The Canadian Journal of Statistics
Vol. 48, No. 3, 2020, Pages 366–385
La revue canadienne de statistique

A backward procedure for change-point detection with applications to copy number variation detection

Seung Jun SHIN[1], Yichao WU[2*], and Ning HAO[3]

[1]*Department of Statistics, Korea University, Seoul, South Korea*
[2]*Department of Mathematics, Statistics, and Computer Science, The University of Illinois at Chicago, Chicago, IL, U.S.A.*
[3]*Department of Mathematics, The University of Arizona, Tuscon, AZ, U.S.A.*

Key words and phrases: Backward detection; copy number variation; mean change-point model; multiple change points; Short signal.
MSC 2010: Primary 62F10; secondary 62P10.

Abstract: Change-point detection regains much attention recently for analyzing array or sequencing data for copy number variation (CNV) detection. In such applications, the true signals are typically very short and buried in the long data sequence, which makes it challenging to identify the variations efficiently and accurately. In this article, we propose a new change-point detection method, a backward procedure, which is not only fast and simple enough to exploit high-dimensional data but also performs very well for detecting short signals. Although motivated by CNV detection, the backward procedure is generally applicable to assorted change-point problems that arise in a variety of scientific applications. It is illustrated by both simulated and real CNV data that the backward detection has clear advantages over other competing methods, especially when the true signal is short. *The Canadian Journal of Statistics* 48: 366–385; 2020 © 2020 Statistical Society of Canada

Résumé: La détection de points de rupture gagne en popularité pour la détection de variations du nombre de copies (VNC) avec des données de micro-puces ou de séquençage. Dans de telles applications, un signal habituellement court se trouve dans une longue séquence de données, ce qui complique sa détection efficace et précise. Les auteurs proposent une nouvelle méthode de détection du point de rupture, une procédure à rebours, qui en plus d'être suffisament simple et rapide pour exploiter des données en haute dimension, offre de très bonnes performances dans la détection de signaux courts. Même si elle a été motivée par la détection de VNC, la méthode à rebours est généralement applicable à une panoplie de problèmes de détection du point de rupture qui émergent de différentes applications scientifiques. Les auteurs illustrent les avantages clairs de la méthode à rebours, notamment lorsque le signal est court, par rapport aux autres méthodes, autant sur des données de VNC simulées que réelles. *La revue canadienne de statistique* 48: 366–385; 2020 © 2020 Société statistique du Canada

1. INTRODUCTION

As a classic topic, change-point detection regains much attention recently in the context of uncovering structural change in big data. In particular, a normal mean change-point model and its variants have been applied to analyze high-throughput data for DNA copy number variation (CNV) detection. The CNV is defined as duplication or deletion of a segment of DNA sequences

* *Author to whom correspondence may be addressed.*
E-mail: yichaowu@uic.edu

TABLE 1: Data types for CNV information and the corresponding resolutions.

Data	Measurement	Resolution
Array CGH	Fluorescence intensity ratio	10–100× kb
SNP array	Fluorescence intensity ratio/B-allele frequency	kb
NGS	Read depth/distance of paired end	b

compared to the reference genome, and can cause significant effects at molecular levels and be associated with susceptibility (or resistance) to disease (Feuk, Carson & Scherer, 2006; Freeman et al., 2006; McCarroll & Altshuler, 2007).

