Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2016, Article ID 8343187, 16 pages
http://dx.doi.org/10.1155/2016/8343187

Research Article

A Fast Framework for Abrupt Change Detection Based on Binary Search Trees and Kolmogorov Statistic

Jin-Peng Qi,[1] Jie Qi,[1] and Qing Zhang[2]

[1]*College of Information Science & Technology, Donghua University, Shanghai 201620, China*
[2]*Australia e-Health Research Centre, Csiro Computation Informatics, Brisbane, QLD 4060, Australia*

Correspondence should be addressed to Jin-Peng Qi; qipengkai@126.com

Received 28 September 2015; Accepted 28 April 2016

Academic Editor: Hiroki Tamura

Change-Point (CP) detection has attracted considerable attention in the fields of data mining and statistics; it is very meaningful to discuss how to quickly and efficiently detect abrupt change from large-scale bioelectric signals. Currently, most of the existing methods, like Kolmogorov-Smirnov (KS) statistic and so forth, are time-consuming, especially for large-scale datasets. In this paper, we propose a fast framework for abrupt change detection based on binary search trees (BSTs) and a modified KS statistic, named BSTKS (binary search trees and Kolmogorov statistic). In this method, first, two binary search trees, termed as BSTcA and BSTcD, are constructed by multilevel Haar Wavelet Transform (HWT); second, three search criteria are introduced in terms of the statistic and variance fluctuations in the diagnosed time series; last, an optimal search path is detected from the root to leaf nodes of two BSTs. The studies on both the synthetic time series samples and the real electroencephalograph (EEG) recordings indicate that the proposed BSTKS can detect abrupt change more quickly and efficiently than KS, t-statistic (t), and Singular-Spectrum Analyses (SSA) methods, with the shortest computation time, the highest hit rate, the smallest error, and the highest accuracy out of four methods. This study suggests that the proposed BSTKS is very helpful for useful information inspection on all kinds of bioelectric time series signals.

1. Introduction

Abrupt change detection is to identify abrupt changes in the statistical properties of a signal series, which occur at unknown instants [1–3]. These changes are interesting because they are indicative of qualitative transitions in the data generation mechanism (DGM) underlying the signals. Currently, CP detection has attracted considerable attention in the fields of data mining and statistics, and it has been widely studied in many real-world problems, such as atmospheric and financial analyses [1], fault detection in engineering system [4, 5], climate change detection [6], genetic time series analyses [7], signal segmentation [8, 9], and intrusion detection in computer network [4].

In community of statistics, some nonparametric approaches for CP detection have been widely explored. For example, KS statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution or between the empirical distribution function of two samples [10, 11]. Also, KS statistic and its modified versions are broadly investigated on many application fields, for example, testing hypotheses regarding activation in blood oxygenation level-dependent functional MRI data [12], modeling the cumulative distribution function of rub-induced AE signals, quantifying the goodness of fit to offer a suitable signal feature for diagnosis [13], as well as abrupt change detecting on EEG signals [14], and gene expression time series [15]. Meanwhile, as for the model-related statistic approaches, some modified cumulative sum (CUSUM) methods provide the asymptotic distributions of test statistics and the consistency of procedures and behave better in finite samples and have a higher stability with respect to the time of change than ordinary CUSUM procedures [16]. The CUSUM method and its revised versions have been widely applied to detect the structural breaks in the parameters of stochastic models, as well as the abrupt changes in the regression parameters of multiple time series regression models, such as multiple

CP detection in biological sequences [17], abrupt change detection in the regression parameters of a set of capital asset pricing data related to the Fama-French extension of the CAPM [16], and abrupt change detection in a shape-restricted regression model [18].

On the other hand, SSA is a powerful technique for time series analyses. SSA is nonparametric and requires no prior knowledge on the properties of time series signal [19]. The main idea of SSA is applied in the principal component analyses on the trajectory matrix with subsequent reconstruction of the original time series. SSA has been proved to be very successful and has already become a standard tool in the analyses of climatic [10], meteorological, and geophysical time series [11, 19]. Currently, SSA has been successfully applied in the real time series recordings, for example, abrupt change analyses on EMG-onset detection [12] and CP detection in time series [13]. Although SSA is a model-free method, it is not scalable to large-scale datasets, because it is time-consuming and sometimes invalid for time series analyses with less significant data fluctuation.

In addition, Wavelet Transform (WT) is another important tool for time series analyses [14, 15, 20–23]. WT has been widely applied in anomaly detection, time series prediction, image processing, and noise reduction [15, 23–25]. WT can represent general function at different scales and positions in a versatile and sophisticated manner, so that the data distribution features can be easily extracted from different time or space scales [25, 26]. As a simple WT, Haar Wavelet (HW) owns some attractive features including fast implementation and ability to analyze the local features. HW is very useful to find abrupt changes of discontinuity and high frequency in time series, so it is a potential candidate in modern electrical and computer engineering applications, such as signal and image compression, eye detection [27], abnormality detection on time series [28, 29], and abrupt change detection on autoregressive conditional heteroscedastic processes [30].

However, all of these methods above are time-consuming and sometime invalid for abrupt change detection near the left or the right boundary, especially for insignificant data fluctuation in large-scale time series. To resolve these problems, we propose a fast framework for CP detection based on binary search trees and a modified KS statistic, termed BSTKS for short. In this novel method, first, two BSTs are derived from a diagnosed time series. Second, three search criteria are introduced in terms of the statistic and variance fluctuations between two adjacent time series segments, and then an optimal search path is detected from the root to leaf nodes of two BSTs. Last, the proposed BSTKS and other KS, t, and SSA methods are tested on both the synthetic time series and real EEG recordings and evaluated in terms of computation time, hit rate, error, accuracy, and area under curve (AUC) of Receiver Operating Characteristic (ROC) curve analyses.

In general, for a certain bioelectric signal, an abrupt change means an important transition of biological functions or health states before and after a strong attack or an acute perturbation from internal or external environment. Therefore, it is very necessary to not only discern abrupt change from all kinds of physiological and psychological time series signals, but also inspect the significant fluctuation between adjacent time series segments with different scales. The following sections focused on not only presenting the framework of the proposed BSTKS method through theoretical foundation, simulation, and evaluation, but also discussing how it can more quickly and efficiently detect abrupt change on both synthetic and real bioelectric EEG signals than other existing KS, t, and SSA methods. The rest of this paper is organized as follows. Section 2 gives the preliminary of abrupt change by introducing the statistic and variance fluctuations between two adjacent time series segments. Section 3 implements the integrated framework of the BSTKS method in terms of three search criteria in detail. Section 4 provides some representative experiments by using the synthetic time series and real EEG recordings and then analyzes the performance of BSTKS by comparing with other KS, t, and SSA methods. Section 5 gives summary and conclusion from previous sections.

2. Preliminary

2.1. Statistic Fluctuation. KS statistic is sensitive to differences in both location and shape of the cumulative distribution functions (c.d.f) of two samples. The null distribution of KS statistic is calculated under the null hypothesis that the two samples are drawn from the same distribution or one sample is drawn from the reference distribution. To detect an abrupt change from a diagnosed time series Z, we define the statistic fluctuation between two adjacent segments within Z by means of KS statistic as follows [1, 4, 19].

Definition 1. Supposing a time series sample, $Z = \{z_1, \ldots, z_N\}$, one observes

$$Z = f\left(\frac{i}{n}\right) + X, \quad i = 1, \ldots, N, \tag{1}$$

where $X = \{x_i\}_{i=1,\ldots,N}$ is a set of the discrete and centred i.i.d random variables and f is a noisy mean signal with unknown distribution. The statistic fluctuation between two adjacent segments $Z_L = \{z_a, \ldots, z_c\}$ and $Z_R = \{z_{c+1}, \ldots, z_b\}$ is defined as

$$S_{mn}(x) = \left(\frac{mn}{m+n}\right)^{1/2} \sup_{x \in R} \left|F_m(x) - G_n(x)\right|, \tag{2}$$

in which $F_m(x)$ and $G_n(x)$ are the c.d.f of Z_L and Z_R, respectively; $m = c - a$, $n = b - c - 1$, and $m + n \leq N$. Supposing the hypothesized $F_m(x)$ and $G_n(x)$ in (2) are not available, we can derive the empirical cumulative distribution functions (e.c.d.f) of $F_m(x)$ and $G_n(x)$ from Z_L and Z_R. Then, $F_m(x)$ and $G_n(x)$ can be redefined as

$$\begin{aligned} F_m(x) &= P_m\left(Z_L \leq x\right) = \frac{1}{m}\sum_{i=a}^{c} I\left(z_i \leq x\right), \\ G_n(x) &= P_n\left(Z_R \leq x\right) = \frac{1}{n}\sum_{j=c+1}^{N} I\left(z_j \leq x\right), \end{aligned} \tag{3}$$

where $F_m(x)$ and $G_n(x)$ count the proportion of the sample points below level x.

Hypothesis 1. In order to discern an abrupt change on Z in terms of statistic fluctuation defined above, we introduce KS test for two adjacent segments Z_L and Z_R in Z as

(H_0) if $S_{mn}(z_c) \le \delta$, no abrupt change occurs in Z;

(H_1) if $S_{mn}(z_c) > \delta$, abrupt change occurs in Z,

in which $\delta \in R$ is a threshold of the statistic fluctuation within Z belonging to an identical distribution. Then, we test (H_0) against (H_1) from observations. If an abrupt change c occurs in Z, there exists a value c satisfying $S_{mn}(z_c) > \delta, z_c \in [z_1, z_N]$, and $\delta \in R$. In this hypothesis, we assume that the number, the location, and the size of the function f in (1) are unknown, and the upper bound of the statistic fluctuation δ is supposed to be known.

2.2. Variance Fluctuation. Provided the statistic fluctuation defined in (2) is insignificant enough, it is difficult to detect abrupt change near the left or the right boundary within Z, especially when sample size N gets smaller. Therefore, we need to introduce another variable to calculate the variance fluctuation between two adjacent parts within a time series sample.

Definition 2. Supposing two adjacent segments $Z_L = \{z_a, \ldots, z_c\}$ and $Z_R = \{z_{c+1}, \ldots, z_b\}$ in $Z = \{z_1, \ldots, z_N\}$, the variance fluctuation between Z_L and Z_R is defined as

$$D_{mn}(c) = \sup_{1 \le L,R \le N} \left| \frac{1}{m} \sum_{L=a}^{c} z_L - \frac{1}{n} \sum_{R=c+1}^{b} z_R \right|, \tag{4}$$

in which $m = c - a$, $n = b - c - 1$, and $m + n \le N$.

