytically, it possesses the optimal non-parametric convergence rate, it has the optimal numerical stability, and it has higher robustness than KLIEP. We further proposed to use RuLSIF, a novel divergence estimation paradigm emerged in the machine learning community recently. RuLSIF inherits good properties of uLSIF, and moreover it possesses an even better non-parametric convergence property. Through extensive experiments on artificial datasets and real-world datasets including human-activity sensing, speech, and Twitter messages, we demonstrated that the proposed RuLSIF-based change-point detection method is promising.

Though we estimated a density ratio between two consecutive segments, some earlier researches (Basseville and Nikiforov, 1993; Gustafsson, 1996, 2000) introduced a hyper-parameter that controls the size of a margin between two segments. In our preliminary experiments, however, we did not observe significant improvement by changing the margin. For this reason, we decided to use a straightforward model that two segments have no margin in between.

Through the experiment illustrated in Figure 6 in Section 4.1, we can see that the performance of the proposed method is affected by the choice of hyper-parameters n and k. However, discovering optimal values for these parameters remains a challenge, which will be investigated in our future work.

RuLSIF was shown to possess a better convergence property than uLSIF (Yamada et al., 2013) in terms of density ratio estimation. However, how this theoretical advantage in density ratio estimation can be translated into practical performance improvement in change detection is still not clear, beyond the intuition that a better divergence estimator gives a better change score. We will address this issue more formally in the future work.

Although the proposed RuLSIF-based change-point detection was shown to work well even for multi-dimensional time-series data, its accuracy may be further improved by incorporating *dimensionality reduction*. Recently, several attempts were made to combine dimensionality reduction with direct density-ratio estimation (Sugiyama et al., 2010, 2011b; Yamada and Sugiyama, 2011). Our future work will apply these techniques to change-point detection and evaluate their practical usefulness.

Compared with other approaches, methods based on density ratio estimation tend to be computationally more expensive because of the cross-validation procedure for model selection. However, thanks to the analytic solution, the RuLSIF- and uLSIF-based methods are computationally more efficient than the KLIEP-based method that requires an iterative optimization procedure (see Figure 9 in Kanamori et al. (2009) for the detailed

time comparison between uLSIF and KLIEP). Our important future work is to further improve the computational efficiency of the RuLSIF-based method.

In this paper, we focused on computing the change-point score that represents the plausibility of change points. Another possible formulation is hypothesis testing, which provides a useful threshold to determine whether a point is a change point. Methodologically, it is straightforward to extend the proposed method to produce the p-values, following the recent literatures (Sugiyama et al., 2011a; Kanamori et al., 2012a). However, computing the p-value is often time consuming, particularly in a non-parametric setup. Thus, overcoming the computational bottleneck is an important future work for making this approach more practical.

Recent reports pointed out that Twitter messages can be indicative of real-world events (Petrović et al., 2010; Sakaki et al., 2010). Following this line, we showed in Section 4.3 that our change-detection method can be used as a novel tool for analyzing Twitter messages. An important future challenge along this line includes automatic keyword selection for topics of interests.

Acknowledgements

SL was supported by NII internship fund and the JST PRESTO program. MY and MS were supported by the JST PRESTO program. NC was supported by NII Grand Challenge project fund.

References

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, arXiv, 2007. arXiv:0710.3742v1 [stat.ML].

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28(1):131–142, 1966.

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, USA, 1961.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 81–88, 2007.

B. Brodsky and B. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1993.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

M. Csörgö and L. Horváth. 20 nonparametric methods for changepoint problems. In P. R. Krishnaiah and C. R. Rao, editors, *Handbook of Statistics*, volume 7, pages 403–425. Elsevier, Amsterdam, the Netherlands, 1988.

F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, 2005.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, NY, USA, 1993.

R. Garnett, M. A. Osborne, and S. J. Roberts. Sequential Bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 345–352, 2009.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, Cambridge, MA, USA, 2009.

V. Guralnik and J. Srivastava. Event detection from time series data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 33–42, 1999.

F. Gustafsson. The marginalized likelihood ratio test for detecting abrupt changes. *IEEE Transactions on Automatic Control*, 41(1):66–78, 1996.

F. Gustafsson. *Adaptive Filtering and Change Detection*. Wiley, Chichester, UK, 2000.

Z. Harchaoui, F. Bach, and E Moulines. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pages 609–616, 2009.

R. E. Henkel. *Tests of Significance*. SAGE Publication, Beverly Hills, CA, USA, 1976.

S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2):309–336, 2011.

T. Ide and K. Tsuda. Change-point detection using Krylov subspace learning. In *Proceedings of the SIAM International Conference on Data Mining*, pages 515–520, 2007.

N. Itoh and J. Kurths. Change-point detection of climate time series by nonparametric method. In *Proceedings of the World Congress on Engineering and Computer Science 2010*, volume 1, 2010.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

T. Kanamori, T. Suzuki, and M. Sugiyama. f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720, 2012a.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012b.

T. Kanamori, T. Suzuki, and M. Sugiyama. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, 2013. to appear.

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.

Y. Kawahara, T. Yairi, and K. Machida. Change-point detection in time-series data based on subspace identification. In *Proceedings of the 7th IEEE International Conference on Data Mining*, pages 559–564, 2007.

A. Keziou. Dual representation of ϕ-divergences and applications. *Comptes Rendus Mathematique*, 336(10):857–862, 2003.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

V. Moskvina and A. Zhigljavsky. Application of the singular-spectrum analysis to change-point detection in time series. *Journal of Sequential Analysis*, 2003a. In submission.

V. Moskvina and A. Zhigljavsky. Change-point detection algorithm based on the singular-spectrum analysis. *Communications in Statistics: Simulation and Computation*, 32: 319–352, 2003b.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. In *Proceedings of IEEE International Symposium on Information Theory*, pages 2016–2020, Nice, France, 2007.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

U. Paquet. Empirical Bayesian change point detection. *Graphical Models*, 1995:1–20, 2007.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189, 2010.

J. Reeves, J. Chen, X. L. Wang, R. Lund, and Q. Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, USA, 1970.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010.

B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

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

M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

M. Sugiyama, T. Suzuki, Y. Itoh, T. Kanamori, and M. Kimura. Least-squares two-sample test. *Neural Networks*, 24(7):735–751, 2011a.

M. Sugiyama, M. Yamada, P. von Bünau, T. Suzuki, T. Kanamori, and M. Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183–198, 2011b.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density ratio matching under the Bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012a.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, Cambridge, UK, 2012b.

J. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from non-stationary time series data. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):482–492, 2006.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

Y. Wang, C. Wu, Z. Ji, B. Wang, and Y. Liang. Non-parametric change-point method for differential gene expression detection. *PLoS ONE*, 6(5):e20060, 2011.

M. Yamada and M. Sugiyama. Direct density-ratio estimation with dimensionality reduction via hetero-distributional subspace analysis. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 549–554, Aug. 7–11 2011.

M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 2013. to appear.

K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 320–324, 2000.