Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Change-Point Detection with Feature Selection in High-Dimensional Time-Series Data

Makoto Yamada, Akisato Kimura, Futoshi Naya, and Hiroshi Sawada
NTT Communication Science Laboratories, 2-4, Hikaridai, Seika-cho, Kyoto, 619-0237, Japan
myamada0321@gmail.com, {kimura.akisato, naya.futoshi, sawada.hiroshi}@lab.ntt.co.jp

Abstract

Change-point detection is the problem of finding abrupt changes in time-series, and it is attracting a lot of attention in the artificial intelligence and data mining communities. In this paper, we present a supervised learning based change-point detection approach in which we use the separability of past and future data at time t (they are labeled as $+1$ and -1) as plausibility of change-points. Based on this framework, we propose a detection measure called the *additive Hilbert-Schmidt Independence Criterion* (aHSIC), which is defined as the weighted sum of the HSIC scores between features and its corresponding binary labels. Here, the HSIC is a kernel-based independence measure. A novel aspect of the aHSIC score is that it can incorporate feature selection during its detection measure estimation. More specifically, we first select features that are responsible for an abrupt change by using a supervised approach, and then compute the aHSIC score by employing the selected features. Thus, compared with traditional detection measures, our approach tends to be robust as regards noise features, and so the aHSIC is suitable for a use with high-dimensional time-series change-point detection problems. We demonstrate that the proposed change-point detection method is promising through extensive experiments on synthetic data sets and a real-world human activity data set.

1 Introduction

Change-point detection, which is the problem of detecting abrupt changes in time-series data, is attracting a lot of attention in the artificial intelligence and data mining communities [Basseville *et al.*, 1993; Brodsky and Darkhovsky, 1993; Kifer *et al.*, 2004], and there are various types of real-world applications such as fraud detection in cellular systems [Murad and Pinkas, 1999], intrusion detection in computer networks [Yamanishi *et al.*, 2000], irregular-motion detection in vision systems [Ke *et al.*, 2007], music segmentation [Desobry *et al.*, 2005], and sentiment analysis from Twitter data [Liu *et al.*, 2013]. Recently, the problem of change-point detection from high-dimensional time-series data such as music and multi-sensor data has been attracting increasing attention [Desobry *et al.*, 2005; Liu *et al.*, 2013]. Compared with traditional change-point detection problems, the number of features d tends to be larger than the number of data points n, and it includes a large number of noise features. This makes the change-point detection problem from high-dimensional time-series data challenging.

An effective change-point detection approach would be to use the divergence between probability distributions of data in the past and data in the future at time t. Specifically, we regard the time t point as a change-point if the divergence between two distribution is significantly large. Various change-point detection methods have been proposed based on this concept including generalized likelihood ratio (GLR) and cumulative sum approaches [Gustafsson, 1996; Basseville *et al.*, 1993]. In these approaches, *the logarithm of the likelihood ratio* between two probability distributions is used as a measure of change-point detection, where each probability density is estimated independently by density estimation. However, since density estimation is known to be a difficult problem [Härdle *et al.*, 2004; Huang *et al.*, 2007], density estimation based approaches tend to perform poorly. Moreover, since high-dimensional time-series data includes a large number of noise features, the density estimation accuracy tends to be degraded by noise.

To avoid using density estimation, direct density-ratio based change-point detection approaches have been proposed [Kawahara and Sugiyama, 2009; Liu *et al.*, 2013]. These approaches estimate the ratio of probability distributions directly without using density estimation. Direct density-ratio estimation has been actively studied by the machine learning community and techniques include kernel mean matching (KMM) [Huang *et al.*, 2007], the Kullback-Leibler importance estimation procedure (KLIEP) [Sugiyama *et al.*, 2008], WKV [Nguyen *et al.*, 2010], and unconstrained least-squares importance fitting (uLSIF) and its robust extension called relative uLSIF (RuLSIF) [Kanamori *et al.*, 2009; Yamada *et al.*, 2011]. These direct density-ratio estimation methods have exhibited the optimal convergence rate for non-parametric density-ratio estimation. However, as with density estimation based methods, the accuracy of the density-ratio estimation is likely to be degraded by noise features.

Change-point detection with stationary subspace analysis (SSA), which is a dimensionality reduction method, is

a promising change-point detection method for multivariate time-series data [Blythe *et al.*, 2012]. SSA factorizes a multivariate time-series data into stationary and non-stationary sources, and the change-points can be detected in a non-stationary subspace. Since SSA can reduce the dimensionality of data without losing the abrupt change characteristic, it can significantly improve change-point detection performance. However, since SSA needs to compute the log of a covariance matrix which is singular when $(d < n)$, it needs a large number of training samples $(n \gg d)$ to accurately factorize stationary and non-stationary sources. Therefore, an SSA based change-point detection algorithm is not applicable to high-dimensional change-point detection problems.