Hypothesis 2. (H_0) If $D_{mn}(c) \le \beta$, no abrupt change occurs at c in Z; (H_1) if $D_{mn}(c) > \beta$, abrupt change occurs at c in Z.

Here, $\beta \in R$ is a variance threshold of time series Z which obeys an identical distribution. If there exists a value c satisfying $D_{mn}(c) > \beta$, $z_c \in [z_1, z_N]$, then an abrupt change occurs at c in Z.

3. Method

3.1. Two BSTs' Construction. In the first part of the proposed BSTKS method, two BSTs, that is, BSTcA and BSTcD, are constructed from a time series sample Z, by using multilevel HWT. Generally, as shown in Figure 1, a discrete time series signal $Z = \{z_1, z_2, \ldots, z_N\}$ can be decomposed into the kth-level trend cA^k and k-level fluctuations, that is, $\mathrm{cD}^1, \mathrm{cD}^2, \ldots, \mathrm{cD}^k$, $k = 1, 2, \ldots, \log_2 N$. The k-level HWT is the mapping H_k defined as [13]

$$Z \xrightarrow{H_k} \left(\mathrm{cA}^k \mid \mathrm{cD}^k \mid \mathrm{cD}^{k-1} \mid \cdots \mid \mathrm{cD}^2 \mid \mathrm{cD}^1\right), \tag{5}$$

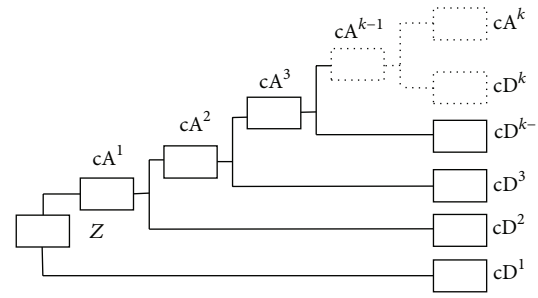


Figure 1: The diagram of a discrete time series Z decomposition by k-level HWT, which is composed of k-level cA and cD, that is, the average and difference coefficient vectors.

and then, the mapping H_k can be represented by the approximation and detail coefficient matrices, termed McA and McD as follows:

$$\mathrm{McA} = \begin{bmatrix} \mathrm{cA}_{1,1} & \cdots & \mathrm{cA}_{1,N} \\ \vdots & \mathrm{cA}_{k,j} & \vdots \\ \mathrm{cA}_{M,1} & 0 & 0 \end{bmatrix},$$
$$\mathrm{McD} = \begin{bmatrix} \mathrm{cD}_{1,1} & \cdots & \mathrm{cD}_{1,N} \\ \vdots & \mathrm{cD}_{k,j} & \vdots \\ \mathrm{cD}_{M,1} & 0 & 0 \end{bmatrix}, \tag{6}$$

where $0 \le k \le M = \log_2 N$ and $1 \le j \le N/2^k$.

Supposing the size of a diagnosed Z is divisible k times by 2, the jth element $\mathrm{cA}_{k,j}$ in cA^k and the jth element $\mathrm{cD}_{k,j}$ in cD^k can be denoted as

$$\mathrm{cA}_{k,j} = \frac{1}{\left(\sqrt{2}\right) \wedge k} \left(\sum_{i=a}^{b} z_i \right),$$
$$\mathrm{cD}_{k,j} = \frac{1}{\left(\sqrt{2}\right) \wedge k} \left(\sum_{L=a}^{c} z_L - \sum_{R=c+1}^{b} z_R \right), \tag{7}$$

where $1 \le k \le \log_2 N$ and $2^k(j-1) + 1 \le i \le j * 2^k$; $a = 2^k(j-1) + 1$, $c = 2^k(j-1) + 2^{(k-1)}$, and $b = 2^k * j$.

During two BSTs' construction, as shown in Figure 2, the non-leaf nodes in BSTcA and BSTcD are assembled by the k-level coefficient vectors of McA and McD, respectively; and then the leaf nodes are derived directly from the original time series Z. Therefore, the features of abrupt change in Z can be reflected and distributed into the different non-leaf nodes of BSTcA and BSTcD, in accordance with the k level coefficient vectors in McA and McD.

3.2. CP Detection Based on Three Search Criteria. To find an optimal path towards the potential CP within a given time series Z quickly and efficiently, some search criteria need to be introduced, and then the data exceptions can be detected from the root to leaf nodes of two BSTs. As for the statistic

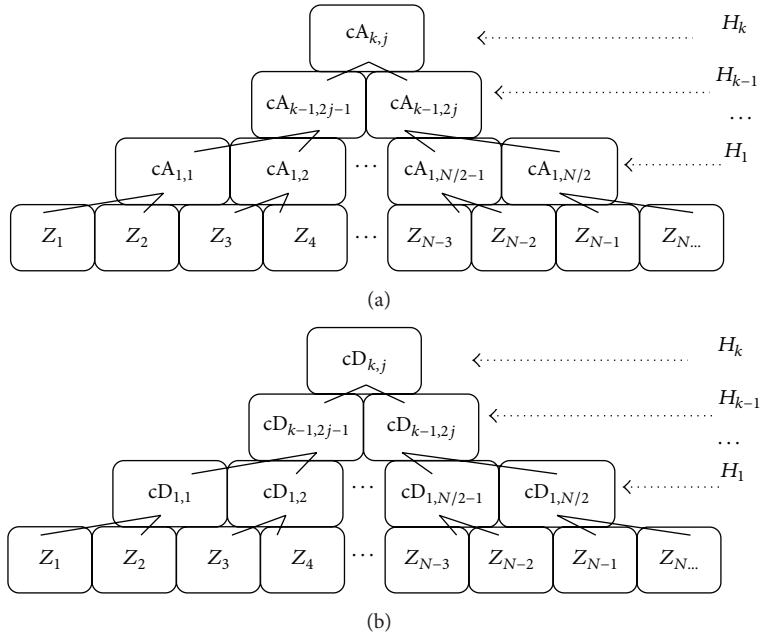


Figure 2: The diagrams of two binary trees, BSTcA and BSTcD, which are constructed by McA and McD, as well as the original time series Z.

fluctuation within BSTcA, first, a new variable $z_{k,j}$ is defined according to a current non-leaf node $\mathrm{cA}_{k,j}$ in BSTcA,

$$z_{k,j} = \frac{1}{2^k}\left(\sum_{i=a}^{b} z_i\right) = \frac{\left(\sqrt{2}\right) \wedge k}{2^k}\mathrm{cA}_{k,j}, \tag{8}$$

where $1 \le k \le \log_2 N$, $1 \le j \le N/2^k$; $a = 2^k(j-1)+1$, $b = 2^k * j$, and $a \le i \le b$. Then, the statistic fluctuation between two adjacent segments $Z_L = \{z_a, \ldots, z_c\}$ and $Z_R = \{z_{c+1}, \ldots, z_b\}$ can be defined by a modified KS statistic as

$$S'_{mn}(k,j) = \left(\frac{nm}{n+m}\right)^{1/2} \cdot \left|\left\{\frac{1}{n}\sum_{iL=a}^{c} I\left(z_{iL} \le z_{k,j}\right) - \frac{1}{m}\sum_{iR=c+1}^{b} I\left(z_{iR} \le z_{k,j}\right)\right\}\right|, \tag{9}$$

in which $z_{k,j}$ is a new element defined in (8); m and n stand for the sizes of Z_L and Z_R, respectively; $1 \le k \le \log_2 N$, $1 \le j \le N/2^k$; $a = 2^k(i-1)+1$, $b = 2^k j$, and $c = 2^k(j-1)+2^{(k-1)}$. $S'_{mn}(k,j)$ measures the e.c.d.f difference between Z_L and Z_R, and the larger $S'_{mn}(k,j)$ means the more significant statistic fluctuation between Z_L and Z_R. Therefore, a potential abrupt change might occur at c in Z with more probability.

Definition 3. For a current non-leaf node $\mathrm{cA}_{k,j}$ in BSTcA, with its left and right-child nodes $\mathrm{cA}_{k-1,2j-1}$ and $\mathrm{cA}_{k-1,2j}$, the distance of e.c.d.f, $S_{k,j;L}$, and $S_{k,j;R}$ can be defined as

$$\begin{aligned} S_{k,j;L} &= S'_{mn}(k,j;k-1,2j-1) = \left(\frac{nm}{n+m}\right)^{1/2} \cdot \left|\frac{1}{n}\left(\sum_{i=a}^{b} I\left(z_i \le z_{k,j}\right)\right) - \frac{1}{m}\sum_{iL=a}^{c} I\left(z_{iL} \le z_{k,j}\right)\right| \\ &= \left(\frac{nm}{n+m}\right)^{1/2} W\left|\frac{1}{n}\left(\sum_{i=a}^{b} I\left(z_i \le \mathrm{cA}_{k,j}\right)\right) - \frac{1}{m}\sum_{iL=a}^{c} I\left(z_{iL} \le \mathrm{cA}_{k,j}\right)\right|, \\ S_{k,j;R} &= S'_{mn}(k,j;k-1,2j) = \left(\frac{nm}{n+m}\right)^{1/2} \cdot \left|\frac{1}{n}\left(\sum_{i=a}^{b} I\left(z_i \le z_{k,j}\right)\right) - \frac{1}{m}\sum_{iR=c+1}^{b} I\left(z_{iR} \le z_{k,j}\right)\right| \\ &= \left(\frac{nm}{n+m}\right)^{1/2} W\left|\frac{1}{n}\left(\sum_{i=a}^{b} I\left(z_i \le \mathrm{cA}_{k,j}\right)\right) - \frac{1}{m}\sum_{iR=c+1}^{b} I\left(z_{iR} \le \mathrm{cA}_{k,j}\right)\right|, \end{aligned} \tag{10}$$

where $2 \le k \le \log_2 N$, $1 \le j \le N/2^k$; $a = 2^k(j-1)+1$, $b = 2^k j$, $c = 2^k(j-1)+2^{(k-1)}$; $n = 2^k$, $m = 2^{k-1}$; and $W = (\sqrt{2}) \wedge k/2^k$. To estimate an optimal path towards the potential change position within Z, without loss of generality, the first search criterion is introduced based on the statistic fluctuations $S_{k,j;L}$ and $S_{k,j;R}$.

