Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 8371085, 10 pages
https://doi.org/10.1155/2018/8371085

Research Article

A Fast Screen and Shape Recognition Algorithm for Multiple Change-Point Detection

Dan Zhuang [1] **and Youbo Liu** [2]

[1] *School of Statistics, Southwestern University of Finance and Economics, China*
[2] *School of Electrical Engineering and Information, Sichuan University, China*

Correspondence should be addressed to Dan Zhuang; zhdan@2014.swufe.edu.cn

Received 2 August 2018; Accepted 13 September 2018; Published 11 October 2018

Academic Editor: Erik Cuevas

A Fast Screen and Shape Recognition (FSSR) algorithm is proposed with complexity down to $O(\sqrt{n})$ for the multiple change-point detection problems. The proposed FSSR algorithm includes two steps. First, by dividing the data into several subsegments, FSSR algorithm can quickly lock some small subsegments that are likely to contain change-points. Second, through a point by point search in each selected subsegment, FSSR algorithm determines the precise location of the change-point. The simulation study shows that FSSR has obvious speed and stability advantages. Particularly, the sparser the change-points is, the better result will be achieved from FRRS. Finally, we apply FSSR to two real applications to demonstrate its feasibility and robustness. One is the problem of DNA copy number variations identifying; another is the problem of operation scenarios reduction for renewable integrated electrical distribution network.

1. Introduction

The change-point detection has been studied in various fields including environtology [1], climatology [2], agricultural economy [3], bioinformatics [4], and public economics [5]. In this paper, the basic and canonical normal model with multiple mean change-points [6–8] is considered as follows:

$$X_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right), \ i = 1, \ldots, n, \tag{1}$$

where $\mu_1 = \mu_2 = \cdots = \mu_{\tau_1} \neq \mu_{\tau_1+1} = \cdots = \mu_{\tau_2} \neq \mu_{\tau_2+1} = \cdots = \mu_{\tau_N} \neq \mu_{\tau_N+1} = \cdots = \mu_n$) is assumed and $\{\mu_1, \mu_2, \ldots, \mu_n\}^T$ is piecewise-constant vector. Besides, $\tau = (\tau_1, \ldots, \tau_N)'$ is the location vector of the change-points and N is the number of change-points. It is also assumed that $N \ll n$ and any two change-points are "not too close to each other".

A class of classical methods is to estimate the number and locations of change-points by fitting criterion, such as AIC [9] and BIC [6, 10, 11]. However, the computational complexity of these methods is very high. Braun et al. [12] and Bai and Perron [13] employed a dynamic programming algorithm to reduce the computational cost to the order of $O(n^2)$. Based on a minimum description length information criterion, Lu et al. [2] proposed an information theory approach from a nontraditional view, by using genetic algorithms tool to optimize the objective function. But it is still unfavourable for large n [14]. Nevertheless, several algorithms are available to detect multiple change-points for big data. Antoch and Jaruskova [15] focus on an effective calculation of critical for large sample, by minimizing costs over segmentation and using dynamic modelling principles, some other methods by segmenting data to speed up algorithms (such as [16, 17]) or using regularization techniques [18].

To reduce the computational complexity, some stepwise approaches are proposed. Since being proposed, LASSO [19] has become a very popular statistical approach. After a reparametrization $\theta_i = \mu_{i+1} - \mu_i$, Huang et al. [20] used LASSO-type model and the LARS algorithm to find the solution in time complexity of $O(n\log(n))$. Moreover, some LASSO-type methods were proposed to improve the adaptability and robustness (see, e.g., [21–23]).

Binary segmentation (BS) algorithm is another classical stepwise technique for multiple change-point detection by combining with a CUSUM statistic [24]. Due to its low

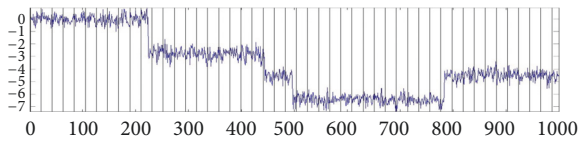

Figure 1: A time series with four change-points.

computational complexity of $O(n\log(n))$ and the fact that the execution of this algorithm is easy, BS has been widely studied and used. However, the stopping rule was difficult to compute in practice due to influence by the previously detected change-points. For example, Chen and Gupta [25] studied the problem of covariance change-point by embedding SIC into the BS procedure. In theoretical side, BS is only consistent when the minimum spacing between any two adjacent change-points is of order almost $n^{3/4}$ [8]. Circular Binary Segmentation (CBS) and Wild Binary Segmentation (WBS) were proposed to overcome the defect that BS cannot detect a small changed segment buried in the middle of large segments [8, 26, 27]. Some authors proposed many multidimensional approaches based on CUSUM statistics [28–31]. By extending the CUSUM to kernel function, Cabrieto et al. [32] detected correlation changes in multivariate systems.

Recently, by using a sliding fixed window approach, Niu and Zhang [7] proposed a very efficient (the computational complexity can even reach $O(n)$) and effective screening method known as Screening and Ranking algorithm (SaRa). Chu [33] presented two online, sliding window segmentation algorithms for single change-point detection problem. As far as we know, SaRa is the fastest algorithm at present for multiple change-points detection, because the local CUSUM statistic and forward scan algorithm are used. However, the bandwidth h is a crucial parameter for accurately identifying the change-points. To select a good bandwidth h, Niu and Zhang [7] suggested that one can try several bandwidths, respectively. The performance of SaRa will be disturbed if the bandwidth is too large or too small [34]. Xiao et al. [34] proposed a modified SaRa (mSaRa) by applying the quantile normalization and a mixture of model-based clustering. Yau and Zhao [35] also used a scan method to construct confidence intervals for multiple change-points in time series.

Although the computational complexity of SaRa is down to $O(n)$, there are still some rooms for reducing the computation cost under the assumption that $N \ll n$. In the existing algorithms, it is necessary to determine one point is change-point at least to compare all CUSUM statistics near this point. When this point becomes the maximum value and exceeds the threshold value, it can be finally confirmed as a change-point. Computation and comparison of CUSUM statistics are the main computational cost of those algorithms. However, for data series, if we can quickly lock the areas of change-points through some simple methods, a lot of calculation and comparison of CUSUM statistics will be avoided.