In this paper to deal with high-dimensional time-series data, we integrate feature selection into change-point detection measure estimation. More specifically, we present a supervised learning based change-point detection approach in which we use the separability of past and future data at time t (they are labeled as $+1$ and -1) as a plausibility of a change-points [Desobry *et al.*, 2005; Hido *et al.*, 2008]. If we can separate the past and future data sets easily, we regard them as having different probability distributions. On the other hand, if two data sets are not separable, we regard them as having the same probability distribution. Using this framework as a basis, we propose a detection measure called the *additive Hilbert-Schmidt independence criterion* (aHSIC), which is given as the weighted sum of the HSIC scores where HSIC is a kernel-based independence measure [Gretton *et al.*, 2005]. Here, each HSIC score is computed from samples of a feature and its corresponding pseudo binary labels. An advantage of the aHSIC score over existing detection measures is that it can incorporate feature selection during detection measure estimation. That is, we first select features that are responsible for an abrupt change in a supervised manner, and then compute an aHSIC score by using those selected features. Thus, compared with traditional detection measures, aHSIC is more robust as regards noise features than existing measures and it is suited for high-dimensional time-series data. Experiments on synthetic and real-world human activity data showed that the proposed methods are promising.

2 Problem Formulation

In this section, we formulate our change-point detection problem based on supervised learning framework [Desobry *et al.*, 2005; Hido *et al.*, 2008].

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

$$\mathcal{X}(t) := \{\boldsymbol{x}(t-i+1)\}_{i=1}^{n}$$

are samples with length n extracted from time series data in a sliding-window manner at time t.

Let us consider two non-overlapped sequences $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$, and we annotate the samples in $\mathcal{X}(t)$ as $y = 1$ and the samples in $\mathcal{X}(t+n)$ as $y = -1$. We denote the augmented sequences as

$$\mathcal{Z}(t) := \{(\boldsymbol{x}(t+n-i+1), y_i)\}_{i=1}^{2n},$$

where $y_i \in \{+1,\ -1\}$ is a pseudo binary label. See Figure 1 for an illustrative example of two intervals $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$. Note, in high-dimensional time-series data, the number of dimensions d is usually bigger than that of samples $2n$ (i.e., $d > 2n$).

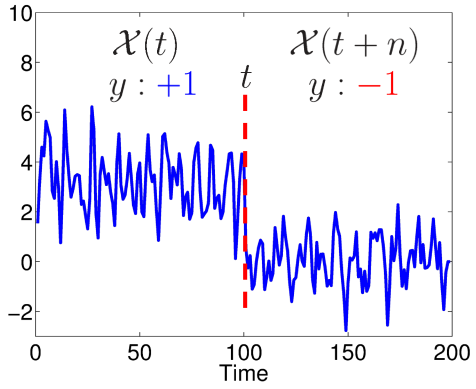


Figure 1: Illustrative example of two intervals $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$ for 1D time series.

Then, our change-point detection strategy is to compute a dependency score between input $\boldsymbol{x}$ and output y from data $\mathcal{Z}(t)$, and use it as the change-point plausibility:

$$\begin{cases} D(\mathcal{Z}(t)) < \tau & \text{(No abrupt change occurs)} \\ D(\mathcal{Z}(t)) \geq \tau & \text{(An abrupt change occurs)}, \end{cases}$$

where $D(\mathcal{Z}(t))$ is a dependency measure that takes a large value if $\boldsymbol{x}$ and y are dependent and τ is a threshold that controls the sensibility/robustness tradeoff. Note, since y can be regarded as a step function like sequence that exhibits a change at time t, a large dependency value means that $\mathcal{X}(t)$ and $\mathcal{X}(t+1)$ are separable. That is, we can regard $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$ as samples from different distributions. In contrast, if the dependency score is small, then we can regard $\mathcal{X}(t)$ and $\mathcal{X}(t+n)$ as samples from the same distribution.

3 Proposed Change-Point Detection Method

The detection performance depends strongly on the dependency measure. In particular, since high-dimensional data tends to include a large number of noise features, a key issue is to compute a detection measure for an abrupt change using only important features. Thus, we incorporate feature selection into our change-point detection measure. To the best of our knowledge, this is the first work to perform feature selection in change-point detection for high-dimensional data.

In this section, we first propose our dependency measure, and then show a way to estimate it.

3.1 Additive Hilbert-Schmidt Independence Criterion

We propose an *additive HSIC* (aHSIC) score as a dependency measure $D(\mathcal{Z}(t))$:

$$\text{aHSIC}(\mathcal{Z}(t)) := \sum_{k=1}^{d} \alpha_k \text{HSIC}(\boldsymbol{u}_k(t), \boldsymbol{y}),$$