Criterion 1. Given two statistic fluctuation variables $S_{k,j;L}$ and $S_{k,j;R}$ in accordance with two non-leaf child nodes $\mathrm{cA}_{k-1,2j-1}$ and $\mathrm{cA}_{k-1,2j}$ of the current selected node $\mathrm{cA}_{k,j}$ in BSTcA, and $2 \le k \le \log_2 N$,

(a) if $(S_{k,j;L} > S_{k,j;R}) \wedge (S_{k,j;L} > C(\alpha))$ holds true, then the left-child node $\mathrm{cA}_{k-1,2j-1}$ is selected and involved into the current search path; meanwhile, the right-child $\mathrm{cA}_{k-1,2j}$ is discarded;

(b) if $(S_{k,j;R} > S_{k,j;L}) \wedge (S_{k,j;R} > C(\alpha))$ holds true, then the right-child node $\mathrm{cA}_{k-1,2j}$ is selected and involved into the current search path; meanwhile, the left-child $\mathrm{cA}_{k-1,2j-1}$ is discarded.

Proof. For a selected non-leaf node $\mathrm{cA}_{k,j}$ in BSTcA, as shown in Figure 3, the original time series Z is divided equally into two adjacent segments Z_L and Z_R, which are covered by two non-leaf child nodes $\mathrm{cA}_{k-1,2j-1}$ and $\mathrm{cA}_{k-1,2j}$, respectively. According to the definitions of $S_{k,j;L}$ and $S_{k,j;R}$ in (10), the satisfied $S_{k,j;L} > S_{k,j;R}$ indicates that the statistic fluctuation within Z_L is more significant than that one within Z_R; that is, a potential abrupt change might be contained in Z_L with more probability than in Z_R, and vice versa. Furthermore, if $S_{k,j;L} > C(\alpha)$ holds true, then (H_1) of Hypothesis 1 is satisfied; that is, abrupt change occurs in Z_L, and vice versa, where $C(\alpha)$ is the critical value predefined in an identical distribution and α is the significance level. Therefore, one of the two child nodes $\mathrm{cA}_{k-1,2j-1}$ and $\mathrm{cA}_{k-1,2j}$ is selected and involved into the current search path; meanwhile, the remaining one is discarded. Once the statistic fluctuation is significant enough, an optimal search path can be detected by Criterion 1 from the top to the last non-leaf level in BSTcA. However, the search procedure is probably forced to cease because the statistic fluctuation is so insignificant that Criterion 1 is invalid for detecting it, especially for the left or the right boundary when sample Z is with smaller size N. Therefore, it is necessary to introduce another search criterion based on the variance fluctuations within BSTcD. □

Definition 4. For a current non-leaf node $\mathrm{cD}_{k,j}$ in BSTcD, with its left and right-child nodes $\mathrm{cD}_{k-1,2j-1}$ and $\mathrm{cD}_{k-1,2j}$, respectively, the variance fluctuations $D_{k,j;L}$ and $D_{k,j;R}$ are defined in terms of (4) as

$$D_{k,j;L} = D'_{mn}(k, j; k-1, 2j-1) = \left(\frac{nm}{n+m}\right)^{1/2} \cdot \left|\frac{1}{n}\left\{\left(\sum_{iL=a}^{c} z_{iL}\right) - \left(\sum_{iR=c+1}^{b} z_{iR}\right)\right\}\right| - \left|\frac{1}{m}\left\{\left(\sum_{La=a}^{lc} z_{La}\right) - \left(\sum_{Lb=lc+1}^{c} z_{Lb}\right)\right\}\right| = \left(\frac{nm}{n+m}\right)^{1/2} \left|\left|N'\left(\mathrm{cD}_{k,j}\right)\right| - \left|M'\left(\mathrm{cD}_{k-1,2j-1}\right)\right|\right|,$$

$$D_{k,j;R} = D'_{mn}(k, j; k-1, 2j) = \left(\frac{nm}{n+m}\right)^{1/2} \cdot \left|\frac{1}{n}\left\{\left(\sum_{iL=a}^{c} z_{iL}\right) - \left(\sum_{iR=c+1}^{b} z_{iR}\right)\right\}\right| - \left|\frac{1}{m}\left\{\left(\sum_{Ra=c+1}^{rc} z_{Ra}\right) - \left(\sum_{Rb=rc+1}^{b} z_{Rb}\right)\right\}\right| = \left(\frac{nm}{n+m}\right)^{1/2} \left|\left|N'\left(\mathrm{cD}_{k,j}\right)\right| - \left|M'\left(\mathrm{cD}_{k-1,2j}\right)\right|\right|, \quad (11)$$

where $2 \le k \le \log_2 N$, $1 \le j \le N/2^k$; $a = 2^k(j-1)+1$, $b = 2^k j$, $c = 2^k(j-1)+2^{(k-1)}$; $lc = 2^k(j-1)+2^{(k-2)}$, $rc = c+2^{(k-2)}$; $n = 2^k$, $m = 2^{k-1}$; and $N' = (\sqrt{2}) \wedge k/2^k$, $M' = (\sqrt{2}) \wedge (k-1)/2^{(k-1)}$.

Suppose Criterion 1 is invalid as $(S_{k,j;L} = S_{k,j;R}) \parallel (\max(S_{k,j;L}, S_{k,j;R}) \le C(\alpha))$ holds true; the second search criterion needs to be introduced in terms of the two variance fluctuation variables $D_{k,j;L}$ and $D_{k,j;R}$ as follows.

Criterion 2. Given two variance fluctuation variables $D_{k,j;L}$ and $D_{k,j;R}$ according to the two non-leaf child nodes $\mathrm{cD}_{k-1,2j-1}$ and $\mathrm{cD}_{k-1,2j}$ of the selected node $\mathrm{cD}_{k,j}$ in BSTcD, and $2 \le k \le \log_2 N$,

(a) if $(D_{k,j;L} > D_{k,j;R}) \wedge (D_{k,j;L} > C(\beta))$ holds true, then the left-child node $\mathrm{cA}_{k-1,2j-1}$ in BSTcA is accordingly selected and involved into the current search path; meanwhile the right one is ignored;

(b) if $(D_{k,j;L} < D_{k,j;R}) \wedge (D_{k,j;R} > C(\beta))$ holds true, then the right-child node $\mathrm{cA}_{k-1,2j}$ in BSTcA is accordingly selected and involved into the current search path; meanwhile the left one is ignored.

Proof. Similarly, as illustrated in Figure 4, the satisfied $(D_{k,j;L} > D_{k,j;R})$ in Criterion 2 means that the variance fluctuations within Z_L are stronger than that one within Z_R, in terms of the definitions of $D_{k,j;L}$ and $D_{k,j;R}$ in (11). That is, a potential abrupt change might exist in Z_L with more probability than in Z_R, and vice versa. Meanwhile, if $D_{k,j;L} > C(\beta)$ holds true, then (H_1) in Hypothesis 2 is satisfied; that is, abrupt change occurs in Z_L, and vice versa, where $C(\beta)$ is the critical value predefined in an identical

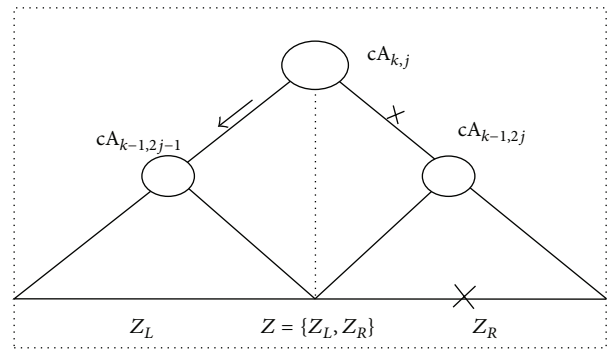


Figure 3: The scheme of Criterion 1 based on the statistic fluctuations within BSTcA. In terms of this criterion, the left or right-child node, that is, $\mathrm{cA}_{k,2j-1}$ or $\mathrm{cA}_{k,2j}$, might be selected to be involved in the current search path; meanwhile the remaining one is discarded. Thereafter, an optimal path towards the potential abrupt change in Z is expected to be obtained from BSTcA, after $\log_2 N$ binary search steps.

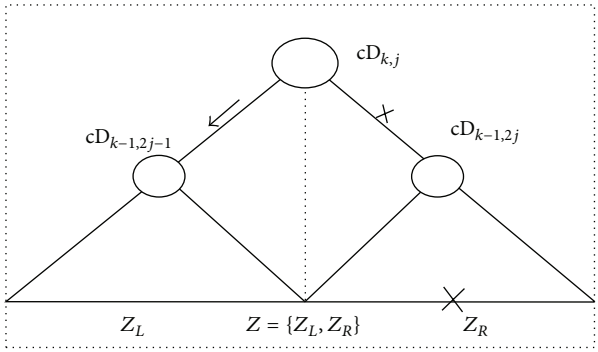


Figure 4: The scheme of Criterion 2 based on the variance fluctuations within BSTcD. Supposing Criterion 1 is invalid for insignificant statistic fluctuation within BSTcA, Criterion 2 ensures that one of the two non-leaf child nodes $cA_{k-1,2j-1}$ and $cA_{k-1,2j}$ can also be selected from BSTcA, in accordance with the variance fluctuation within BSTcD. Therefore, the search procedure can keep going forward, to the potential abrupt change in Z.

distribution and β is the significance level. As a result, one of the two non-leaf child nodes $cA_{k-1,2j-1}$ and $cA_{k-1,2j}$ in BSTcA can be accordingly selected, and the other one is neglected. Therefore, if Criterion 1 is invalid for less significant statistic fluctuation within BSTcA, Criterion 2 ensures that the search procedure can keep going forward to the potential abrupt change in Z, especially when abrupt change occurs near the left or the right boundary of Z with smaller size N. □

Based on Criterions 1 and 2 above, a search path can be obtained from the top root to the last non-leaf levels of BSTcA. In order to estimate an abrupt change from the original elements of Z, another criterion needs to be introduced to discern which one can be selected from two adjacent leaf nodes in BSTcA.

Definition 5. Supposing the current node $cA_{k,j}$ is selected in the last non-leaf level of BSTcA, $k = 1$, with two child leaf nodes z_{2j-1} and z_{2j}, two statistic fluctuation variables D_L and D_R are defined based on KS test as

$$\begin{aligned} D_L &= D_{mn}(z_L) = \left(\frac{mn}{m+n}\right)^{1/2} \left|F_m(z_L) - G_n(z_L)\right| \\ &= \left(\frac{mn}{m+n}\right)^{1/2} \\ &\quad \cdot \left| \left\{ \frac{1}{m}\sum_{i=1}^{2_{j-1}} I(z_i \le z_L) - \frac{1}{n}\sum_{h=2_j}^{N} I(z_h \le z_L) \right\} \right|, \\ D_R &= D_{mn}(z_R) = \left(\frac{mn}{m+n}\right)^{1/2} \left|F_m(z_R) - G_n(z_R)\right| \\ &= \left(\frac{mn}{m+n}\right)^{1/2} \\ &\quad \cdot \left| \left\{ \frac{1}{m}\sum_{i=1}^{2j} I(z_i \le z_R) - \frac{1}{n}\sum_{h=2j+1}^{N} I(z_h \le z_R) \right\} \right|, \end{aligned} \tag{12}$$

where $z_L = z_{2j-1}$ and $z_R = z_{2j}$; $F_m(z)$ and $G_n(z)$ refer to the e.c.d.f of $Z_L = \{z_1, \ldots, z_m\}$ and $Z_R = \{z_{m+1}, \ldots, z_N\}$, respectively; $m = 2j-1$ or $2j$ and $n = N - m$.