In this paper, we make two contributions. First, we show that our FSSR algorithm can make the computational complexity of the algorithm far less than $O(n)$. If $N \ll \sqrt{n}$, the computational complexity of FSSR can be reduced to $\sqrt{n}$. Second, in order to enhance the robustness, we embed a single-peak recognition mechanism into our algorithm. Furthermore, we also found that the proposed method has a more favorable performance when the change-points are sparser.

The paper is organized as follows. Our motivation is described in Section 2. In Section 3, the FSSR algorithm is introduced. The performances of FSSR, SaRa, and mSaRa are compared by a simulation study in Section 4. In Section 5, the proposed methodology is used in DNA copy number variations identifying and a practical engineering task involving electric power system, and we validate the effectiveness of our FSSR algorithm.

2. Motivation

Our motivation can be shown by Figure 1. Dividing the data into several small subsegments, we find that, in most small subsegments, the data is normal white noise with no change-point. The shape of data sequence is different only in a few subsegments which cover change-points. It is important to find these subsegments that contain change-points quickly. In addition, it is easy to pick out small subsegments that do not contain a change-point. Excluding these subsegments, the rest subsegments are likely to contain a real change-points.

Because two adjacent subsegments which do not contain a change-point have common mean, the difference between two CUSUM statistics of these two adjacent subsegments should be small. On the other hand, if a small subsegment covers a change-point, the difference between the CUSUM statistics of this small subsegment and the adjacent subsegment will be significant. Then we can identify subsegments with change-points through a suitable threshold. Let K be the number of subsegments. Therefore, to lock the subsegments containing change-points, we only need to calculate CUSUM statistics K times. Once we find out these small subsegments that contain change-points, we only need to search for change-points in these small subsegments.

3. Fast Screen and Shape Recognition Algorithm

In this section, we give a brief description of the FSSR.

3.1. FSSR Algorithm. First, for a given positive integer K, we split the data series $X = (X_1, X_2, \ldots, X_n)'$ into $K+1$ subsegments $Y_1, Y_2, \ldots, Y_{K+1}$ with almost equal length where $Y_i = (X_{n_i}, \ldots, X_{n_{i+1}-1})'$ and $1 = n_1 < n_2 < \cdots < n_{K+1} < n_{K+2} = n+1$. By setting $U_i = \{Y_i, Y_{i+1}\}$ $(i = 1, 2, \ldots, K)$, we get a set of subsegments $U = \{U_1, U_2, \ldots, U_K\}$.

Second, for each pair of two adjacent subsegments, the local CUSUM statistic is defined as follows.

$$D_i = \left|MU_{i+1} - MU_i\right|, \quad i = 1, 2, \ldots, K-1 \tag{2}$$

where MU_i is the mean of subsegment U_i. We select an index set $S = \{s_j\}$ $(j = 1, \ldots, T)$ by a given thresholding rule $D_{s_j} > \lambda$, where $S \subset \{1, 2, \ldots, K-1\}$, λ is usually a quantile of D_i under the assumption that there is no change-point. In the