where $\boldsymbol{u}_k(t) = [x_k(t-n+1), x_k(t-n+2), \ldots, x_k(t+n-1)]^\top \in \mathbb{R}^{2n}$ is the k-th feature for all samples, $^\top$ is the matrix transpose, $\boldsymbol{y} = [1, \ldots, 1, -1, \ldots, -1]^\top \in \mathbb{R}^{2n}$ is the pseudo binary label vector, $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{y}) = \mathrm{tr}(\bar{\boldsymbol{K}}^{(k)}\bar{\boldsymbol{L}})$ is a kernel-based independence measure called the (empirical) *Hilbert-Schmidt independence criterion* (HSIC) [Gretton *et al.*, 2005], $\mathrm{tr}(\cdot)$ denotes the trace, $\alpha_1, \ldots, \alpha_d \geq 0$, $\sum_{k=1} \alpha_k = 1$, $\bar{\boldsymbol{K}}^{(k)} = \boldsymbol{\Gamma}\boldsymbol{K}^{(k)}\boldsymbol{\Gamma}$ and $\bar{\boldsymbol{L}} = \boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}$ are centered and normalized Gram matrices, $\boldsymbol{K}^{(k)}_{i,j} = K(x_{k,i}, x_{k,j})$ and $\boldsymbol{L}_{i,j} = L(y_i, y_j)$ are Gram matrices, $K(x, x')$ and $L(y, y')$ are kernel functions, $\boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix, $\boldsymbol{I}_n$ is the n-dimensional identity matrix, and $\mathbf{1}_n$ is the n-dimensional vector with all ones.

Note, HSIC always takes a non-negative value, and is zero if and only if two random variables are statistically independent when a *universal reproducing kernel* [Steinwart, 2001] such as a Gaussian kernel is used. That is, if the k-th feature $\boldsymbol{u}_k$ is independent of $\boldsymbol{y}$ (i.e., the k-th feature is not important for an abrupt change), $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{y})$ takes a small value.

A novelty of the aHSIC score is that it is possible to measure dependency based solely on features that are related to an output $\boldsymbol{y}$ if we set $\boldsymbol{\alpha}$ appropriately. That is, if we can select features that are responsible for abrupt changes, the aHSIC score is independent of noise features.

3.2 HSIC Lasso

In aHSIC, the choice of $\boldsymbol{\alpha}$ parameter is a key issue. A simple heuristic is to set $\frac{1}{n}$ for all α. However, for high-dimensional time series data, a few features are important and rest are noise. That is, equal weighting of the HSIC scores is not a suitable choice for high-dimensional time-series.

In this paper, we propose using HSIC Lasso for estimating the $\boldsymbol{\alpha}$ parameter [Yamada *et al.*, 2012]:

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \quad & \|\bar{\boldsymbol{L}} - \sum_{k=1}^{d} \alpha_k \bar{\boldsymbol{K}}^{(k)}\|^2_{\mathrm{Frob}} + \lambda\|\boldsymbol{\alpha}\|_1, \\ \text{s.t.} \quad & \alpha_1, \ldots, \alpha_d \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|_{\mathrm{Frob}}$ is the Frobenius norm and $\|\cdot\|_1$ is the ℓ_1 norm.

The first term in Eq.(1) means that we are regressing the output kernel matrix $\bar{\boldsymbol{L}}$ by a linear combination of feature-wise input kernel matrices $\{\bar{\boldsymbol{K}}^{(k)}\}_{k=1}^d$. After estimating $\boldsymbol{\alpha}$, we normalize each element of $\boldsymbol{\alpha}$ as $\alpha_k \leftarrow \alpha_k / \sum_{k=1}^d \alpha_k$.

The first term in Eq.(1) can be rewritten as

$$\begin{aligned} \|\bar{\boldsymbol{L}} - \sum_{k=1}^{d} \alpha_k \bar{\boldsymbol{K}}^{(k)}\|^2_{\mathrm{Frob}} = \mathrm{HSIC}(\boldsymbol{y}, \boldsymbol{y}) - 2\sum_{k=1}^{d} \alpha_k \mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{y}) \\ + \sum_{k,l=1}^{d} \alpha_k \alpha_l \mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{u}_l). \end{aligned} \tag{2}$$

Thus, if all features $\boldsymbol{u}_k$ are mutually independent (i.e., $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{u}_l) = 0, \forall\, k, l$), we can rewrite Eq.(2) as

$$\|\bar{\boldsymbol{L}} - \sum_{k=1}^{d} \alpha_k \bar{\boldsymbol{K}}^{(k)}\|^2_{\mathrm{Frob}} \propto -\mathrm{aHSIC}(\mathcal{Z}).$$

This means that, minimizing the objective function of HSIC Lasso corresponds to maximizing the $\mathrm{aHSIC}(\mathcal{Z})$ score. Thus, using HSIC Lasso to estimate $\boldsymbol{\alpha}$ is a natural choice.

Statistical Interpretation of HSIC Lasso: From Eq.(2), if the k-th feature $\boldsymbol{u}_k$ has high dependence on output $\boldsymbol{y}$, $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{y})$ takes a large value and thus α_k should also be large. On the other hand, if $\boldsymbol{u}_k$ and $\boldsymbol{y}$ are independent, $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{y})$ is close to zero and thus such α_k tends to be removed by the ℓ_1-regularizer. That is, relevant features that have strong dependence on output $\boldsymbol{y}$ tends to be selected by the HSIC Lasso. That is, features that are important for an abrupt change are selected.