Consider that the largest statistic fluctuation between $F_m(x)$ and $G_n(x)$ is achieved either before or after one of the jumps, that is,

$$\begin{aligned} &\sup_{x \in \mathrm{R}} \left|G_n(x) - F_m(x)\right| \\ &= \max_{1 \le i \le n} \begin{cases} \left|F_m(z_i^-) - G_n(z_i^-)\right| & \text{before the } i\text{th jump} \\ \left|F_m(z_i) - G_n(z_i)\right| & \text{after the } i\text{th jump.} \end{cases} \end{aligned} \tag{13}$$

Then, another two variables D_L^- and D_R^- are defined as

$$\begin{aligned} D_L^- &= D_{mn}(z_L^-) = \left(\frac{mn}{m+n}\right)^{1/2} \left|F_m(z_L^-) - G_n(z_L^-)\right| \\ &= \left(\frac{mn}{m+n}\right)^{1/2} \\ &\quad \cdot \left| \left\{ \frac{1}{m}\sum_{i=1}^{2j-1} I(z_i < z_L) - \frac{1}{n}\sum_{h=2j}^{N} I(z_h < z_L) \right\} \right|, \\ D_R^- &= D_{mn}(z_R^-) = \left(\frac{mn}{m+n}\right)^{1/2} \left|F_m(z_R^-) - G_n(z_R^-)\right| \\ &= \left(\frac{mn}{m+n}\right)^{1/2} \\ &\quad \cdot \left| \left\{ \frac{1}{m}\sum_{i=1}^{2j-1} I(z_i < z_R) - \frac{1}{n}\sum_{h=2j}^{N} I(z_h < z_R) \right\} \right|. \end{aligned} \tag{14}$$

Therefore, the maximal statistic fluctuations D_L' and D_R' can be selected from D_L^- and D_L, as well as D_R^- and D_R. Then, the third search criterion is introduced in terms of D_L' and D_R' as follows.

Criterion 3. Given D_L' and D_R' in accordance with two child leaf nodes z_{2j-1} and z_{2j} of the selected non-leaf node $cA_{k,j}$ in BSTcA, $k = 1$,

(a) if $(\max(D_L', D_R') = D_L') \wedge (D_L' > C(\gamma))$ holds true, then the left leaf node z_{2j-1} in Z is taken as the estimated CP, and the right one is neglected;

(b) if $(\max(D_L', D_R') = D_R') \wedge (D_R' > C(\gamma))$ holds true, then the right leaf node z_{2j} in Z is taken as the estimated CP, and the left one is neglected;

(c) otherwise, no abrupt change is detected from Z.

Proof. Obviously, if $\max(D_L', D_R') > C(\gamma)$ is satisfied in Criterion 3, then the statistic fluctuation overtakes the critical value $C(\gamma)$ which is given in an identical data distribution, and γ is the significance level. Therefore, one of the two leaf nodes z_{2j-1} and z_{2j} is taken as the estimated CP within Z. □

Supposing a non-leaf node $cA_{k,j}$ is selected in BSTcA, the statistic and variance fluctuations are accordingly calculated

between two adjacent segments Z_L and Z_R. Meanwhile, the search procedure is implemented from the root to non-leaf nodes in the last second level of BSTcA, in terms of Criterions 1 and 2. Then, the estimated CP can be obtained from the leaf nodes in BSTcA, by using Criterion 3. Thereafter, an optimal path towards a potential CP within Z is detected from BSTcA, after about $\log_2 N$ binary search steps.

3.3. Methods Compared with BSTKS. There are many methods proposed for abrupt change detection in time series, and the following are some typical methods, to evaluate the proposed BSTKS framework.

KS Statistic (see [31]). In this method, a diagnosed time series Z is divided into two adjacent segments $Z_L = \{z_1, z_2, \ldots, z_m\}$ and $Z_R = \{z_{m+1}, z_{m+2}, \ldots, z_N\}$, and then KS statistic is applied to calculate the statistic distance between Z_L and Z_R as

$$D_{mn}(x) = \left(\frac{mn}{N}\right)^{1/2} \sup_{x \in R} \left|F_n(x) - G_m(x)\right| = \left(\frac{mn}{N}\right)^{1/2} \sup_{x \in R} \left| \sum_{R=m+1}^{n} I(z_R < x) - \sum_{L=1}^{m} I(z_L < x) \right|, \quad (15)$$

where $F_n(x)$ and $G_m(x)$ stand for the e.c.d.f of Z_L, and Z_R, respectively; $N = m + n$, N is the total length of Z, and m refers to a current test position within Z.

t-Statistic (see [32]). t also known as Welch's t-test is used only when the two population variances are assumed different (the two sample sizes may or may not be equal) and hence must be estimated separately. Suppose a diagnosed Z is divided into $Z_L = \{z_1, z_2, \ldots, z_m\}$ and $Z_R = \{z_{m+1}, z_{m+2}, \ldots, z_N\}$. Then, t-statistic is calculated as

$$t = \frac{\overline{Z}_L - \overline{Z}_R}{S_{\overline{Z}_L - \overline{Z}_R}}, \quad S_{\overline{Z}_L - \overline{Z}_R} = \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}, \quad (16)$$

where $\overline{Z}_L$ and $\overline{Z}_R$ are the sample means of Z_L and Z_R, respectively; S is an unbiased estimator of the standard deviation, $N = m + n$, and m and n are the sizes of two segments Z_L and Z_R, respectively.

SSA (see [12, 13, 33]). In SSA method, a windowed portion is chosen within a time series $Z = \{z_1, z_2, \ldots, z_N\}$, where N is large enough and a window width m and the lag parameter M are set such that $M = m/2$, $K = m - M + 1$. For each $n = 0, 1, \ldots, N - m - M$, this method takes an interval of the time series $[n + 1, n + m]$ and then defines the $M \times K$ trajectory matrix Xn and describes the structure of the windowed portion as an L-dimensional subspace. If the structure changes further, it will not be well described by the computed subspace. Then, the distance between this subspace and the new trajectory vectors will increase; therefore, this increase will signal that an abrupt change occurs in Z.

4. Results and Discussion

In this section, the proposed BSTKS is evaluated on the synthetic time series and real EEG recordings with different size N. By comparing with existing KS, t, and SSA methods, the efficiency, sensitivity, and performance are analyzed in terms of the computation time, error and accuracy, hit rate, and AUC of ROC analyses. Furthermore, the novelty of our algorithm and necessity for real application are discussed in the following paragraphs.

Table 1: The averaged results on four methods with datasets G_1 to G_7.

	Time	Hit rate	Error	Accuracy	AUC
BSTKS	*.0063*	*.4797*	*38.5268*	*.9018*	*.8922*
KS	.3537	.0841	38.7321	.8804	.8984
t	1.0068	.0168	56.3036	.8878	.7960
SSA	1.5218	.0583	41.9464	.8762	.9941

4.1. CP Detection on Synthetic Time Series. In our simulations, some typical time series samples were derived from the normally distributed datasets (mean, $u = 0$, and standard deviation, sd = 1). Each diagnosed sample of size N is composed of a normal segment of size k and an adjacent segment of size $N-k$, in which the abnormal part is simulated by adding a constant variation v into the random numbers of size $N - k$. The proposed BSTKS and other three methods, namely, KS, t, and SSA, were tested, respectively, on 200 samples which were derived from each time series group G_i with $N_i = 2 \wedge (4+i), i = 1, 2, \ldots, 7$, and $v_i = d(1+\log_2(k-4))$, where $k = \log_2(N_i)$ and $d = 1.0$. For each sample in G_i, a series of test positions were arranged by $CPK_j = j * (2 \wedge (k-4))$, $k \geq 5$, and $j = 1, 2, \ldots, 15$.

First, simulations were carried out according to different value of sample size N_i and test position CPK_j. The average analyses on four methods were listed in Table 1, and the results of simulations on datasets G_1–G_7 were illustrated in Figure 5. In general, our BSTKS is the most promising with the shortest computation time, the highest hit rate, the smallest error, and the highest accuracy out of all four methods. Particularly, as sample size N increases from N_1 to N_7, all four methods take longer time for bigger N, and BSTKS is always the fastest one. Meanwhile, BSTKS owns the highest level of hit rate against the low tracks of other three methods; and BSTKS is much more efficient with the smallest error and the highest accuracy, though all four methods tend to be better with N increasing. However, BSTKS has smaller AUC of ROC analyses, that is, bigger search space, than SSA and KS.