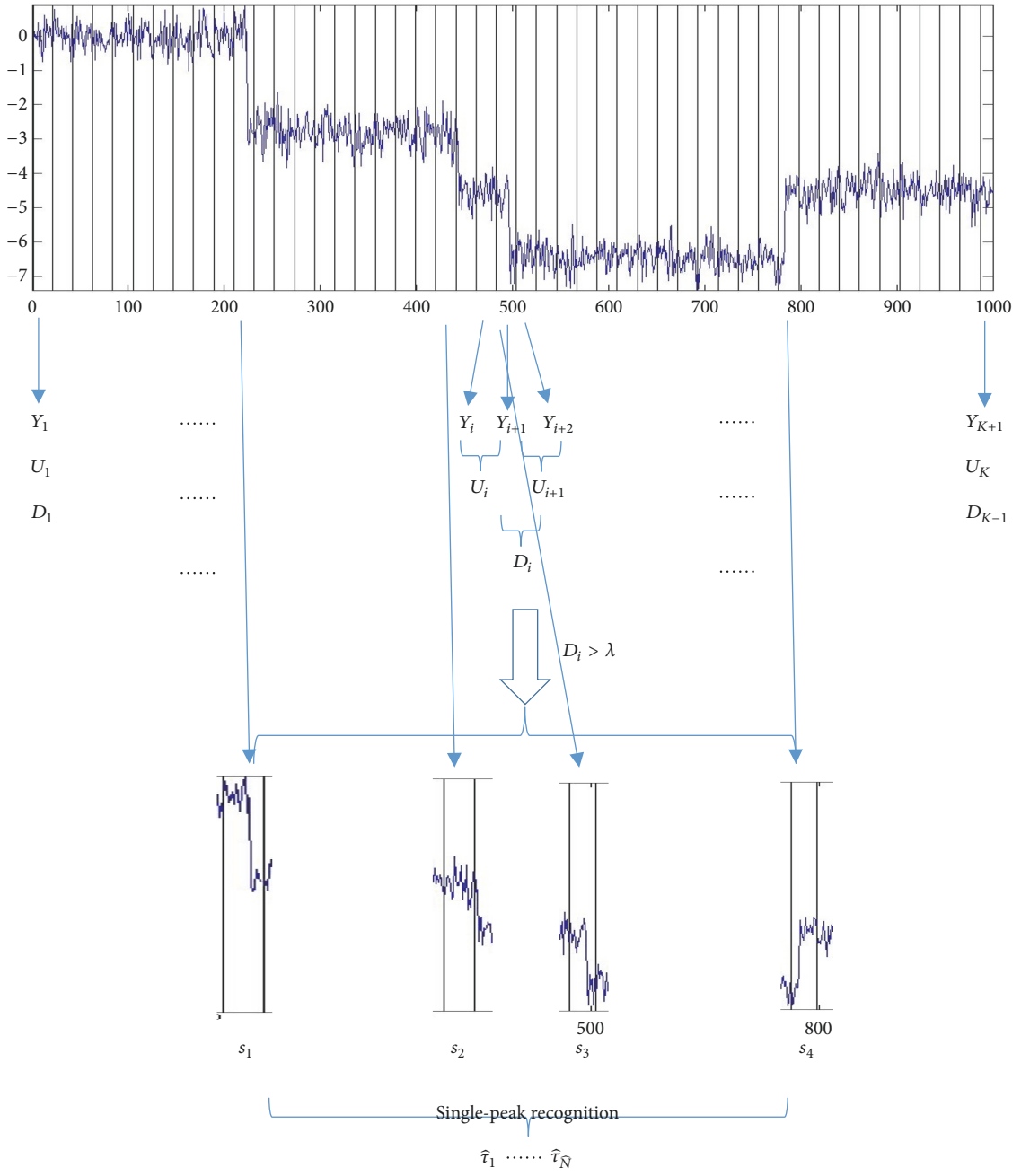

Figure 2: Flow chart of FSSR algorithm.

paper, we let $\lambda = \sqrt{K+1}U_{1-\alpha}\widehat{\sigma}/\sqrt{8n}$ where $U_{1-\alpha}$ is the upper $1-\alpha$ quantile of standard normal distribution, $\widehat{\sigma} = \sum_{i=1}^{K} \widehat{\sigma}_i/K$, and $\widehat{\sigma}_i$ is the sample standard deviation of subsegment Y_i. The selected s_j implies that change-point is likely to be covered by the subsegment Y_{s_j+1}.

Third, based on the front screen, there is no change-point in the most subsegments, then we only need to search change-points in each selected subsegment Y_{s_j+1} $(j = 1, 2, \ldots, T)$. To detect the exact location of a change-point, it is needed to search in the each selected subsegment point by point. Let $h = [n/(K+1)]$ mean that h is largest integer less than $n/(K+1)$. Let $C(x,h) = |\sum_{i=1}^{h} X_{x+1-k}/h - \sum_{i=1}^{h} X_{x+k}/h|$ be a local CUSUM statistic to detect change-points. For all points $x \in (n_{s_j+1}, n_{s_j+2})$, if $C(\widehat{\tau}_k, h)$ is the h-local maximizer, $\widehat{\tau}_k$ is an estimator of change-point τ_k. Put all h-local maximizers together; we get the final estimator for location vector of change-points $\widehat{\tau} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{N}})'$ where $\widehat{N}$ is the estimator of N after single-peak recognition.

A flow chart of FSSR algorithm is given in Figure 2.

3.2. Robustness. The good performance of CUSUM statistic is based on the normal assumption of error. In practice, the

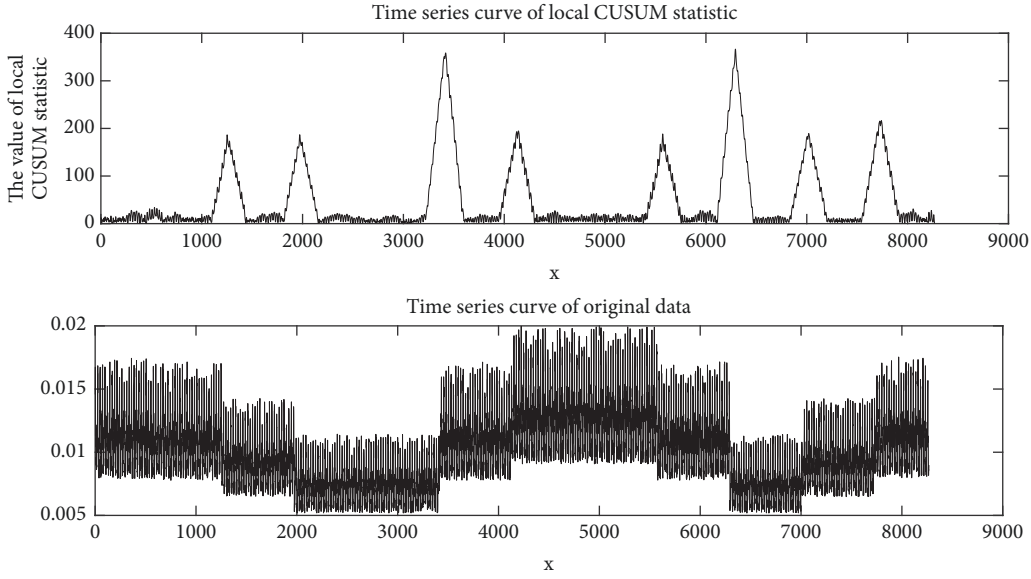

Figure 3: A sketch for local CUSUM statistic with 8 change-points.

data does not necessarily obey a normal distribution. Xiao et al. [34] used the Quantile normalization (QN) on the original intensities to seek the requirement of normality. Then, two robust processes embedded into our FSSR algorithm.

First, QN is used to make the data close to follow a normal distribution at each subsegment. In the procedure of FSSR, we rank the data in each subsegment. Then a sample with the same size as each subsegment from the standard normal distribution $N(0, 1)$ is simulated. At last, we replace the data of each subsegment by the simulated sample from $N(0, 1)$ and run our algorithm on the new data series.

Second, a single-peak recognition is used to enhance the robustness of the local maximizer. In most algorithms (such as BS, WBS, SaRa, and mSaRa), local maximum principle and threshold are used to confirm the change-point. In practice, the choice of threshold is very sensitive and has great influence on the result. From Figure 3, we can see that the local CUSUM statistic indicates a single-peak at each change-point. In this paper, to further improve the robustness of change-point detection, we define a simple single-peak principle. For any local maximum point x, let $D_{xl} = \sum_{i=1}^{h} I_{(C(x-i+1,h)>C(x-i,h))}$ and $D_{xr} = \sum_{i=1}^{h} I_{(C(x+i,h)>C(x+i+1,h))}$. If $(D_{xl} + D_{xr})/2h > \gamma$, a cut-off value, the point x is confirmed as a change-point. Obviously, the bigger γ, the stricter our rule. We find that FSSR algorithm performs well when $\gamma \in (0.7, 0.75)$ through some simulation experiments. In practice, we use $\gamma = 0.7$ in order to identify as many potential change-points as possible.