Furthermore, if $\boldsymbol{u}_k$ and $\boldsymbol{u}_l$ are strongly dependent (i.e., redundant features), $\mathrm{HSIC}(\boldsymbol{u}_k, \boldsymbol{u}_l)$ is large and thus either α_k or α_l tends to be zero. Thus, redundant features tend to be removed by HSIC Lasso.

Overall, HSIC Lasso tends to find non-redundant features with strong dependence on output $\boldsymbol{y}$, which is a preferable property for a change-point detection measure.

Computational Property: An important computational property of HSIC Lasso is that the first term in Eq.(1) can be rewritten as

$$\frac{1}{2}\|\mathrm{vec}(\bar{\boldsymbol{L}}) - [\mathrm{vec}(\bar{\boldsymbol{K}}^{(1)}), \ldots, \mathrm{vec}(\bar{\boldsymbol{K}}^{(d)})]\boldsymbol{\alpha}\|_2^2,$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator. This is the same form as plain Lasso with n^2 and d are the numbers of samples and optimization parameters, respectively.

To solve this Lasso problem, a technique called the *dual augmented Lagrangian* (DAL) was shown to be computationally highly efficient [Tomioka *et al.*, 2011]. Because DAL can also incorporate the non-negativity constraint without losing its computational advantages, we can directly use DAL to solve our HSIC Lasso problem.

A MATLAB® implementation of the HSIC Lasso is available from http://www.kecl.ntt.co.jp/icl/ls/members/myamada/hsiclasso.html.

3.3 Group HSIC Lasso

If we have a prior knowledge of the features, we can utilize this knowledge to select features in HSIC Lasso. For example when detecting a change in human activity from sensors attached to the hands and legs, it is reasonable to select a group of features (i.e., sensors) that are important in relation to an abrupt change.

To this end, we propose using the group-type lasso regularizer [Meier *et al.*, 2008; Zou and Hastie, 2005] for estimating $\boldsymbol{\alpha}$ as

$$\begin{aligned} \min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \quad & \frac{1}{2}\|\bar{\boldsymbol{L}} - \sum_{k=1}^{d} \alpha_k \bar{\boldsymbol{K}}^{(k)}\|^2_{\mathrm{Frob}} + \lambda \sum_{g=1}^{G} \|\boldsymbol{\alpha}_g\|_2, \\ \text{s.t.} \quad & \alpha_1, \ldots, \alpha_d \geq 0, \end{aligned}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_G^\top]^\top$, $\boldsymbol{\alpha}_g$ is the gth group of variables, and G is the number of groups. This group-type lasso problem can also be efficiently solved by the DAL package with the non-negativity constraint.

3.4 Kernel Selection

HSIC is a kernel based independence measure; the independence criterion changes with respect to a kernel parameter and/or kernel types. That is, the input kernel parameters should be fixed for all features. Thus, we first normalize the input features with a standard deviation at 1 and then use the same kernel parameters for all kernels.

For input x, we use the Gaussian kernel,

$$K(x, x') = \exp\left(-\frac{(x-x')^2}{2\sigma_{\mathrm{x}}^2}\right),$$

where we set $\sigma_{\mathrm{x}} = 1$. For output y, we use the delta kernel,

$$L(y, y') = \begin{cases} 1 & \text{if } y = y', \\ 0 & \text{otherwise.} \end{cases}$$

4 Related Methods

Here, we review related change-point detection measures.

4.1 KLIEP

Let us assume that samples in $\mathcal{X}(t)$ are drawn i.i.d. from a distribution with a density $p(\boldsymbol{x})$ and samples in $\mathcal{X}(t+n)$ are drawn i.i.d. from a distribution with density $p'(\boldsymbol{x})$. Then, we can use *divergence* as the plausibility of the change-points. More specifically, if the divergence $D(p(\boldsymbol{x})||p'(\boldsymbol{x}))$ is larger than the threshold τ, we can regard the point as the change-point.

A popular choice for the divergence function is *Kullback-Leibler (KL) divergence*:

$$KL[p(\boldsymbol{x})||p'(\boldsymbol{x})] = -\int p'(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}.$$

A naive approach for estimating the KL-divergence is to estimate the probability densities $p(\boldsymbol{x})$ and $p'(\boldsymbol{x})$ separately using a kernel density estimation [Härdle *et al.*, 2004] and then take the ratio. However, density estimation is known to be a hard problem, and the KL-divergence estimation can be poor. To mitigate this, an alternative KL divergence estimation method is proposed, where the density-ratio $\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$ is directly estimated without going through the density estimations by using the Kullback-Leibler Importance Estimation Procedure (KLIEP), which has been proved to achieve the optimal non-parametric convergence rate in a mini-max sense [Sugiyama *et al.*, 2008].