Second, simulations were carried out based on the datasets G_1, G_4, and G_7. The proposed BSTKS and other three methods were tested according to the different value of variance $v = d(1 + \log 2(k - 4))$, $k = 5, 8, 11$, and $d = 0.5, 1.0, 2.0, 3.0$, respectively. The average results of four methods on G_1, G_4, and G_7 were summarized in Table 2, and the typical simulations were selected on G_4 and represented in Figure 6. Generally, when v gets larger, all four methods get better hit rate, accuracy, and AUC of ROC analysis, except for longer computation time for bigger size N. Compared with other three methods, the proposed BSTKS is more encouraging because of the shortest computation time, especially when N gets bigger, as well as the highest hit rate and accuracy,

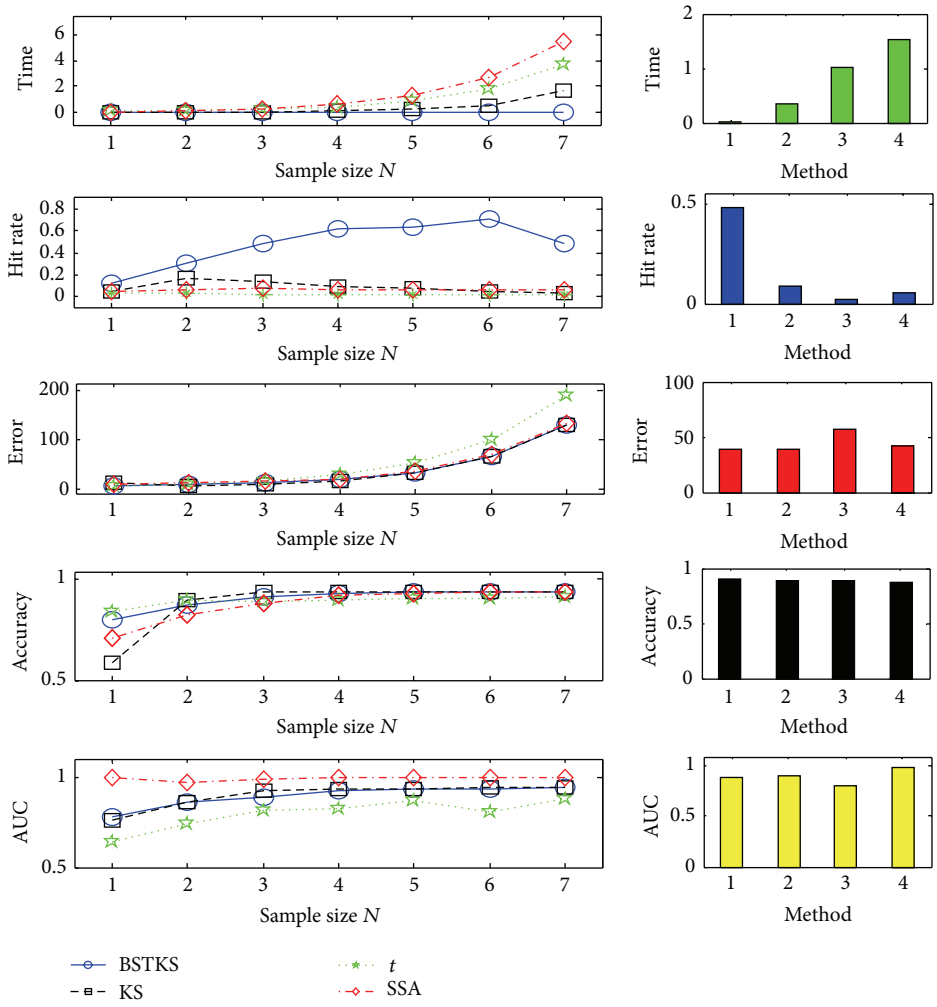


FIGURE 5: The simulations on G_1 to G_7 with size N from 2^5 to 2^{11}. (a) The results in terms of computation time, hit rate, error and accuracy, and AUC of ROC analyses. (b) The average analyses on BSTKS and other three methods. In the histograms, "1," "2," "3," and "4" stand for BSTKS, KS, t, and SSA, respectively.

especially when N gets smaller. Moreover, the simulations on G_4 with different variance v (Figure 6) explicitly illustrate that BSTKS has the best performance when v gets larger, in terms of the shortest time and the biggest increase of the hit rate out of four methods. For the accuracy and AUC, both BSTKS and KS keep higher sensitivity than t and SSA, as v increases from 0.5 to 3.0. Moreover, the simulations on G_1 and G_7 were omitted, because similar results can be obtained like G_4 above.

Third, simulations were implemented based on different CP test positions within G_1 and G_4. The proposed BSTKS and other three methods were analyzed according to the different value of test position CPK and variance v. The results of simulations on G_1 and G_4 were illustrated in Figure 7, and the results near the left and right boundaries in G_1 and G_4 were summarized in Table 3. In general, all four methods tend to be better, when N increases under a fixed v, or when v increases under a fixed N. Meanwhile, for test position CPK near the left and right boundaries, the proposed BSTKS produces better performance than other three methods, because of the highest hit rate, the smallest error in all four methods, and higher accuracy and AUC than t and SSA. Moreover, the simulations on G_1 and G_4 near the left and right boundaries were illustrated in Figure 8 in detail. In terms of the distribution of estimated CP (e-CP), PDF of e-CP, and AUC of ROC analysis, these simulations indicate that BSTKS is more sensitive for both left and right boundaries than other three methods, especially when sample size N and variance v get smaller.

Therefore, all simulation results above suggest that our proposed BSTKS is an encouraging and efficient method for abrupt change detection from the synthetic time series datasets, because of the shortest computation time, the highest hit rate, and accuracy out of four methods, especially for less significant statistic fluctuation when N gets smaller, as well as for less significant variance fluctuation when N gets bigger, and v gets smaller.

Table 2: The summary of simulations according to different variances in G_1, G_4, and G_7.

Items		Methods											
		$N = 2^5$				$N = 2^8$				$N = 2^{11}$			
		BSTKS	KS	*t*	SSA	***BSTKS***	KS	*t*	SSA	***BSTKS***	KS	*t*	SSA
$d = 0.5$	Time	***.018***	*.035*	*.227*	*.116*	***.029***	*.335*	*1.85*	*2.81*	***.060***	6.96	17.2	24.6
	Hit rate	***.046***	.005	.010	.038	***.093***	.093	.005	.025	***.106***	.056	.006	.034
	Accuracy	***.792***	.515	.792	.748	***.984***	.995	.905	.899	***.999***	.999	.944	.884
	AUC	***.694***	.694	.644	.997	***.954***	.951	.978	.971	***.983***	1.00	.999	1.00
$d = 1.0$	Time	***.018***	.034	.223	.113	***.031***	.356	1.97	2.98	***.061***	7.09	18.1	26.5
	Hit rate	***.086***	.013	.007	.035	***.041***	.001	.099	.142	***.135***	.045	.000	.035
	Accuracy	***.846***	.552	.839	.756	***.998***	.998	.939	.974	***.999***	.999	.940	.986
	AUC	***.695***	.695	.851	.997	***.993***	.992	.997	.991	***.992***	1.00	.998	1.00
$d = 2.0$	Time	***.018***	.035	.229	.115	***.031***	.345	1.91	2.88	***.065***	7.60	19.7	29.1
	Hit rate	***.165***	.061	.007	.049	***.181***	.090	.000	.049	***.167***	.028	.000	.053
	Accuracy	***.927***	.737	.958	.765	***.998***	.997	.971	.983	***.999***	.999	.984	.997
	AUC	***.754***	.754	.908	.997	***.998***	.999	.996	1.00	***.995***	1.00	.999	1.00
$d = 3.0$	Time	***.019***	.037	.245	.125	***.034***	.382	2.08	3.28	***.067***	8.08	20.5	31.4
	Hit rate	***.225***	.086	.002	.037	***.189***	.084	0.00	.046	***.169***	.035	0.00	.045
	Accuracy	***.942***	.818	.952	.773	***.997***	.996	.986	.983	***.999***	.999	.991	.998
	AUC	***.857***	.938	.655	.997	***1.00***	.996	.728	.100	***1.00***	.999	.467	1.00

Table 3: The summary of simulations on G_1 and G_4 near the left and right boundaries according to different variance d.

Items		Methods															
		$N = 2^5, CPK = 8$				$N = 2^5, CPK = 24$				$N = 2^8, CPK = 16$				$N = 2^8, CPK = 240$			
		BSTKS	KS	*t*	SSA	***BSTKS***	KS	*t*	SSA	***BSTKS***	KS	*t*	SSA	***BSTKS***	KS	*t*	SSA
$d = 0.5$	Hit rate	***.200***	*.055*	*.036*	*.160*	***.215***	*.060*	*.010*	*0.0*	***.230***	.160	0.0	0.0	***.265***	.165	0.0	.065
	Error	***2***	20	8	3	***3***	7	8	12	***24***	1	134	49	***28***	21	102	33
	Accuracy	***.937***	.375	.750	.906	***.906***	.781	.750	.625	***.960***	.996	.476	.808	***.891***	.918	.601	.871
	AUC	***.635***	.963	.641	.978	***.656***	.987	.599	.978	***.599***	.946	.780	.797	***.750***	.884	.564	.797
$d = 1.0$	Hit rate	***.515***	.160	.040	.295	***.485***	.190	.005	0.0	***.490***	.195	0.0	0.0	***.535***	.175	.025	.100
	Error	***0***	10	4	2	***0***	3	8	12	***1***	0	112	10	***0***	1	90	1
	Accuracy	***1.0***	.687	.875	.937	***1.0***	.906	.750	.625	***.996***	1.0	.562	.960	***1.00***	.996	.648	.996
	AUC	***.831***	.883	.641	.978	***.922***	.927	.599	.978	***.864***	.986	.780	.829	***.999***	.988	.657	.911
$d = 2.0$	Hit rate	***.655***	.240	0.0	.220	***.725***	.175	0.0	0.0	***.510***	.160	0.0	0.0	***.530***	.150	.005	.065
	Error	***0***	6	1	2	***0***	1	4	12	***0***	0	107	5	***0***	1	36	4
	Accuracy	***1.00***	.812	.968	.937	***1.0***	.968	.875	.625	***1.0***	1.0	.582	.980	***1.0***	.996	.859	.984
	AUC	***.976***	.979	.770	.978	***.938***	.875	.808	.978	***.978***	.985	.780	.999	***1.0***	.990	.752	.995
$d = 3.0$	Hit rate	***.715***	210	0.0	.265	***.730***	.215	0.0	0.0	***.530***	.195	0.0	0.0	***.545***	.145	0.0	.060
	Error	***0***	6	1	2	***0***	1	1	13	***0***	0	119	5	***0***	2	11	4
	Accuracy	***1.0***	.812	.968	.937	***1.0***	.968	.968	.593	***1.0***	1.0	.535	.980	***1.0***	.992	.957	.984
	AUC	***.999***	.960	.770	.978	***.996***	.822	.808	.978	***.999***	.940	.780	.999	***1.0***	.990	.752	.998

4.2. Abrupt Change Analyses on EEG Recordings. To verify the proposed method further, we take some representative samples from the CHBMIT Scalp EEG Database. In the PhysioBank platform, the CHBMIT Scalp EEG Database (CHBMIT) was collected at the Children's Hospital Boston; it consists of EEG recordings from pediatric subjects with intractable seizures [34, 35]. In this CHBMIT EEG database, some subjects were monitored up to several days after withdrawal of antiseizure medication in order to characterize their seizures and assess their candidacy for surgical intervention. Based on these EEG recordings in the CHBMIT EEG database, as well as some existing experiments in [36–39],