3.3. Computational Complexity. The time complexity in the FSSR is twofold. First, in the scan step, it is only needed to calculate K local CUSUM statistics. Then the computational complexity of this step is $O(K)$. In the second step, to detect the exact location of change-point, we need to calculate the local CUSUM statistic at each point of each selected subsegment. The computational complexity of this step is $O(nT/K)$. Then the computational complexity of this algorithm is $O(nT/K + K)$. If the change-points are sparse enough to satisfy $N = o(\sqrt{n})$, we can assume that $T = o(\sqrt{n})$ because T is the number of the selected subsegments and is very close to N. For example, if $T = [M\log(n)]$ (where $M > 0$ is a constant), the computational complexity reduces down to $O(\sqrt{n})$ by setting $K = \sqrt{n}$. In practice, we use $K = [\sqrt{n}]$.

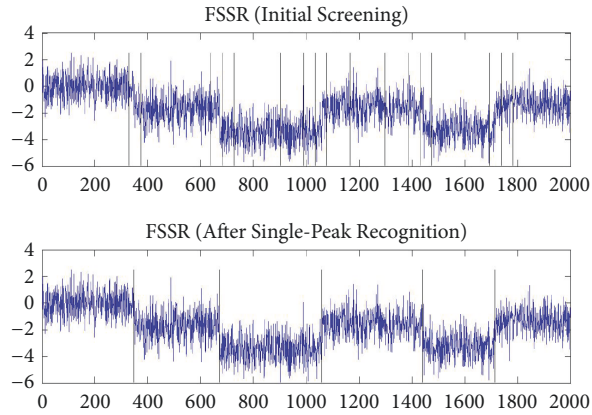

Figure 4: Process of FSSR algorithm.

4. Simulation Study

Many papers show that SaRa and mSaRa are better than those BS-type methods, such as Niu and Zhang [7], Xiao et al. [34], and Song et al. [36]. Then, in this section, the performance of FSSR against SaRa and mSaRa should be useful to examine.

4.1. An Example. Before conducting large-scale simulation experiments, we first demonstrate the implementation process and effect of our FSSR algorithm through an example. We consider an example with $n = 2000$ and $N = 5$. In Figure 4, the top plot is the initial result based on screening

Table 1: Distribution of $\widehat{N} - N$ for the various competing algorithms and sample sizes (100 simulated sample paths), with BIC values and run times (normal error case).

			$\widehat{N} - N$								
Sample Size		Method	≤ −3	-2	-1	0	1	2	≥ 3	BIC	Time
n=500	N=5	FSSR	6	20	48	23	3	0	0	0.0662×10^3	0.0113
		SaRa	0	1	5	14	22	32	26	0.1495×10^3	0.0231
		mSaRa	48	30	11	10	0	1	0	0.1644×10^3	0.3802
n=3000	N=10	FSSR	1	5	37	45	10	2	0	0.1445×10^3	0.0583
		SaRa	0	0	2	13	7	2	76	0.6343×10^3	0.1339
		mSaRa	31	8	11	9	12	8	21	0.6626×10^3	1.3432
	N=15	FSSR	20	22	29	24	5	0	0	0.2644×10^3	0.0593
		SaRa	5	3	9	15	6	8	54	0.9871×10^3	0.1101
		mSaRa	88	6	2	1	1	0	2	0.8995×10^3	1.0129
n=5000	N=10	FSSR	0	1	19	67	10	3	0	0.1357×10^3	0.0794
		SaRa	0	0	1	10	4	2	83	0.9874×10^3	0.1869
		mSaRa	11	2	5	10	7	5	60	0.8727×10^3	2.3090
	N=20	FSSR	15	22	28	27	6	1	1	0.3125×10^3	0.1280
		SaRa	7	3	4	24	6	6	50	1.5053×10^3	0.2049
		mSaRa	86	1	2	2	1	0	8	1.4307×10^3	2.0621
	N=30	FSSR	44	6	9	16	7	9	9	0.5242×10^3	0.1317
		SaRa	30	8	12	20	10	5	15	3.1588×10^3	0.1973
		mSaRa	100	0	0	0	0	0	0	1.9524×10^3	1.4564
n=8000	N=10	FSSR	0	0	25	55	17	3	0	0.1759×10^3	0.0920
		SaRa	0	0	6	9	5	6	74	1.9034×10^3	0.3024
		mSaRa	2	1	1	7	5	3	81	1.0359×10^3	2.3814
	N=20	FSSR	13	11	28	37	10	1	0	0.3645×10^3	0.2373
		SaRa	7	4	3	18	6	4	58	2.2859×10^3	0.4487
		mSaRa	64	8	7	3	2	0	16	1.8060×10^3	3.2446
	N=30	FSSR	12	4	19	11	12	7	35	0.5457×10^3	0.2480
		SaRa	17	3	11	32	5	3	29	3.9068×10^3	0.3850
		mSaRa	95	0	1	0	1	0	3	2.7051×10^3	2.5822
	N=50	FSSR	95	3	1	0	1	0	0	1.0258×10^3	0.3051
		SaRa	51	12	14	14	2	4	3	6.9394×10^3	0.3918
		mSaRa	99	1	0	0	0	0	0	3.3326×10^3	3.0239

and the second lower plot is the final result after single-peak recognition. By the screening process, we identify 17 points which are very close to change-points. Based on these points, we carry out local search and finally get 5 change-points through single-peak recognition. The all detected change-points are marked by vertical lines.

From this example, we can see that our FSSR algorithm can quickly and accurately find the change-points. In order to show more comparisons, we consider the normal error case in Section 4.3 and t error case in Section 4.4, respectively.

4.2. Simulation Design. Before presenting the detailed comparison, we give the simulation design.

First, the generation of basic data comes from the standard normal distribution and a student $t(3)$ distribution.