There are multiple sources of data that provide us with copy number information. The microarray Comparative Genomic Hybridization (aCGH) techniques have been widely used for CNV detection (Urban et al., 2006). The aCGH is helpful to detect long CNV segments with tens of kilobases (kb) or more, but is not able to locate small-scale CNVs with length shorter than its minimal resolution (>10 kb), which are common in the human genome (Sebat et al., 2004; Carter, 2007; Wong et al., 2007). In addition to the aCGH approach, the single nucleotide polymorphism (SNP) genotyping array has become an alternative in CNV detection because of its improved resolution (Sun et al., 2009). For example, popular SNP array platforms such as Illumina (Peiffer et al., 2006) and Affymetrix (McCarroll et al., 2008) allow detection of CNVs with kilobase-resolution. In SNP arrays, CNV information is measured by the total fluorescent intensity signal ratios from both alleles at each SNP locus referred to as the log-*R*-ratio (LRR). It also allows us to obtain the relative ratio of the fluorescent signals between two alleles, known as B allele frequency. Finally, in very recent applications, aligned DNA sequencing data with even higher resolution can be directly used for CNV detection. The next generation sequencing (NGS) techniques typically produce millions of short reads that are to be aligned with the reference genome. Both the associated read depth (RD) and distances of paired-end (DPE) from the aligned sequence are important sources of inferring CNV (Medvedev, Stanciu & Brudno, 2009; Abyzov et al., 2011; Duan et al., 2013; Chen et al., 2017). Note that there is a trade-off between the resolution and data size. With higher resolution data, it is possible to discover shorter CNVs; at the same time, the larger data size brings great challenge in computation. This might be one of the reasons why the SNP array has been most popular in recent CNV studies since aCGH data have low resolutions and RD or DPE data from NGS are too large to be handled directly. However, as related computing technologies advance, the NGS data are getting more attention in recent applications. Finally, we summarize popular data sources for CNV detection in Table 1.

There have been many methods developed for CNV detection. Different approaches are applied to different types of data sources. As one of the most popular approaches, the CNV detection problem can be regarded as an application of the change-point model which has been actively studied in statistics. For example, both the LRR from SNP array and $\log_2$ ratio from aCGH have a mean value of zero for normal copy number, while negative (resp. positive) for the deletion (resp. duplication). Similar ideas can also be employed for the RD data in the sense that one may observe less (resp. more) read counts in a region with deleted (resp. duplicated) copy number. In all the examples, the data structure changes at CNVs. Naturally, one may infer CNVs by checking the subregions where the LRR or read-depth is significantly different from the mean of rest.

The change-point model has a long history that traces back to the 1950s. See Page (1955, 1957), Chernoff & Zacks (1964), Gardner (1969) and Sen & Srivastava (1975) for the early developments of the change-point model with at most one change point. The data size

considered in those papers is also small. However, new applications call for more flexible models capable of detecting multiple change points scattered along a huge sequence. Recent developments include circular binary segmentation (CBS, Olshen et al., 2004; Venkatraman & Olshen, 2007), ℓ_1 penalization (Huang et al., 2005; Tibshirani & Wang, 2008; Zhang et al., 2010), total-variation-penalized estimation (TVP, Harchaoui & Lévy-Leduc, 2010), fragment assembling algorithm (FASeg, Yu et al., 2007), screening and ranking algorithm (SaRa, Niu & Zhang, 2012; Hao, Niu & Zhang, 2013), likelihood ratio selection (LRS, Jeng, Cai & Li, 2010), simultaneous multiscale change point estimator (SMUCE, Frick, Munk & Sieling, 2014) and wild binary segmentation (WBS, Fryzlewicz, 2014) among many others. Hidden Markov model is another popular approach for CNV detection (Fridlyand et al., 2004; Wang et al., 2007; Szatkiewicz et al., 2013). Yet the hidden Markov model relies on some application specific-assumptions valid only for certain copy-number data. Zhang (2010) provided a comprehensive overview on CNV detection as an application of the change-point model. Roy & Reif (2013) compared the performance of a few recent CNV detection methods under various scenarios.

As an application of the change-point model, inferring CNV is regarded as a very challenging problem since the CNV subregions are usually very short and hidden in a very long sequence. The size of detectable CNVs typically ranges from a thousand to millions of base pairs (bp). The International HapMap 3 Consortium (2010) shows that the average size of total CNVs in the individual genome is 3.5 ± 0.5 Mbp (0.1%). As an illustration, we analyze the SNP array data collected from the Autism Genetics Resource Exchange (AGRE; Bucan et al., 2009) which contain three parallel LRR sequences of a father--mother–offspring trio. Figure 1a depicts the LRR sequence of the mother's whole genome and clearly illustrates that it is impossible to pick CNV out by eye. Figure 1b shows a zoom-in plot of one of the detected CNVs from the mother's sequence whereas Figure 1c displays a histogram of sizes (in terms of the number of biomarkers) of CNVs which are commonly detected by the methods considered in this article, respectively. The size of the entire LRR sequence in Figure 1a is 561,466, while the one CNV depicted in Figure 1b has size 6. We would like to remark that the most of CNVs are very short (as shown in Figure 1c) and hidden in a long and noisy sequence, which makes the detection non-trivial. We will revisit this data in Section 6 where a complete analysis is illustrated.