In [Kawahara and Sugiyama, 2009], a KLIEP based KL-divergence estimator was used for change-point detection, and it outperformed existing density estimation based methods. However, since KLIEP uses all the features to compute the divergence, the KL-divergence estimator tends to perform poorly when there are many noise features.

4.2 uLSIF/RuLSIF

Recently, computationally efficient density-ratio estimation method called relative unconstrained Least-Squares Importance Fitting (RuLSIF) has been proposed for estimating *relative Pearson-divergence*[Yamada *et al.*, 2011]:

$$\begin{aligned} PE_\alpha[p(\boldsymbol{x})||q_\alpha(\boldsymbol{x})] &:= PE[p(\boldsymbol{x})||\alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})] \\ &= \int \left(\frac{p(\boldsymbol{x})}{q_\alpha(\boldsymbol{x})} - 1\right)^2 q_\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \end{aligned}$$

where $q_\alpha(\boldsymbol{x}) = \alpha p(\boldsymbol{x}) + (1-\alpha)p'(\boldsymbol{x})$ is called *α-mixed density* and r_α is the *relative density ratio*:

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

The novelty of the relative density-ratio is that it is always bounded above by $\frac{1}{\alpha}$, and it has been shown that the convergence rate for estimating relative density ratio is faster than that of the standard density-ratio $\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$ [Yamada *et al.*, 2011]. This is a favorable property in practice. Note, when $\alpha = 0$, it is reduced to $\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})}$ (uLSIF) [Kanamori *et al.*, 2009].

It has been shown that the relative PE divergence estimator is suited to change-point detection [Liu *et al.*, 2013]. However, as with the KL-divergence estimator, estimation of relative PE divergence tends to be poor when there are many noise features and a small number of training samples.

4.3 Kernel Change Detection

Kernel change detection (KCD) is also a detection method that does not require density estimations [Desobry *et al.*, 2005].

Let $\boldsymbol{\beta}_{t,1}$ and $\boldsymbol{\beta}_{t,2}$ be coefficients of a one-class support vector machine (OSVM) computed from $\mathcal{X}(t)$ and $\mathcal{X}(t+1)$ respectively, $\boldsymbol{K}_{t,12}$ is the Gram matrix computed from support vectors obtained from $\mathcal{X}(t)$ and $\mathcal{X}(t+1)$, $\boldsymbol{K}_{t,11}$ is the Gram matrix computed from support vectors obtained from $\mathcal{X}(t)$, and $\boldsymbol{K}_{t,22}$ is the Gram matrix computed from support vectors obtained from $\mathcal{X}(t+1)$. Then, the dissimilarity measure used in KCD is given as

$$D(\mathcal{X}(t), \mathcal{X}(t+1)) = \frac{\boldsymbol{\beta}_{t,1}^\top \boldsymbol{K}_{t,12} \boldsymbol{\beta}_{t,2}}{\sqrt{\boldsymbol{\beta}_{t,1}^\top \boldsymbol{K}_{t,11} \boldsymbol{\beta}_{t,1}} \sqrt{\boldsymbol{\beta}_{t,2}^\top \boldsymbol{K}_{t,22} \boldsymbol{\beta}_{t,2}}}.$$

This dissimilarity measure is shown to be asymptotically equivalent to the Fisher ratio in the Gaussian case.

KCD is robust to outliers, since outliers are removed by OSVM. However, since KCD uses all the features for detecting change-points, it can perform poorly when there are many noisy features.

4.4 Change-Point Detection using SSA

Change-point detection using stationary subspace analysis (SSA) is a promising method for multivariate time-series data [Blythe *et al.*, 2012].

Let us divide the entire time series $\mathcal{X}$ into $\mathcal{X}_1, \ldots, \mathcal{X}_N$ sets and define the hypothesis of testing non-stationarity as

$$\begin{aligned} H_0 &: \mathcal{X}_1, \ldots, \mathcal{X}_N \sim N(\boldsymbol{0}, \boldsymbol{I}) \\ H_1 &: \mathcal{X}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, \mathcal{X}_N \sim N(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N). \end{aligned}$$

where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

Then, the goal of the SSA based change-point detection is to test whether time-series data are stationary. More specifically, if we can reject the null hypothesis H_0, we can consider that there is a change-point. On the other hand, if we cannot reject the null hypothesis H_0, we consider there to be no change-point.

The test statistics based on SSA can be written as

$$\Lambda(\mathcal{X}) = -\frac{d_s}{2}N + \frac{1}{2}\sum_{i=1}^{N} N_i \left(-\log\det \widehat{\boldsymbol{\Sigma}}_i^{\widehat{\boldsymbol{B}}} + (\widehat{\boldsymbol{\mu}}_i^{\widehat{\boldsymbol{B}}})^\top \widehat{\boldsymbol{\mu}}_i^{\widehat{\boldsymbol{B}}}\right)$$