Second, the jump size of change-point $\delta_i = \mu_{\tau_i} - \mu_{\tau_{i+1}}$ is generated by a random mechanism. We set $\delta_i = (2B_i - 1)(qU_i + (1 - q))$ where $q \in (0, 1)$ is a variable that controls the degree of heterogeneity of δ_i (in this paper we chooce $q = 0.2$), $U_i \sim N(0, 1)$, $B_i \sim b(1, 0.5)$, and B_i and U_i are independent of each other.

Third, we consider four sample sizes (n=500, 3000, 5000, 8000) and five change-points numbers (N=5, 10, 15, 20, 30, 50). The change-points are scattered in the data segment according to a random mechanism. $N + 1$ random numbers are extracted from the uniform distribution on interval (1, 5) and are recorded as $L_1, L_2, \ldots, L_{N+1}$. We let the location of change-point τ_i be $[(\sum_{j=1}^{i} L_j / \sum_{j=1}^{N+1} L_j)n]$ (i=1,...,N).

4.3. Performance on Normal Data. In this case, because the error is normal, the QN process is not embed into our algorithm. From Table 1, there are some observations as follows.

(1) It is obvious that FSSR has a significant speed advantage. For fixed n, the speed advantage of FSSR is more significant as N becomes smaller. For fixed N, the speed

Table 2: Distribution of $\widehat{N} - N$ for the various competing methods and sample sizes (100 simulated sample paths) with BIC values and run times (t error case).

			$\widehat{N} - N$								
Sample Size		Method	≤ -3	-2	-1	0	1	2	≥ 3	BIC	Time
n=500	N=5	FSSR	23	21	27	20	8	0	1	0.3016×10^3	0.0266
		SaRa	9	11	37	21	18	2	2	0.4719×10^3	0.0233
		mSaRa	39	23	26	11	1	0	0	0.4054×10^3	0.2920
n=3000	N=10	FSSR	18	32	19	24	5	0	2	1.6737×10^3	0.1009
		SaRa	37	9	5	12	3	12	22	2.7135×10^3	0.1253
		mSaRa	47	11	8	8	8	3	15	2.0120×10^3	0.6151
	N=15	FSSR	55	28	12	2	2	1	0	1.6977×10^3	0.1240
		SaRa	68	3	9	5	2	1	12	3.1348×10^3	0.1283
		mSaRa	80	6	6	3	4	1	0	2.1958×10^3	0.6368
n=5000	N=10	FSSR	18	19	30	20	11	2	0	2.7748×10^3	0.1771
		SaRa	35	8	6	5	5	4	37	4.3186×10^3	0.2001
		mSaRa	38	4	8	10	6	4	30	3.3002×10^3	0.8318
	N=20	FSSR	66	23	9	7	1	0	0	2.7637×10^3	0.1921
		SaRa	77	4	4	1	6	2	6	5.0548×10^3	0.1948
		mSaRa	91	2	4	0	2	0	1	3.5509×10^3	0.8705
	N=30	FSSR	99	1	0	0	0	0	0	2.8936×10^3	0.2050
		SaRa	90	0	4	1	0	2	3	6.3680×10^3	0.1879
		mSaRa	99	1	0	0	0	0	0	4.2754×10^3	0.8828
n=8000	N=10	FSSR	10	15	37	27	9	2	0	4.4204×10^3	0.2831
		SaRa	17	5	3	5	6	5	59	6.8527×10^3	0.3153
		mSaRa	15	7	6	4	4	9	55	4.9348×10^3	1.3940
	N=20	FSSR	30	30	28	10	2	0	0	4.3621×10^3	0.3262
		SaRa	61	5	5	4	4	5	16	8.6200×10^3	0.3047
		mSaRa	76	4	4	5	2	1	8	5.6735×10^3	1.2837
	N=30	FSSR	93	5	2	0	0	0	0	4.4635×10^3	0.3484
		SaRa	90	2	4	1	1	0	2	10.5598×10^3	0.3003
		mSaRa	95	2	0	0	2	0	1	6.6622×10^3	1.1897
	N=50	FSSR	100	0	0	0	0	0	0	4.7080×10^3	0.3608
		SaRa	98	1	0	1	0	0	0	12.0999×10^3	0.3109
		mSaRa	100	0	0	0	0	0	0	7.3868×10^3	1.2347

advantage of FSSR is more significant as n becomes larger. In summary, the more sparse the change-points are, the more obvious the speed advantage of FSSR is.

(2) For fixed n, the consistency of change-point detection becomes better as N becomes smaller. For example, the probability of $|\widehat{N} - N| \leq 1$ is up to 0.97.

(3) Under the BIC criterion, the change-point detection result based on our FSSR algorithm is always the best one for segmental fitting data.

4.4. Robustness on t-Distribution. To investigate the effect of our FSSR on the thick tail errors, we set the errors to obey the t distribution with 3 degrees of freedom.

Besides the advantages similar to the normal case, we get some new discoveries in Table 2.

(1) As the QN process is used, the speed advantage of our FSSR is weakened and sometimes it is even slower than SaRa. Compared to mSaRa, the speed advantage of our FSSR algorithm is still very obvious.

(2) Under the same conditions except for the error distribution, the consistency of change-point detection of all algorithms is not as good as those in Table 1.

5. Real Data

5.1. Application to Coriel Data. Several methods based on change-point (e.g., [7, 26]) have been widely studied and applied in copy number variation (CNV) detection.