In this context, a desirable CNV-detection method should be not only accurate enough to detect such short CNVs but also computationally fast enough to get estimates within a practically manageable time limit even for a very long data sequence. Toward this, we propose a new

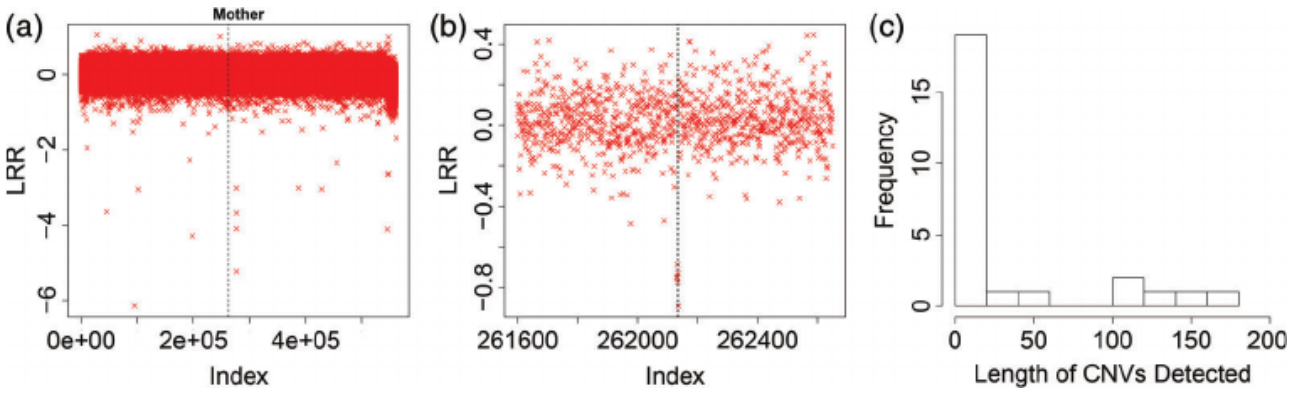

FIGURE 1: The subpanel (a) shows the whole sequence of the mother's LRR; (b) depicts one of them from the mother's LRR sequence of the AGRE trio SNP array data; and (c) is a histogram of length of CNVs commonly detected by several methods.

change-point detection method called backward detection (BWD) whose name comes from the backward variable selection in linear regression. BWD is computationally efficient, with a complexity $O(n \log n)$ to analyze a sequence of length n. Moreover, it performs very well in picking out closely located change-points. Therefore, it is an ideal tool for CNV detection. The idea of the BWD is conceptually similar to Wald's agglomerative clustering (Wald, 1963), but different in that the location information of the sequence data is employed naturally. We also note that our method for change-point detection can be viewed as a "bottom-up" strategy which has not been studied as extensively in the literature as "top-down" ones (such as binary segmentation) mainly on account of the computational intensity of "top-down" methods. Recently, Fryzlewicz (2018) proposed an efficient "bottom-up" method for the general multiple change-point detection problem by using what he calls the tail-greedy unbalanced Haar (TGUH) transformation. Yet, our numerical simulation shows that the TCUH often fails to detect short and sparse signals commonly encountered in CNV detection.

The rest of the article is organized as follows. In Section 2, a normal mean change-point model and several popular detection strategies are introduced. In Section 3, the BWD method is described in detail. A stopping rule for BWD is developed in Section 4. The numerical performance of the proposed method is evaluated in Section 5, and illustrations to both $\log R$ ratio data from SNP arrays and the RDs from aligned sequence data are given in Section 6. A discussion follows in Section 7.