where $\widehat{\boldsymbol{\mu}}_i^{\widehat{\boldsymbol{B}}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\boldsymbol{\Sigma}}_i^{\widehat{\boldsymbol{B}}} = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{\Sigma}}_i\widehat{\boldsymbol{B}}^\top$, $\widehat{\boldsymbol{\mu}}_i = \frac{1}{|\mathcal{X}_i|}\sum_{t\in\mathcal{X}_i}\boldsymbol{x}(t)$, $\widehat{\boldsymbol{\Sigma}}_i = \sum_{t\in\mathcal{X}_i}(\boldsymbol{x}(t)-\widehat{\boldsymbol{\mu}}_i)(\boldsymbol{x}(t)-\widehat{\boldsymbol{\mu}}_i)^\top$, and $\widehat{\boldsymbol{B}} \in \mathbb{R}^{m\times d}$ $(m < d)$ is given as the solution of the following optimization problem:

$$\widehat{\boldsymbol{B}} = \underset{\boldsymbol{B}\boldsymbol{B}^\top=\boldsymbol{I}}{\operatorname{argmax}} \sum_{i=1}^{N}\left(-\log\det \widehat{\Sigma}_i^{\boldsymbol{B}} + (\widehat{\boldsymbol{\mu}}_i^{\boldsymbol{B}})^\top \widehat{\boldsymbol{\mu}}_i^{\boldsymbol{B}}\right).$$

It has been shown that SSA-based change-point detection performs well on multivariate time-series data. However, the SSA based change-point detection algorithm needs a large number of training samples (i.e., $n \gg d$), since it includes the log of the covariance matrix which is singular when $(d < n)$. In addition, SSA needs several non-overlapped time-series. Thus, the SSA-based change-point detection algorithm is not applicable to high-dimensional change-point detection problems.

5 Experiments

In this section, we investigate experimentally the performance of the proposed and existing feature selection methods using synthetic and real-world human activity datasets.

5.1 Setup

We compare the performance of the proposed methods with that of RuLSIF, KLIEP, and KCD. With RuLSIF and KLIEP, we use publicly available codes. For KCD, we use LIBSVM [Chang and Lin, 2011] to compute OSVM, where we use a Gaussian kernel with $\sigma = \sqrt{d/2}$ and the regularization parameter of OSVM $\nu = 0.5$. We experimentally fix λ at 0.01 for HSIC Lasso and λ at 10.0 for Group HSIC Lasso. For all the methods, we fix the window size at $2n = 40$.

In this paper, we compare the performance of change-point detection methods objectively in terms of the *receiver operating characteristic (ROC) curves* and the area under the ROC curve (AUC) values. Note that, detection at t is regarded as correct if there exists a true alarm at step t^* such that $t \in [t^* - 10, t^* + 10]$.

5.2 Synthetic Datasets

In this section, we illustrate the behavior of the proposed additive HSIC based change-point detection method using synthetic datasets. We generate time-series data so that one feature includes abrupt changes and the remaining features are noise.

We use the following two synthetic multivariate time-series datasets, which contain manually inserted change-points:

(a) Data1 (Jumping mean): The following 1-dimensional auto-regressive model borrowed from [Yamanishi and Takeuchi, 2002] is used to generate 1000 samples

$$x_1(t) = 0.6x_1(t-1) - 0.5x_1(t-2) + \mu_M + \epsilon_t,$$

where $\epsilon_t \sim N(0,1)$. The initial values are set as $x_1(1) = x_1(2) = 0$. A change-point is inserted at every 100 times steps by setting the noise mean μ at time t as

$$\mu_M = \begin{cases} 0 & M = 1 \\ \mu_{M-1} + 3 & M = 2,\ldots,49, \end{cases}$$

where M is a change-point index such that $100(M-1)+1 \le t \le 100M$.

Then, we generate a noise vector $(x_2(t),\ldots,x_{50}(t))^\top \sim N(\boldsymbol{0}_{49}, \boldsymbol{I}_{49})$ and concatenate it to $x_1(t)$ as $\boldsymbol{x}(t) = [x_1(t), x_2(t), ..., x_{50}(t)]^\top$.

(b) Data2 (Scaling variance): The same auto-regressive model is used as with Data1, but a change-point is inserted at every 100 time steps by setting the noise standard deviation σ at time t as

$$\sigma = \begin{cases} 1 & M = 1,3,\ldots,49 \\ 5 & M = 2,4,\ldots,48, \end{cases}$$

Then, we generate a noise vector $(x_2(t),\ldots,x_{50}(t))^\top \sim N(\boldsymbol{0}_{49}, \boldsymbol{I}_{49})$ and concatenate it to $x_1(t)$ as $\boldsymbol{x}(t) = [x_1(t), x_2(t), ..., x_{50}(t)]^\top$.

Figure 2-(a),(c),(e),(g) show part of $x_1(t)$ and the corresponding aHSIC score for each data set. The vertical dotted red lines in those figures denote the true change-points. As can be seen, the proposed approach can correctly detect change-points in both cases. Figure 2-(b),(d) show the estimated α_1 values. It is clear that the true feature is successfully selected by HSIC Lasso. Figure 2-(f),(h) show the ROC curves. The experimental results show that proposed method compares favorably with existing methods.