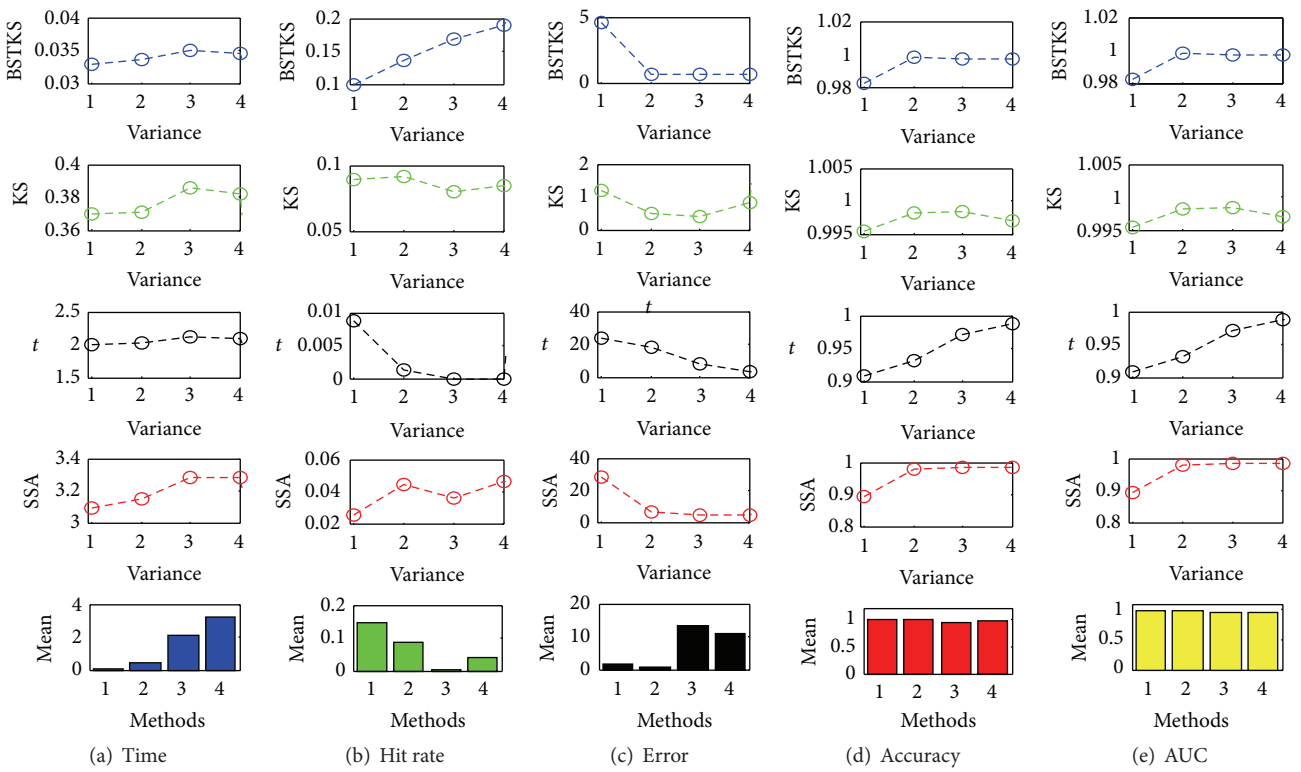


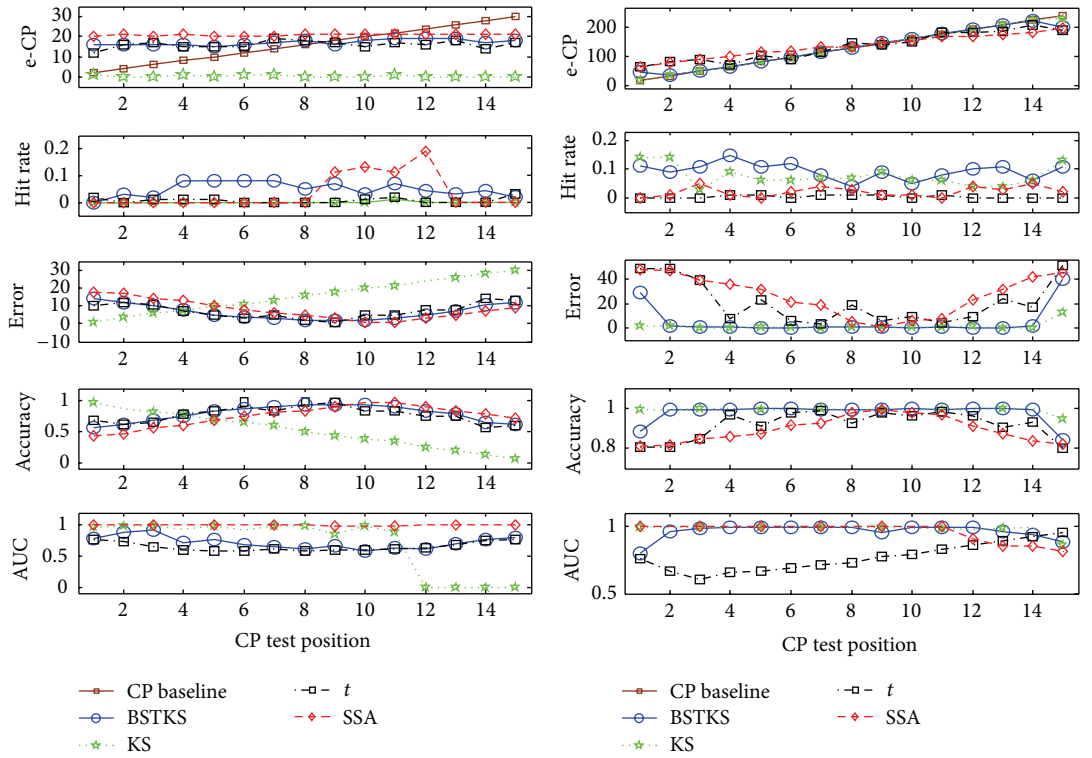
FIGURE 6: The simulations on 200 samples in G_4 with different variances. Under different variances v from 0.5 to 3.0, (a) the computation time, (b) the hit rate, (c) the error, (d) the accuracy, and (e) the AUC of ROC analysis, for BSTKS, KS, t, and SSA, respectively. In all "mean" histograms, "1," "2," "3," and "4" in x-axis stand for BSTKS, KS, t, and SSA methods, respectively.

the proposed BSTKS and other three methods were tested according to different value of test position CPK and sample size N.

First, a diagnosed EEG sample $Z = [Z_L, Z_R]$ was assembled from two significantly different segments, in which $Z_L = \{z_1, \ldots, z_{CPK}\}$ and $Z_R = \{z_{CPK+1}, \ldots, z_N\}$ were derived from chb01_04_edfm and chb01_05_edfm, respectively. Then, BSTKS and other three methods were tested on the assembled EEG recordings Z_1–Z_8, respectively, according to the different value of assigned test position CPK and sample size N. The results of abrupt change detection on these assembled EEG samples were illustrated in Figure 8 and summarized in Table 4. Generally speaking, all four methods can roughly estimate the assigned test position from each assembled EEG recording and then divide it into two adjacent segments Z_L and Z_R. It is worth stressing that the proposed BSTKS can discern the different EEG segments accurately with the smallest error and the highest accuracy out of four methods. Also, BSTKS is the most efficient and encouraging with the shortest time in all four methods.

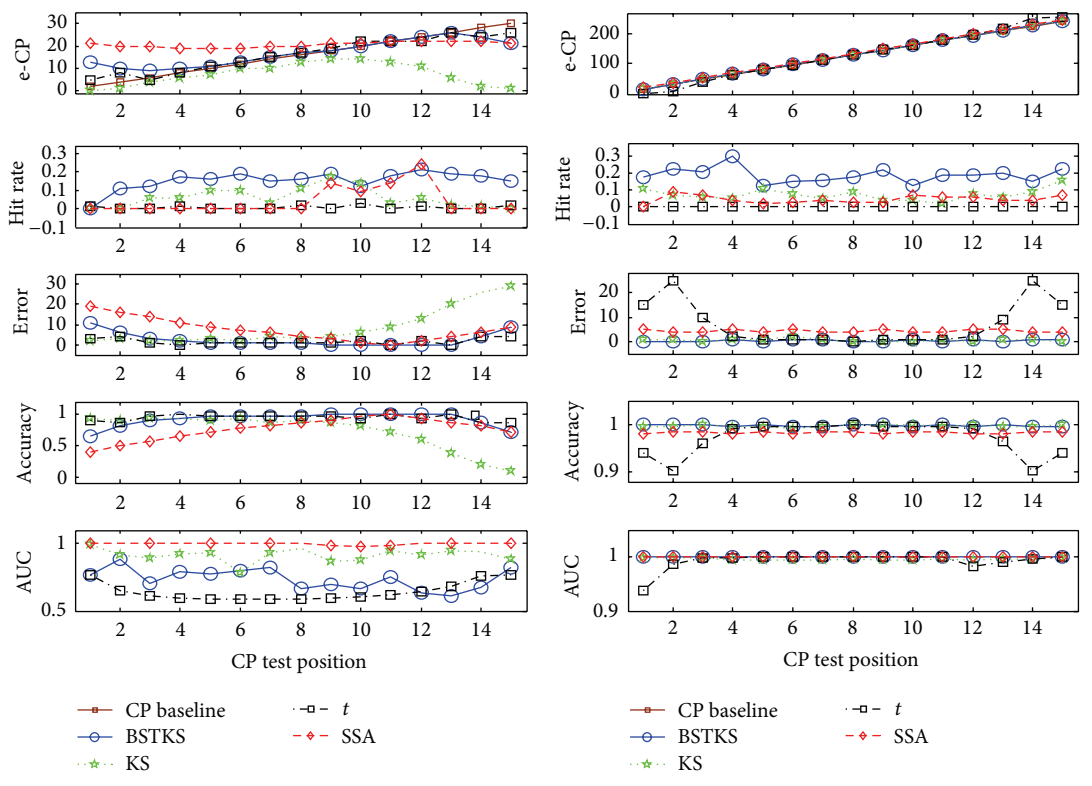
Moreover, for CPK near the left and right boundaries in Z_1–Z_8, BSTKS has much better sensitivity than other KS, t, and SSA methods because of the smallest error and the highest accuracy, especially for less statistic fluctuation when N gets smaller, as well as less significant variance fluctuation when N gets bigger. Supposing the assembled EEG sample indicates that a sharp transition of one's mental situation occurs before and after a sudden attack or acute stimulation, it is meaningful to estimate the location of the abrupt change and the maximal difference of data distribution exists between two adjacent EEG segments. These experiments above suggest that the proposed BSTKS can successfully detect the change position where a sudden change occurs under a potential mental shock, more quickly and efficiently than KS, t, and SSA methods.