2. CHANGE-POINT MODEL

A normal mean change-point model assumes

$$Y_i = \mu_i + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

with ϵ_i being iid $N(0, \sigma^2)$. The means μ_i are assumed to be piecewise constant with K change points at $\mathbf{t} = (t_1, t_2, \ldots, t_K)^T$. Denote $t_0 = 0$ and $t_{K+1} = n$ for convenience. Change points are characterized by the property that $\mu_i = \mu_j$ for any $i, j \in \{t_k + 1, \ldots, t_{k+1}\}$, $0 \le k \le K$, and $\mu_{t_k} \ne \mu_{t_k+1}$ for $1 \le k \le K$. Yet the number of change points K is typically unknown. The goal of change-point detection is to estimate both the number K and the location vector of change points $\mathbf{t}$. Thus CNV detection can be regarded as a direct application of the change-point model (1). However, the CNVs are often very short and buried in a very long data sequence, which makes the problem even more challenging because of the high dimension of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$. The normal mean change-point model (1) is often reasonable in CNV application on account of random noise during the experiments (Barnes et al., 2008). In other cases, data may need to be transformed. For example, raw RD data are discrete and spatially correlated on account of the complicated sequencing process, and hence the normal error assumption is not proper. Nevertheless, the local median transformation can be used to ensure its normality (Cai, Jeng & Li, 2012).

Suppose that the number of change points, K, is known, then the change-point detection problem can be formulated in terms of minimizing the sum of squared errors (SSE). For a set of numbers Y_i with index i in a set $\mathcal{G}$, we define their SSE by $\mathrm{SSE}(\{Y_i, i \in \mathcal{G}\}) = \sum_{i \in \mathcal{G}} (Y_i - \bar{Y}_{\mathcal{G}})^2$ where $\bar{Y}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} Y_i$ and $|\mathcal{G}|$ denotes the cardinality of the index set $\mathcal{G}$. Then change-point detection is to estimate $\mathbf{t}$ by solving the following optimization problem

$$\min \sum_{k=1}^{K} \mathrm{SSE}(\{Y_i : t_k + 1 \le i \le t_{k+1}\}) \tag{2}$$

$$\min_{\mathbf{t}} \sum_{j=0}^{K} \mathrm{SSE}\left(\left\{Y_i : t_j + 1 \le i \le t_{j+1}\right\}\right), \tag{2}$$

$$\text{subject to } 0 = t_0 < t_1 < t_2 < \cdots < t_K < t_{K+1} = n.$$

Note that (2) is inherently a combinatorial problem and very challenging for large n. The total number of different combinations is $\frac{n!}{K!(n-K)!} \geq \left(\frac{n}{K}\right)^K$ which can be huge especially in the application of CNV. This makes it difficult to detect change points by solving (2) directly not to mention the fact that K is typically not known. When n is small and K is bounded, the exhaustive search method has been studied by Yao (1988) and Yao & Au (1989). They showed that an exhaustive search with BIC is consistent for estimating K and $\mathbf{t}$. To improve computational efficiency, dynamic programming can be applied to solve (2) with complexity of $O(n^2)$ (Braun, Braun & Muller, 2000; Jackson et al., 2005). Killick, Fearnhead & Eckley (2012) developed an efficient algorithm named PELT that solves the problem with linear cost $O(n)$, but requires additional assumptions which might not be practical in certain applications. In general, these methods have not been widely applied in CNV applications.