Table 1 shows the mean computational time of the proposed method over the Data1 data set. As can be observed, the computational time of aHSIC score is reasonable. Note that OSVM is implemented with C, while the proposed method is implemented with Matlab. Thus, by implementing the proposed method with C/C++, we can boost the computational speed of the proposed method.

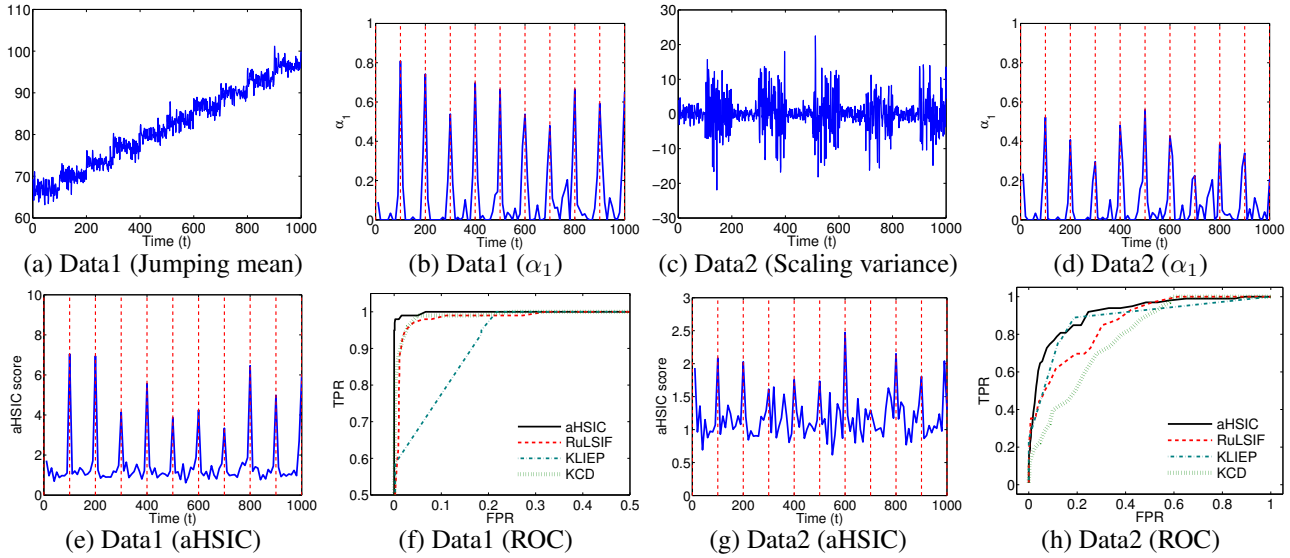
Table 1: Mean computational time of proposed method over Data1 data set.

Method	aHSIC	RuLSIF	KLIEP	KCD
Time (sec)	0.040	0.048	1.584	0.001

5.3 Real-World Human Activity Dataset

The proposed feature selection based approach is very useful for human activity change detection problems. For example when detecting the change from "Standing" to "Brushing teeth", it is reasonable to only use right/left hand information for detection.

In this section, we report the change-point detection of human activity data set [Maekawa and Watanabe, 2011]. The data set contains 14 actions from 61 subjects, where each subject wore three-axis acceleration sensors on the both hands, the waist, and the right thigh. Then, we computed the mean, energy, entropy, and main frequency component (F0) for each axis (12 features for a sensor and 48 features in total). For



(a) Data1 (Jumping mean) (b) Data1 (α_1) (c) Data2 (Scaling variance) (d) Data2 (α_1)

(e) Data1 (aHSIC) (f) Data1 (ROC) (g) Data2 (aHSIC) (h) Data2 (ROC)

Figure 2: Results for synthetic data sets. (a),(c): aHSIC-based change score in Data1 and Data2. (b),(d): Estimated α_1 from Data1 and Data2. (e),(g): aHSIC-based change score in Data1 and Data2. (f),(h): ROC curves in data2. The vertical dotted red lines in (e) and (g) denote the true change points. The AUC values of the proposed, RuLSIF, KLIEP, and KCD in Data1 are 0.999, 0.990, 0.954, and 0.994, and those of in Data2 are 0.913, 0.863 0.882, and 0.789, respectively.

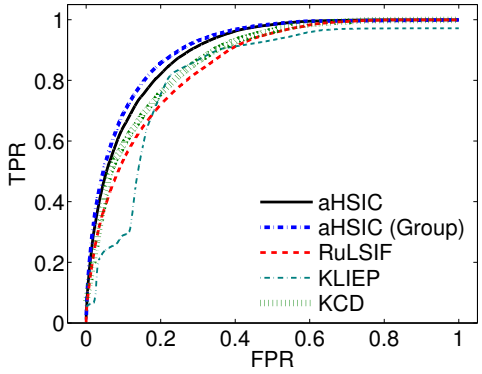


Figure 3: ROC curve of human activity data (averaged over 10 runs). The AUC values of aHSIC, aHSIC (Group), RuLSIF, KLIEP and KCD are 0.896, 0.907, 0.853, 0.820, and 0.869, respectively.