Second, the original EEG samples Z_1–Z_6 were selected directly from different recordings in the chb01_05_edfm; then the proposed BSTKS and other three methods were tested according to different sample size N. Because the distance of e.c.d.f (V.e.c.d.f) can partly reflect the data fluctuation between two adjacent EEG segments, we use this V.e.c.d.f variable to distinguish different performance of BSTKS and other three methods. The results of abrupt change analyses on these original EEG recordings were shown in Figure 9 and summarized in Table 5. For all methods above, they can estimate an abrupt change from each of these original EEG samples Z_1–Z_6 and then divide it into two adjacent EEG segments. Compared with other three methods, the proposed BSTKS is encouraging for the shortest time out of four methods. Moreover, BSTKS has bigger V.e.c.d.f than t and SSA, which means that it can more reasonably distinguish two adjacent EEG segments with different state of mental health. Although KS has the biggest V.e.c.d.f in all four methods, it takes much more search time than BSTKS, especially when



(a) Results of simulation on G_1 with $N = 32$ and $v = 0.5$

(b) Results of simulation on G_4 with $N = 256$ and $v = 0.5$



(c) Results of simulation with $N = 32$ and $v = 2.0$

(d) Results of simulation on G_4 with $N = 256$ and $v = 2.0$

FIGURE 7: The simulations on G_1 and G_4 according to the different variance v and test position *CPK*. The results were shown in (a) G_1 with $N = 32$ and $v = 0.5$, (b) G_4 with $N = 256$ and $v = 0.5$, (c) G_1 with $N = 32$ and $v = 2.0$, and (d) G_4 with $N = 256$ and $v = 2.0$, in terms of e-CP, hit rate, error, accuracy, and AUC, respectively.

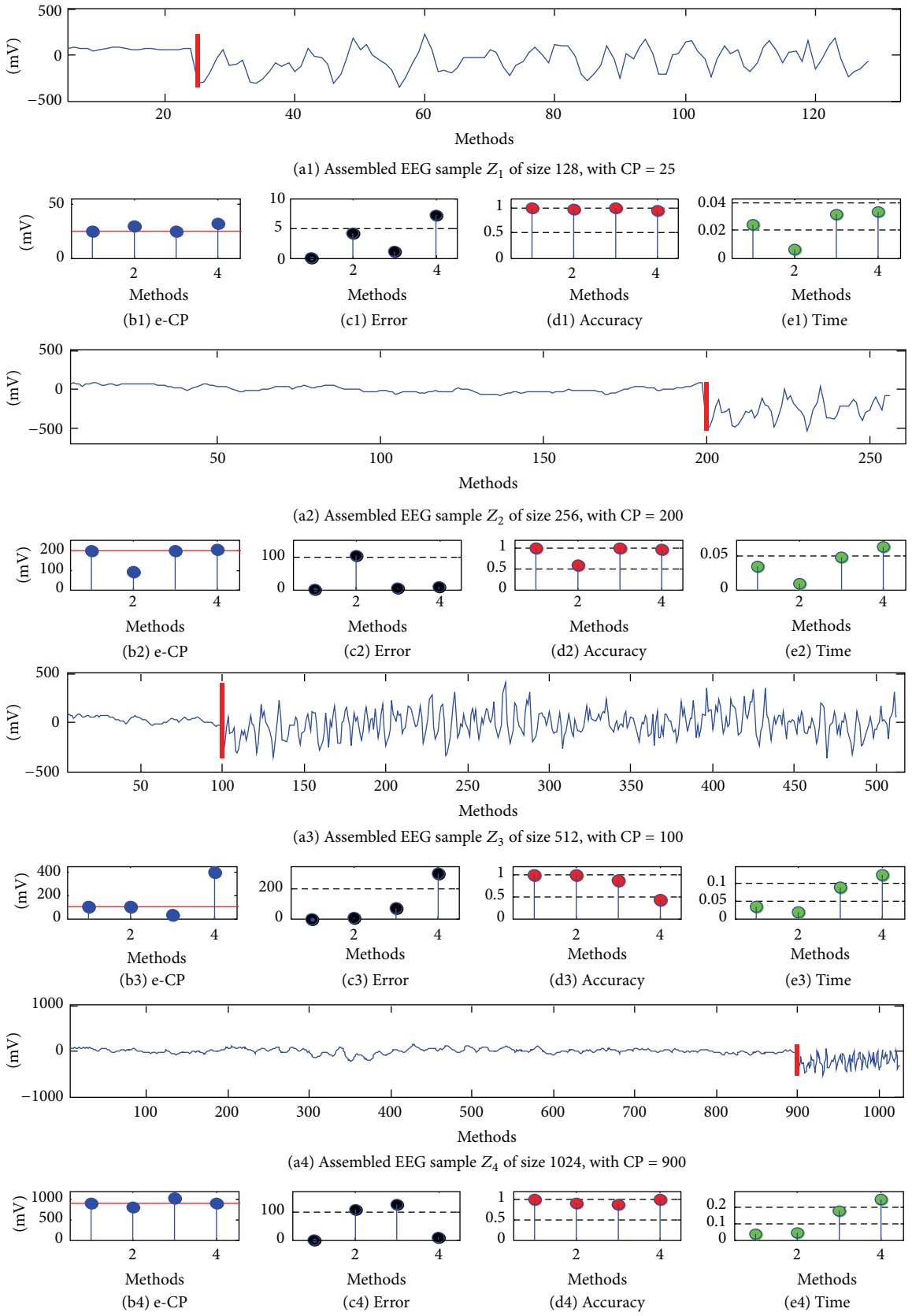


Figure 8: The results of CP detection on the assembled EEG samples Z_1–Z_4 with different value of sample size N and test position CPK. For Z_1–Z_4 with different N from 2^7 to 2^{10}, (a1–a4) the assembled EEG samples Z_1–Z_4 with the assigned test position CPK, (b1–b4) the e-CP, (c1–c4) the error of e-CP, (d1–d4) the accuracy of e-CP, and (e1–e4) the computation time. In the x-axis of (b–e), the methods "1," "2," "3," and "4" stand for BSTKS, KS, t, and SSA, respectively.

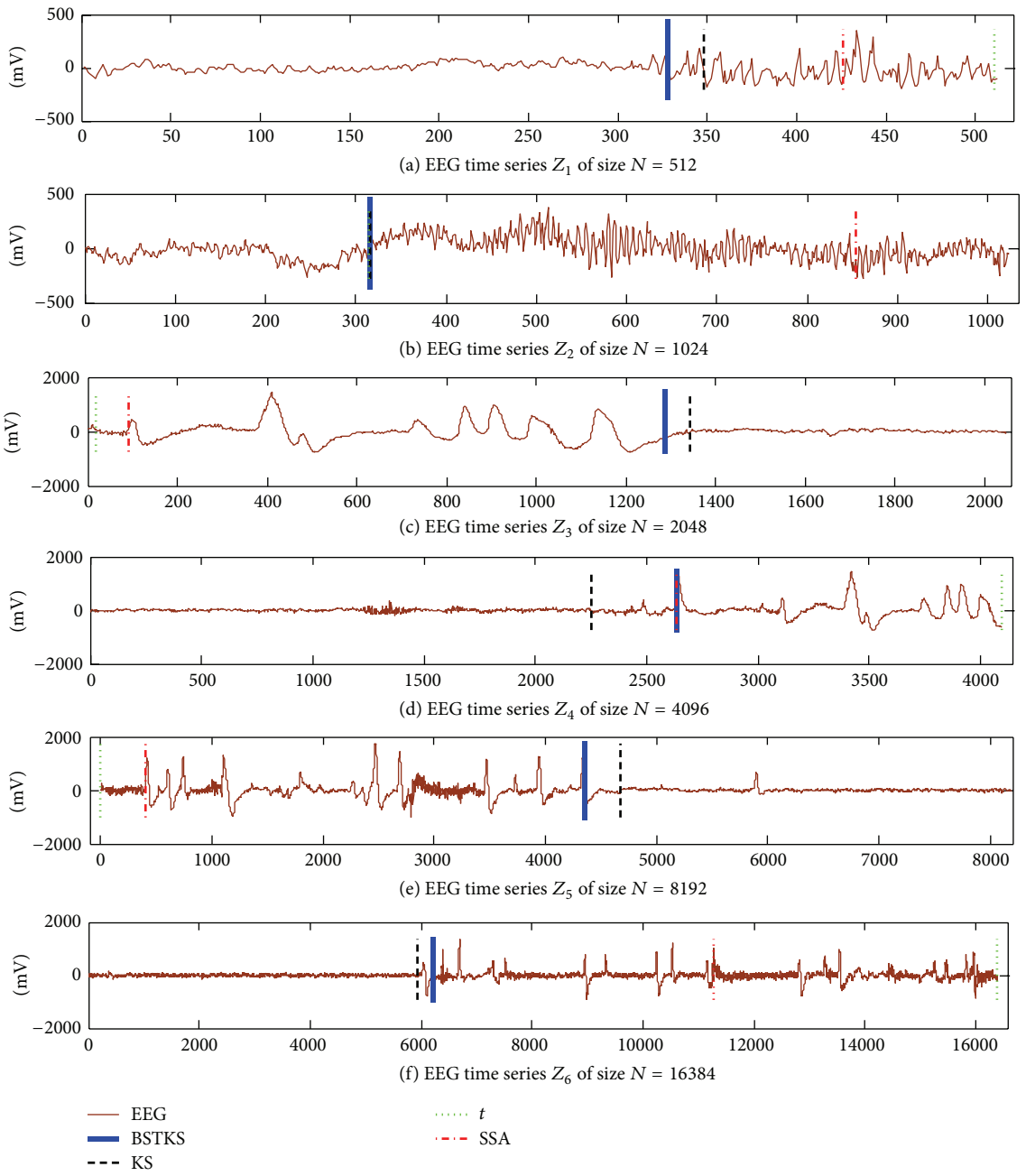


FIGURE 9: The analyses of abrupt change on the original EEG samples, by BSTKS, KS, t, and SSA, respectively. (a–f) The results of CP detection from the original EEG recordings Z_1–Z_6, with N from 2^9 to 2^{14}, respectively.

sample size N is getting larger. In addition, t needs the longest search time out of four methods, and it is invalid occasionally, for example, for Z_1, Z_5, and Z_6.

For these original EEG recordings with intractable seizures, it is of great concern to predict when and where a significant change happens from these EEG signals. This abrupt change probably indicates that a patient encounters a vertical transition from a previous mental status, and it is very important and helpful for diagnosing the patients with intractable seizures. These experiments on original EEG samples above indicate that the proposed BSTKS can not only accurately detect the change position, but also estimate the maximal difference of data distribution existing between two adjacent EEG segments, more quickly and efficiently than existing KS, t, and SSA methods.