We remark that the mean change-point model (1) can be equivalently reformulated as a linear regression model (Huang et al., 2005; Tibshirani & Wang, 2008) and the change-point detection problem is then viewed as a variable selection one. Motivated by backward elimination methods in variable selection, we propose a stepwise procedure called BWD to solve the change-point detection problem. Some stepwise methods in the context of change-point detection have been explored. For example, a classical binary segmentation method (BS; Vostrikova, 1981) applies a single change-point detection tool recursively, identifying one change point at a time, until some stopping criterion is met. We consider BS to be a forward detection method because it starts with a null model with no change point and sequentially detects change points, by analogy to the forward variable selection in a regression context. In spite of its simplicity, as pointed out by Olshen et al., (2004), forward detection is not able to detect short segments buried in a long sequence of observations, which limits its utilization in certain applications such as CNV detection. The CBS (Olshen et al., 2004; Venkatraman & Olshen, 2007) modifies BS by identifying two change points simultaneously and has gained great popularity in CNV detection. However, we observe from limited numerical studies that on the account of the forward detection nature of CBS, it is still unsatisfactory when the true segment (i.e., CNV) is very short. Moreover, CBS has higher computational complexity than the BS, which brings additional burden to dealing with big data.

3. METHOD

3.1. Why Not Forward Detection?

In what follows, we elaborate why forward detection may fail to detect short signals, thus providing a clear motivation for the proposed BWD in CNV detection. Forward detection starts with no change point and tries to detect the very first one by solving the following optimization problem

$$\min_{s_1} \sum_{j=0}^{1} \mathrm{SSE}(\{Y_i, i \in \{s_j + 1, \ldots, s_{j+1}\}\}), \tag{3}$$

$$\text{subject to } 0 = s_0 < s_1 < s_2 = n.$$

The optimizer $\hat{s}_1$ of (3) estimates one of the change points and divides the data into two parts $\{Y_i, i \in \{1, 2, \ldots, \hat{s}_1\}\}$ and $\{Y_i, i \in \{\hat{s}_1 + 1, \ldots, n\}\}$. We may apply (3) to each of these two parts to detect further change points and this can be continued until we have identified all change points.

Note that the total number of combinations for (3) is n and thus the corresponding optimization is feasible. However, as mentioned above the performance of forward detection may not be satisfactory in some situations. For example, if there are only two change points at t_1 and t_2

and $\mu_i = 0$ if $i < t_1$ or $i \geq t_2$ and μ if $t_1 \leq i < t_2$, the forward detection does not work well if the length of signal $L = t_2 - t_1$ is small while both t_1 and $n - t_2$ are large. However, this type of challenging situation is very common in the CNV applications as shown in Figure 1.

To illustrate drawbacks of the forward detection, we consider a simplified scenario in which the locations of the two potential change points t_j, $j = 1, 2$ are known, but it is not clear whether the associated mean μ is actually changed (i.e., $\mu = 0$ or not). Proposition 1 formally states that forward detection asymptotically fails even in this simple scenario unless L is sufficiently large compared to n.

Proposition 1. *Suppose* $\lim_{n\to\infty} t_1/n = c \in (0, 1)$ *and* $L = t_2 - t_1 = O(n^\beta)$ *for some* $\beta \in [0, 1]$. *If* $\beta < 1/2$ *then the forward detection fails as* $n \to \infty$ *for an arbitrary given* (μ, σ^2).

Proof. The first step of forward detection declares that the mean-change occurs at $t_j, j = 1, 2$ if $|\bar{D}_{n,t_j}| = |\bar{Y}_{t_j} - \bar{Y}_{n-t_j}|$ is significantly large enough. Here $\bar{Y}_t = \sum_{i=1}^{t} Y_i/t$ and $\bar{Y}_{n-t} = \sum_{i=t+1}^{n} Y_i/(n-t)$, for a given $t \in \{1, \ldots, n-1\}$.

To test for t_1, the sampling distribution of $\bar{D}_{n,t_1}$ for a given μ is

$$\frac{\bar{D}_{n,t_1} + L\mu/(n - t_1)}{\hat{\sigma}_n \sqrt{\frac{1}{t_1} + \frac{1}{n-t_1}}} \quad \to \quad N(0, 1), \quad \text{in distribution,}$$

where $\hat{\sigma}_n^2$ denotes a consistent estimator of unknown σ^2. Then it can be shown that the associated asymptotic power converges to the nominal level α for any given pair of (μ, σ^2) if $\lim_{n\to\infty} n^{-1/2} L = 0$. A similar result can be shown for t_2 as well, which completes proof. ■

Proposition 1 provides a necessary condition for forward detection in terms of the relative length of the true signal length L as a function of sample size n. The order of L, denoted by β, should be larger than $1/2$ for the original change-point model in which the change points t_1 and t_2 are unknown. Recently, Fryzlewicz (2014) showed that forward selection is consistent for recovering the true change points when β is larger than 3/4.