Generally, as a new source of genetic variation, copy number variation (CNV) plays an important role in phenotypic diversity and evolution. Moreover, many studies have shown that CNV is related to the pathogenicity mechanism of some diseases, including cancer, schizophrenia, and so on [37–40]. Compared with a reference genome assembly [41], CNV usually refers to the deletion or amplification of a region of DNA sequences. Recently, with the significant advances in DNA array technology to detect DNA CNV, various techniques and platforms have been developed for analyzing DNA copy

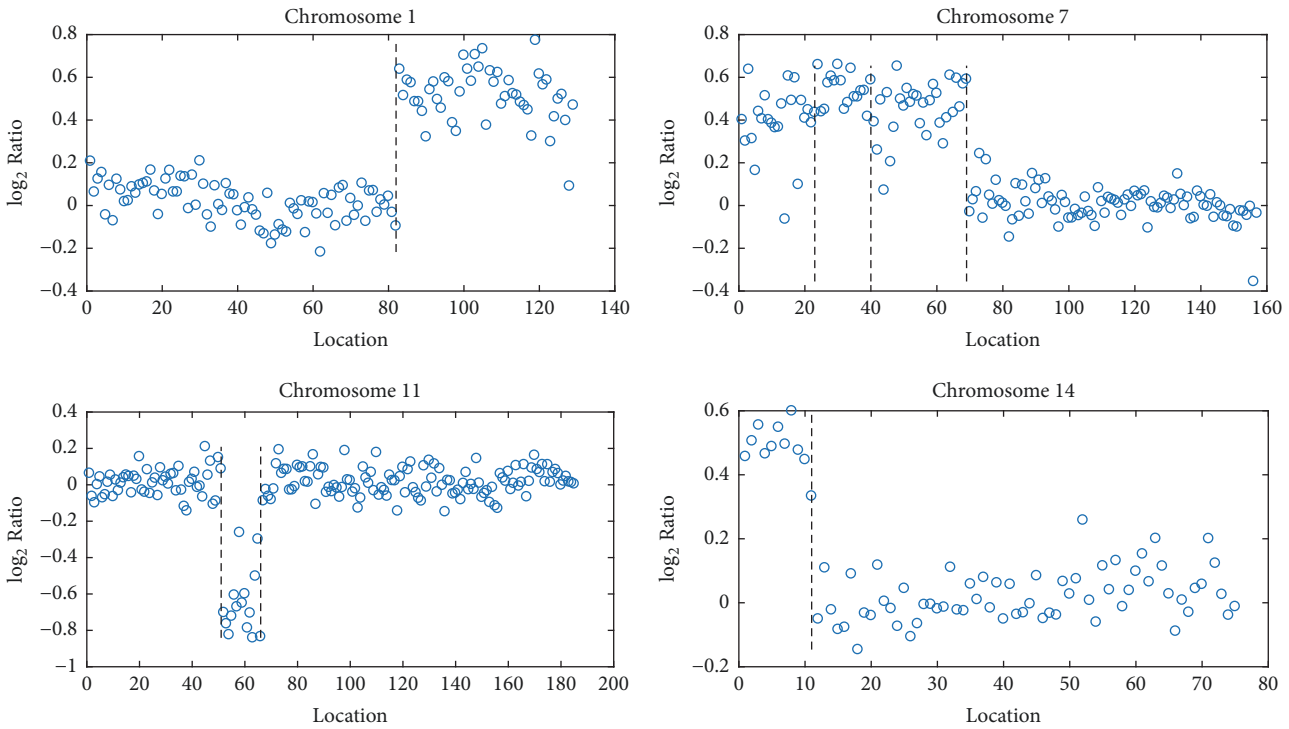

FIGURE 5: A FSSR analysis of the fibroblast cell line on Chromosome 1 of GM13330, Chromosome 7 of GM07081, Chromosome 11 of GM05296, and Chromosome 14 of GM01750.

number, including array comparative genomic hybridization (aCGH), single nucleotide polymorphism (SNP) genotyping platforms, and next-generation sequencing, which provided lots of data. The goal of analyzing of DNA copy number data is to divide the whole genome into segments where copy number vary between contiguous segments and then quantify for each segment. Hence, the target of change-point based methods is to identify the exact locations of copy number changes.

For demonstrating the high efficiency and precision of FSSR, we use the FSSR to analyze the Coriel data set (Download at http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754S1.html), which is firstly studied by Snijders et al. [42]. The well-known data set has been widely used in evaluating CNV detection algorithms ([7, 11, 20, 23, 26, 43, 44] and among others). The data sets consist of a logarithmic ratio of normalized intensities from the disease versus control samples, which are indexed by the physical location of the probes on the genome. The goal is to identify segments of concentrated high or low log ratios. The experiment on 15 fibroblast cell lines makes up the data sets. Each fibroblast cell line contains measurements for 2700 BACs (bacterial artificial chromosome) spotted in triplicate. There are 15 chromosomes with partial alterations and 8 whole chromosomal alterations. All of these alterations but one (Chromosome 15 on GM07801) were confirmed by spectral karyotyping. As shown in Figure 5, we apply FSSR to four chromosomes. They are Chromosome 1 of GM13330, Chromosome 7 of GM07081, Chromosome 11 of GM05296, and Chromosome 14 of GM01750. In the diagram, the points are normalized log ratios, and the dashed lines are locations of change-points detected by our proposed method. As the results show, FSSR identifies all. The results of SaRa or some other methods applied in this real data can consult references ([7, 44] and among others).

5.2. Application to Electric Power System. In this section, we apply the proposed FSSR approach in a real industry application to the electric power system. In the data analysis, the FSSR algorithm can be seen to overperform the SaRa and mSaRa algorithms.

In recent years, the electric distribution network (DN) faces a new challenge to the integration of distributed generations (DGs), after access of distributed scenario energy in the power system. A reasonable and appropriate plan needs to be considered to secure DN for future years. However, in order to save cost, few typical scenarios, which are used to guide in future years, are required to extract from existing massive scenarios. The power load data in the electric power system is typically time series, so the typical scenario reduction can be treated as a problem of detecting change-points.

The real data are collected from the 220kv grade DN of Sichuan province in China. Because the real data can only store for three months in practice, so we intercept data from April 20, 2016, 0:04:00 am, to May 31, 2016, 23:59:00 pm. An observation is recorded every 5 minutes; therefore the sample size is $n = 11802$.