Group HSIC Lasso, we use four groups namely the right/left hands, waist, and right thigh.

In this experiment, we first randomly concatenate all the motion sequences and check whether the change-point is correctly detected. Figure 3 shows the ROC curves (averaged over 10 runs). The AUC values of the aHSIC, aHSIC (Group), RuLSIF, KLIEP, and KCD were 0.896, 0.907, 0.853, 0.820, and 0.869, respectively. Moreover, paired t-tests were conducted, and we observe that aHSIC and its group version outperform existing methods at $p = 0.01(1\%)$. The experimental results show that the proposed method compares favorably with existing methods.

6 Conclusion

In this paper, we proposed a change-point detection method with feature selection for high-dimensional time-series data. We adopted the supervised change-point detection approach [Hido *et al.*, 2008] in which we use the separability of the past and current data sets (they are labeled $+1$ and -1, respectively) as the change-point detection measure. Based on this framework, we proposed a new change-point detection measure called the *additive Hilbert-Schmidt Independence Criterion* (aHSIC), which is defined as the weighted sum of HSIC values between each feature and its corresponding pseudo binary labels. An advantage of the proposed method over existing methods is that it is estimated by using features that are important for an abrupt change. That is, the proposed approach is more robust to noisy features than existing methods. Through extensive experiments on synthetic and real-world human-activity dataset, we demonstrated the promise of the proposed change-point detection method.

Following the current line of research, there are several issues to be pursued if we are to further improve the change point detection performance. For example, in this paper, we assume that separable features are useful for detecting an abrupt change. Thus, if there is a feature that is independent of an abrupt change and has high separability, false positives of the proposed method can be increased. We have already verified experimentally that the proposed method performs well, but the issue definitely constitutes interesting future work. Moreover, investigating the performance of the proposed method over other real datasets such as music change detection is also a challenging future work.

Acknowledgments

The authors thank Dr. Katsuhiko Ishiguro and Mr. Koh Takeuchi for their valuable comments.

References

[Basseville *et al.*, 1993] M. Basseville, I.V. Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, NJ, 1993.

[Blythe *et al.*, 2012] DAJ Blythe, P. von Bunau, FC Meinecke, and K.R. Muller. Feature extraction for change-point detection using stationary subspace analysis. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(4):631–643, 2012.

[Brodsky and Darkhovsky, 1993] B.E. Brodsky and B.S. Darkhovsky. *Nonparametric methods in change point problems*, volume 243. Springer, 1993.

[Chang and Lin, 2011] C-C Chang and C-J Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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

[Gretton *et al.*, 2005] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, pages 63–77, 2005.

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

[Härdle *et al.*, 2004] W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Verlag, Heidelberg, 2004.

[Hido *et al.*, 2008] S. Hido, T. Idé, H. Kashima, H. Kubo, and H. Matsuzawa. Unsupervised change analysis using supervised learning. In *PAKDD*, pages 148–159, 2008.

[Huang *et al.*, 2007] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608. 2007.

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

[Kawahara and Sugiyama, 2009] Y. Kawahara and M. Sugiyama. Change-point detection in time-series data by direct density-ratio estimation. In *SDM*, pages 389–400, 2009.

[Ke *et al.*, 2007] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, pages 1–8, 2007.

[Kifer *et al.*, 2004] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases-Volume 30*, pages 180–191. VLDB Endowment, 2004.

[Liu *et al.*, 2013] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[Maekawa and Watanabe, 2011] T. Maekawa and S. Watanabe. Unsupervised activity recognition with user's physical characteristics data. In *15th Annual International Symposium on Wearable Computers (ISWC)*, pages 89–96. IEEE, 2011.

[Meier *et al.*, 2008] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[Murad and Pinkas, 1999] U. Murad and G. Pinkas. Unsupervised profiling for identifying superimposed fraud. *Principles of Data Mining and Knowledge Discovery*, pages 251–261, 1999.

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

[Steinwart, 2001] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[Sugiyama *et al.*, 2008] M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *NIPS*, pages 1433–1440, 2008.

[Tomioka *et al.*, 2011] R. Tomioka, T. Suzuki, and M. Sugiyama. Super-linear convergence of dual augmented Lagrangian algorithm for sparsity regularized estimation. *Journal of Machine Learning Research*, 12:1537–1586, May 2011.

[Yamada *et al.*, 2011] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *NIPS*, pages 594–602. 2011.

[Yamada *et al.*, 2012] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise non-linear lasso. *Arxiv preprint arXiv:1202.0515*, 2012.

[Yamanishi and Takeuchi, 2002] K. Yamanishi and J. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *ACM SIGKDD*, pages 676–681, 2002.

[Yamanishi *et al.*, 2000] K. Yamanishi, J.I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *ACM SIGKDD*, pages 320–324, 2000.

[Zou and Hastie, 2005] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.