3.2. Backward Detection

Contrary to forward detection, the BWD starts from the opposite extreme that every single position is assumed to be a change point. That is, we begin with n groups corresponding to these $n - 1$ change points and each group contains only one observation. We introduce notation $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_n\}$ with $\mathcal{G}_i = \{i\}$.

The BWD works by repeatedly merging two neighbouring groups into one. Note that the merging of two neighbouring groups will increase the total SSE. For any two neighbouring groups, we use the rise in the SSE to quantify the potential of merging them together. At each merging step, we choose to merge two neighbouring groups with the smallest rise of SSE. Namely, we define

$$R_i = \text{SSE}(\{Y_j, j \in \mathcal{G}_i \cup \mathcal{G}_{i+1}\}) - \text{SSE}(\{Y_j, j \in \mathcal{G}_i\}) - \text{SSE}(\{Y_j, j \in \mathcal{G}_{i+1}\}), \tag{4}$$

where $\text{SSE}(\{Y_j, j \in \mathcal{G}\})$ denotes the SSE for all observations with indices in $\mathcal{G}$.

At the beginning of iteration $m = 0, 1, \ldots, n-2$, there are $n - m$ groups. Denote the current groups by $\mathbb{G}^{(m)} = \{\mathcal{G}_1^{(m)}, \mathcal{G}_2^{(m)}, \ldots, \mathcal{G}_{n-m}^{(m)}\}$ and the corresponding potential of merging two neigh-

bouring groups by $\{R_1^{(m)}, R_2^{(m)}, \ldots, R_{n-m-1}^{(m)}\}$. The superscript is used to represent the mth iteration. Identify $j = \text{argmin}_{i=1,2,\ldots,n-m-1} R_i^{(m)}$. Then we merge groups $\mathcal{G}_j^{(m)}$ and $\mathcal{G}_{j+1}^{(m)}$ into a new group.

Updated grouping is denoted by $\mathbb{G}^{(m+1)} = \{\mathcal{G}_1^{(m)}, \mathcal{G}_2^{(m)}, \ldots, \mathcal{G}_{j-1}^{(m)}, \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}, \mathcal{G}_{j+2}^{(m)}, \ldots, \mathcal{G}_{n-m}^{(m)}\}$ and potentials of merging is updated to $\{R_1^{(m)}, R_2^{(m)}, \ldots, R_{j-2}^{(m)}, R_{j-}^{(m)}, R_{j+}^{(m)}, R_{j+2}^{(m)}, \ldots, R_{n-m-1}^{(m)}\}$, where

$$R_{j-}^{(m)} = \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_{j-1}^{(m)} \cup \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}\}) - \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_{j-1}^{(m)}\}) - \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}\}), \text{ and}$$

$$R_{j+}^{(m)} = \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)} \cup \mathcal{G}_{j+2}^{(m)}\}) - \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}\}) - \mathrm{SSE}(\{Y_j, j \in \mathcal{G}_{j+2}^{(m)}\}).$$

Now, the steps described above are repeatedly applied until a desired stopping rule is satisfied. The associated stopping rule is discussed in the following section. If the procedure is not terminated, only one group will survive at the end of iteration $n - 2$.

Despite their structural similarity, the BWD is substantially different from forward detection, since the null and alternative hypotheses at each step are reversed. At each step, the BWD tests the equivalence between the two group means while the forward tests their difference. Therefore, the BWD tends to stay with more groups with smaller sizes unless there is strong evidence to merge some of them and hence is more powerful to detect short signals buried on a long sequence. We also remark that the BWD starts with solving a series of local problems (each of which focuses on finding structural changes in a small part of the data) and at the end becomes a single global problem that employs the entire sequence. On the other hand, forward detection starts as a global method and divides it into several local problems. This is one of the reasons why BWD is preferred for identifying short signals in lengthy noise sequences, where local methods are known to outperform global methods in general.