We apply FSSR, SaRa, and mSaRa algorithms to the time series of the power data on two transformers, respectively. The results of active power and reactive power are presented

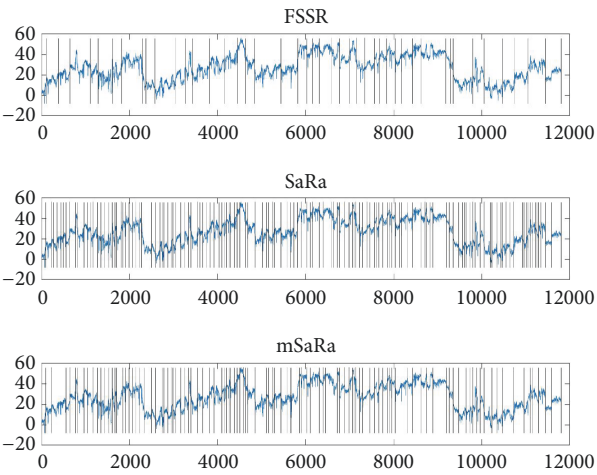

Figure 6: FSSR compared with SaRa and mSaRa applied in power time series data of transformer 1 to extract typical scenarios, with vertical lines corresponding to change-point locations given by the algorithms.

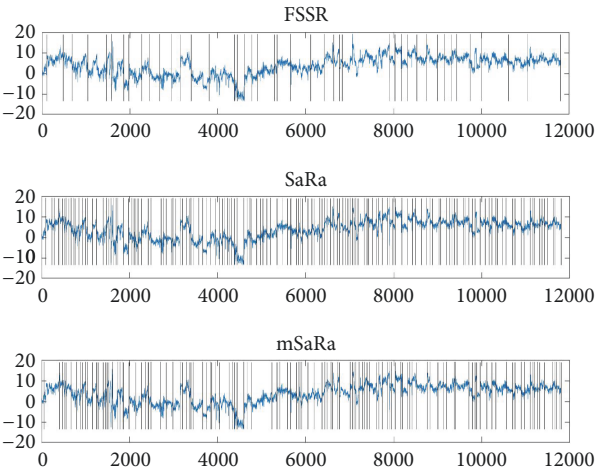

Figure 7: FSSR combined with SaRa and mSaRa applied in power time series data of transformer 2 to extract typical scenarios, with vertical lines corresponding to change-point locations given by the algorithms.

Table 3: The performance of FSSR, SaRa, and mSaRa on transformer 1.

	BIC	Number of Change-points	Time
FSSR	1.5014×10^4	44	0.3148
SaRa	1.6664×10^4	138	0.3055
mSaRa	1.5255×10^4	119	5.2266

in Figures 6 and 7, respectively. In Figures 6 and 7, the vertical line represents the location of change-point given by the algorithms.

Tables 3 and 4 show the fitting effect, number of change-points selected, and running time of three algorithms. The BIC value of FSSR is lowest and the number of change-points given by FSSR is smallest, while the running time of FSSR is almost as short as SaRa and is obviously shorter than mSaRa.

Table 4: The performance of FSSR, SaRa, and mSaRa the transformer 2.

	BIC	Number of Change-points	Time
FSSR	7.0256×10^3	42	0.2950
SaRa	7.2773×10^3	148	0.2214
mSaRa	7.7619×10^3	115	7.2700

6. Concluding Remarks

For the multiple change-point detection problems, an optimal method is mainly evaluated with two aspects: the detecting criterion of change-point and the design of algorithm.

For the criterion of detecting change-points, most of the existing methods are based on the maximization criterion of global CUSUM statistic (such as BS and CBS) or local CUSUM statistic (such as SaRa and mSaRa). From Figure 3, we note that a change-point not only is the local maximum but also should be the local single-peak of the CUSUM statistic distribution. Therefore our FSSR algorithm based on single-peak recognition is more robust than the traditional one by the maximization of the CUSUM statistic. In addition, we use QN on raw data to further enhance robustness.

During the algorithm design, a fast and efficient screening process is considered. We can select the approximate subsegments including change-points at very low computational cost.

Finally, the proposed FSSR has a good performance compared to the comparable existing algorithms according to our simulation and practical application results.

Data Availability

The data used to support the findings of Subsection 5.1 are included within the article, and the data used to support the findings of Subsection 5.2 are included within the supplementary information file.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Youbo Liu gives the practical motivation on the change-point detecting and offers the real data of application in electric power system. Moreover, he provides many good suggestions to revise the manuscript.

Acknowledgments

This research project was supported by the National Natural Science Foundation of China (nos. 11471264, 11401148, and 51437003).

Supplementary Materials

The real data is the power data, which is collected from the integration of distributed generations of Sichuan province in China. The gathering time of the data is from April 20, 2016, 0:04:00 am, to May 31, 2016, 23:59:00 pm. An observation is recorded every 5 minutes, and the sample size is n = 11802. The data contains two parts: active power and reactive power according to the property of the data. Then, we apply FSSR, SaRa, and mSaRa algorithms to the time series of the power data on two transformers, respectively. *(Supplementary Materials)*

References

[1] D. Jarušková, "Change-point detection methods to environmental data," *Environmetrics*, vol. 8, no. 5, pp. 469–483, 1997.

[2] Q. Lu, R. Lund, and T. C. Lee, "An MDL approach to the climate segmentation problem," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 299–319, 2010.

[3] H. J. Jin and D. Miljkovic, "An analysis of multiple structural breaks in US relative farm prices," *Applied Economics*, vol. 42, no. 25, pp. 3253–3265, 2010.

[4] F. Caron, A. Doucet, and R. Gottardo, "On-line changepoint detection and parameter estimation with application to genomic data," *Statistics and Computing*, vol. 22, no. 2, pp. 579–595, 2012.

[5] G. B. Pezzatti, T. Zumbrunnen, M. Bürgi, P. Ambrosetti, and M. Conedera, "Fire regime shifts as a consequence of fire policy and socio-economic development: An analysis based on the change point approach," *Forest Policy and Economics*, vol. 29, pp. 7–18, 2013.

[6] Y.-C. Yao and S. T. Au, "Least-squares estimation of a step function," *Sankhya: The Indian Journal of Statistics, Series A*, vol. 51, no. 3, pp. 370–381, 1989.

[7] Y. S. Niu and H. Zhang, "The screening and ranking algorithm to detect DNA copy number variations," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1306–1326, 2012.

[8] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics*, vol. 42, no. 6, pp. 2243–2281, 2014.