5. Conclusion

In this paper, a novel BSTKS method is proposed based on binary search trees and a modified KS statistic. In this method, two BSTs were constructed from a diagnosed time series by multilevel HWT, and then an optimal search path

Table 4: The summary of abrupt change detection on Z_1–Z_8.

M		Z								
	N	2^7	2^8	2^9	2^{10}	2^7	2^8	2^9	2^{10}	Mean
	CPK	25	50	100	200	100	200	400	900	
e-CP	***BSTKS***	***25***	***50***	***100***	***276***	***100***	***200***	***376***	***900***	***NA***
	KS	29	36	95	206	34	92	301	795	NA
	t	24	255	31	1023	31	199	33	1023	NA
	SSA	32	55	398	1007	106	208	500	907	NA
Err	***BSTKS***	***0***	***0***	***0***	***76***	***0***	***0***	***24***	***0***	***12.5***
	KS	4	14	5	6	66	108	99	105	50.9
	t	1	205	69	823	69	1	367	123	207.3
	SSA	7	5	298	807	6	8	100	7	154.8
Acc	***BSTKS***	***1.0***	***1.0***	***1.0***	***.93***	***1.0***	***1.0***	***.95***	***1.0***	***.98***
	KS	.97	.94	.99	.99	.48	.57	.81	.89	.83
	t	.99	.20	.86	.20	.46	.99	.28	.88	.61
	SSA	.94	.97	.42	.21	.94	.97	.80	.99	.78
Time	***BSTKS***	***.023***	***.031***	***0.034***	***.036***	***.028***	***.033***	***.035***	***.038***	***.032***
	KS	.019	.021	.038	.049	.020	.029	.039	.052	*.033*
	t	.03	.063	.088	.170	.031	.050	.081	.174	*.086*
	SSA	.037	.071	.126	.239	.035	.065	.118	.245	*.117*

Table 5: The summary of CP detection from the original EEG samples Z_1–Z_6.

M		N						Mean
		2^9	2^{10}	2^{11}	2^{12}	2^{13}	2^{14}	
e-CP	***BSTKS***	***328***	***316***	***1286***	***2633***	***4352***	***6224***	***NA***
	KS	348	317	1342	2252	4673	5947	*NA*
	t	511	314	17	4095	10	16383	*NA*
	SSA	426	854	90	2634	408	11271	*NA*
V.e.c.d.f	***BSTKS***	***.0649***	***.2608***	***.2822***	***.0997***	***.1318***	***.0388***	***.1464***
	KS	.4603	.3829	.4407	.3050	.3325	.2234	*.3574*
	t	0	.1257	.5384	0	0	0	*.1106*
	SSA	.1368	.0850	.1260	.0745	.0212	.0012	*.0741*
Time	***BSTKS***	***.020***	***.020***	***.024***	***.030***	***.019***	***.0320***	***.0241***
	KS	.016	.041	.112	.466	1.461	5.638	*1.289*
	t	.072	.137	.281	.913	1.726	4.709	*1.306*
	SSA	.107	.209	.415	1.103	1.769	3.548	*1.192*

is detected from the root to leaf nodes of two BSTs in terms of three search criteria. The novelty of the proposed method is addressed by comparing with other KS, *t*, and SSA methods, and simulations on the synthetic time series indicate that the proposed BSTKS is more efficient due to the shortest time, the highest hit rate, and the smallest error and highest accuracy out of four methods. Meanwhile, BSTKS has better sensitivity than KS near the left and right boundaries, because of shorter search time, higher hit rate, and bigger AUC, especially when sample size N gets smaller with less significant statistic fluctuation. In addition, the necessity of the proposed method in the real domain is analyzed on real EEG recordings, and the results indicate that the proposed method can successfully discern an abrupt change and then obviously distinguish two adjacent EEG segments from the real EEG recordings. Through inspecting the significant fluctuation between adjacent segments signals, it is encouraging further for useful information inspection on all kinds of physiological and psychological time series signals. In a word, our BSTKS is a novel, efficient, and promising method for abrupt change analysis, and it is very helpful for useful information inspection on all kinds of real time series with different scales.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors would like to thank Professor Mohan Karunanithi in the Australia e-Health Research Centre, CSIRO Computation Informatics, for his assistance, support, and advice for this paper. This paper is supported by National Natural Science Foundation of China (no. 13K10414 and no. 61104154) and Specialized Research Fund for Natural Science Foundation of Shanghai (nos. 16ZR1401300 and 16ZR1401200).

References

[1] R. J. Bolton and D. J. Hand, "Statistical fraud detection: a review," *Statistical Science*, vol. 17, no. 3, pp. 235–249, 2002.

[2] B. E. Brodsky and B. S. Darkhovsky, *Nonparametric Methods in Change-point Problems*, vol. 243 of *Mathematics and its Applications*, Springer, Dordrecht, The Netherlands, 1993.

[3] B. E. Brodsky and B. S. Darkhovsky, *Non-Parametric Statistical Diagnosis: Problems and Methods*, vol. 509 of *Mathematics and Its Applications*, Springer, Dordrecht, The Netherlands, 2000.

[4] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne, "Online unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, no. 3, pp. 275–300, 2004.

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

[6] 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*, vol. 46, no. 6, pp. 900–915, 2007.

[7] Y. Wang, C. Wu, Z. Ji, B. Wang, and Y. Liang, "Non-parametric change-point method for differential gene expression detection," *PLoS ONE*, vol. 6, no. 5, Article ID e20060, 2011.

[8] M. E. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, New York, NY, USA, 1993.

[9] J.-P. Qi, Q. Zhang, Y. Zhu, and J. Qi, "A novel method for fast change-point detection on simulated time series and electrocardiogram data," *PLoS ONE*, vol. 9, no. 4, Article ID e93365, 2014.

[10] P. Yiou, E. Baert, and M.-F. Loutre, "Spectral analysis of climate data," *Surveys in Geophysics*, vol. 17, no. 6, pp. 619–663, 1996.

[11] R. Vautard and M. Ghil, "Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series," *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 395–424, 1989.

[12] L. Vaisman, J. Zariffa, and M. R. Popovic, "Application of singular spectrum-based change-point analysis to EMG-onset detection," *Journal of Electromyography and Kinesiology*, vol. 20, no. 4, pp. 750–760, 2010.

[13] V. Moskvina and A. Zhigljavsky, "An algorithm based on singular spectrum analysis for change-point detection," *Communications in Statistics—Simulation and Computation*, vol. 32, no. 2, pp. 319–352, 2003.

[14] F. Gustafsson, *Adaptive Filtering and Change Detection*, vol. 1, John Wiley & Sons, New York, NY, USA, 2000.

[15] V. Alarcon-Aquino and J. A. Barria, "Anomaly detection in communication networks using wavelets," *IEE Proceedings: Communications*, vol. 148, no. 6, pp. 355–362, 2001.

[16] S. Fremdt, "Page's sequential procedure for change-point detection in time series regression," *Statistics*, vol. 49, no. 1, pp. 128–155, 2015.

[17] M. Priyadarshana, T. Polushina, and G. Sofronov, "Hybrid algorithms for multiple change-point detection in biological sequences," *Advances in Experimental Medicine and Biology*, vol. 823, pp. 41–61, 2015.

[18] K. Nosek and Z. Szkutnik, "Change-point detection in a shape-restricted regression model," *Statistics*, vol. 48, no. 3, pp. 641–656, 2014.

[19] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: a toolkit for short, noisy chaotic signals," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1–4, pp. 95–126, 1992.

[20] M. Khalil and J. Duchêne, "Detection and classification of multiple events in piecewise stationary signals: comparison between autoregressive and multiscale approaches," *Signal Processing*, vol. 75, no. 3, pp. 239–251, 1999.

[21] M. Kobayashi, "Wavelets and their applications in industry," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 47, no. 3, pp. 1749–1760, 2001.

[22] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, vol. 4 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge, UK, 2006.

[23] M. Salam and D. Mohamad, "Segmentation of Malay Syllables in connected digit speech using statistical approach," *International Journal of Computer Science and Security*, vol. 2, pp. 23–33, 2008.

[24] J. Qi, Y. Ding, Y. Zhu, and Y. Wu, "Kinetic theory approach to modeling of cellular repair mechanisms under genome stress," *PLoS ONE*, vol. 6, no. 8, Article ID e22228, 2011.

[25] V. S. Tseng, C.-H. Chen, C.-H. Chen, and T.-P. Hong, "Segmentation of time series by the clustering and genetic algorithms," in *Proceedings of the 6th IEEE International Conference on Data Mining Workshops (ICDM '06)*, pp. 443–447, Hong Kong, December 2006.

[26] V. Alarcon-Aquino and J. Barria, "Change detection in time series using the maximal overlap discrete wavelet transform," *Latin American Applied Research*, vol. 39, pp. 145–152, 2009.

[27] S. Chen and C. Liu, "Eye detection using discriminatory Haar features and a new efficient SVM," *Image and Vision Computing*, vol. 33, pp. 68–77, 2015.

[28] B. S. Darkhovski, *Nonparametric Methods in Change-Point Problems: A General Approach and some Concrete Algorithms*, vol. 23 of *Lecture Notes—Monograph Series*, 1994.

[29] J. S. Walker, *A Primer on Wavelets and Their Scientific Applications*, CRC Press, New York, NY, USA, 2002.

[30] P. Fryzlewicz and S. S. Rao, "Multiple-change-point detection for auto-regressive conditional heteroscedastic processes," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 76, no. 5, pp. 903–924, 2014.

[31] R. Simard and P. L'Ecuyer, "Computing the two-sided Kolmogorov-Smirnov distribution," *Journal of Statistical Software*, vol. 39, no. 11, pp. 1–18, 2011.

[32] T. Sørlie, R. Tibshirani, J. Parker et al., "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8418–8423, 2003.

[33] H. Hassani and A. Zhigljavsky, "Singular spectrum analysis: methodology and application to economics data," *Journal of Systems Science and Complexity*, vol. 22, no. 3, pp. 372–394, 2009.

[34] A. L. Goldberger, L. A. Amaral, L. Glass et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[35] A. H. Shoeb, *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*, Massachusetts Institute of Technology, Cambridge, Mass, USA, 2009.

[36] Z. H. Xia, X. H. Wang, X. M. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 2, pp. 340–352, 2016.

[37] S. D. Xie and Y. X. Wang, "Construction of tree network with limited delivery latency in homogeneous wireless sensor networks," *Wireless Personal Communications*, vol. 78, no. 1, pp. 231–246, 2014.

[38] J. Li, X. Li, B. Yang, and X. Sun, "Segmentation-based image copy-move forgery detection scheme," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 507–518, 2015.

[39] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 7, pp. 1403–1416, 2015.