[9] Y. Ninomiya, "Information criterion for Gaussian change-point model," *Statistics & Probability Letters*, vol. 72, no. 3, pp. 237–247, 2005.

[10] Y. C. Yao, "Estimating the number of change-points via schwarzs criterion," *Statistics & Probability Letters*, vol. 6, no. 3, pp. 181–189, 1988.

[11] N. R. Zhang and D. O. Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 63, no. 1, pp. 22–32, 2007.

[12] J. V. Braun and R. K. Braun, "Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation," *Biometrika*, vol. 87, no. 2, pp. 301–314, 2000.

[13] J. Bai and P. Perron, "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, vol. 18, no. 1, pp. 1–22, 2003.

[14] B. Jackson, J. D. Scargle, D. Barnes et al., "An algorithm for optimal partitioning of data on an interval," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 105–108, 2005.

[15] J. Antoch and D. Jaruskova, "Testing for multiple change points," *Computational Statistics*, vol. 28, no. 5, pp. 2161–2183, 2013.

[16] R. Killick and I. A. Eckley, "Changepoint: An R package for changepoint analysis," *Journal of Statistical Software* , vol. 58, no. 3, pp. 1–19, 2014.

[17] R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead, "On optimal multiple changepoint algorithms for large data," *Statistics and Computing*, vol. 27, no. 2, pp. 519–533, 2017.

[18] M. Maciak and I. Mizera, "Regularization techniques in joinpoint regression," *Statistical Papers*, vol. 57, no. 4, pp. 939–955, 2016.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[20] T. Huang, B. Wu, P. Lizardi, and H. Zhao, "Detection of DNA copy number alterations using penalized least squares regression," *Bioinformatics*, vol. 21, no. 20, pp. 3811–3817, 2005.

[21] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, no. 1, pp. 18–29, 2008.

[22] J. Shen, C. M. Gallagher, and Q. Lu, "Detection of multiple undocumented change-points using adaptive Lasso," *Journal of Applied Statistics*, vol. 41, no. 6, pp. 1161–1173, 2014.

[23] Q. Li and L. Wang, "Robust change point detection method via adaptive LAD-LASSO," *Statistical Papers*, vol. 1, pp. 1–13, 2017.

[24] E. S. Venkatraman, *Consistency Results in Multiple Change-Point Situations, [Ph.D. thesis]*, Department of Statistics, Stanford University, 1992.

[25] J. Chen and A. K. Gupta, "Statistical inference on covariance change points in Gaussian model," *Statistics. A Journal of Theoretical and Applied Statistics*, vol. 38, no. 1, pp. 17–28, 2004.

[26] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.

[27] W. R. Lai, M. D. Johnson, R. Kucherlapati, and P. J. Park, "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol. 21, no. 19, pp. 3763–3770, 2005.

[28] A. Batsidis, "Robustness of the likelihood ratio test for detection and estimation of a mean change point in a sequence of elliptically contoured observations," *Statistics*, vol. 44, no. 1, pp. 17–24, 2010.

[29] H. Cho and P. Fryzlewicz, "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: Series B*, vol. 77, no. 2, pp. 475–507, 2015.

[30] N. Hao, Y. S. Niu, and H. Zhang, "Multiple change-point detection via a screening and ranking algorithm," *Statistica Sinica*, vol. 23, no. 4, pp. 1553–1572, 2013.

[31] Z. Chen and Y. Hu, "Cumulative sum estimator for change-point in panel data," *Statistical Papers*, vol. 58, no. 3, pp. 707–728, 2017.

[32] J. Cabrieto, F. Tuerlinckx, P. Kuppens, F. H. Wilhelm, M. Liedlgruber, and E. Ceulemans, "Capturing correlation changes by applying kernel change point detection on the running correlations," *Information Sciences*, vol. 447, pp. 117–139, 2018.

[33] C.-S. J. Chu, "Time series segmentation: A sliding window approach," *Information Sciences*, vol. 85, no. 1-3, pp. 147–173, 1995.

[34] F. Xiao, X. Min, and H. Zhang, "Modified screening and ranking algorithm for copy number variation detection," *Bioinformatics*, vol. 31, no. 9, pp. 1341–1348, 2015.

[35] C. Y. Yau and Z. Zhao, "Inference for multiple change points in time series via likelihood ratio scan statistics," *Journal of the Royal Statistical Society Series B*, vol. 78, no. 4, pp. 895–916, 2016.

[36] C. Song, X. Min, and H. Zhang, "The screening and ranking algorithm for change-points detection in multiple samples," *The Annals of Applied Statistics*, vol. 10, no. 4, pp. 2102–2129, 2016.

[37] S. J. Diskin, C. Hou, J. T. Glessner et al., "Copy number variation at 1q21.1 associated with neuroblastoma," *Nature*, vol. 459, no. 7249, pp. 987–991, 2009.

[38] G. Kirov, "The role of copy number variation in schizophrenia," *Expert Review of Neurotherapeutics*, vol. 10, no. 1, pp. 25–32, 2010.

[39] P. Ibáñez, A.-M. Bonnet, B. Débarges et al., "Causal relation between α-synuclein gene duplication and familial Parkinson's disease," *The Lancet*, vol. 364, no. 9440, pp. 1169–1171, 2004.

[40] J. A. Lee, C. M. B. Carvalho, and J. R. Lupski, "A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders," *Cell*, vol. 131, no. 7, pp. 1235–1247, 2007.

[41] R. Redon, S. Ishikawa, K. R. Fitch et al., "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, 2006.

[42] A. M. Snijders, N. Nowak, R. Segraves et al., "Assembly of microarrays for genome-wide measurement of DNA copy number," *Nature Genetics*, vol. 29, no. 3, pp. 263-264, 2001.

[43] J. Fridlyand, A. M. Snijders, D. Pinkel, D. G. Albertson, and A. N. Jain, "Hidden Markov models approach to the analysis of array CGH data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 132–153, 2004.

[44] X.-L. Yin and J. Li, "Detecting copy number variations from array cgh data based on a conditional random field model," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 2, pp. 295–314, 2010.