Finally, the BWD algorithm can be summarized as follows.

`Input`: $Y_1, \ldots, Y_n$.

1. Initialize $\mathbb{G}^{(1)} = \{\{1\}, \ldots, \{n\}\}$ and $\mathbf{R}^{(1)} = \{R_1^{(1)}, \ldots, R_{n-1}^{(1)}\}$ from (4).
2. At the mth iteration, $m = 1, \ldots, n-1$:

   2–1 Obtain $j = \mathrm{argmin}_i R_i^{(m)}$.

   2–2 Break the loop if $R_j^{(m)}$ is larger than a prespecified cutoff, and go to the next step otherwise.

   2–3 Update

$$\mathbb{G}^{(m+1)} = \{\mathcal{G}_1^{(m)}, \ldots, \mathcal{G}_{j-1}^{(m)}, \mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}, \mathcal{G}_{j+2}^{(m)}, \ldots, \mathcal{G}_{n-m}^{(m)}\},$$

$$\mathbf{R}^{(m+1)} = \{R_1^{(m)}, R_2^{(m)}, \ldots, R_{j-2}^{(m)}, R_{j-}^{(m)}, R_{j+}^{(m)}, R_{j+2}^{(m)}, \ldots, R_{n-m-1}^{(m)}\},$$

$$K = n - m - 1.$$

`Output`: $\mathbb{G}^{(K)}$.

2.2. Modification for Epidemic Change Points

3.3. Modification for Epidemic Change Points

In CNV analysis, most parts of a sequence (normal) have a known baseline mean, say μ_0, and a mean-change away from μ_0 (variant) is followed by a change back to μ_0. This is often referred

to as an epidemic change points (Yao, 1993) and is an important feature of CNV analysis. To take into account such a pairing structure, we modify the algorithm by adding the following Step 2−2′ between Steps 2−2 and 2−3.

2−2′ If the sample average of observations in the merged sets, $\bar{Y}_{\mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}}$ is not significantly different from the baseline mean μ_0, that is, $v^{1/2}\hat{\sigma}_n^{-1}\left|\bar{Y}_{\mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}} - \mu_0\right| > z_\alpha$ where z_α is the upper αth quantile of a standard normal random variable and $v = \left|\mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}\right|$, then update $R_{j-}^{(m)}$ and $R_{j+}^{(m)}$ based on μ_0 instead of $\bar{Y}_{\mathcal{G}_j^{(m)} \cup \mathcal{G}_{j+1}^{(m)}}$.

Finally, we have developed the `bwd` R-package, available on CRAN, that implements the proposed algorithm.

3.4. Computational Complexity

Computational efficiency is of practical interest in CNV applications due to their inherent high-dimensionality. At each iteration in the BWD, the most computationally intensive part is to find $j = \text{argmin}_i R_i^{(m)}$ which takes $O(n)$ at the worst. This gives the total complexity of $O(n^2)$, which is too slow especially when n is very large.

However, it is realized that finding the maximum and corresponding index is straight forward once $\mathbf{R}^{(1)}$ is ordered, which takes $O(n \log n)$ computations. Note that the sorting step is required only once at the initial stage. Once it is sorted, it takes $O(1)$ to find the maximizer index at the mth iteration while we need additional effort to update $\mathbf{R}^{(m+1)}$ in an ordered fashion. However, such an update takes only $O(\log n)$ computations. In particular, we borrow the idea from the bi-section method, a well-known root finding algorithm. We can first compare $R_{j+}^{(m)}$ (or $R_{j-}^{(m)}$) to the median of the values in $\mathbf{R}^{(m)}$. Compare it with the 75th percentile if it is greater than the median and 25th percentile otherwise. We continue this until finding its exact location. Finally, the total computational complexity of the BWD is then reduced to $O(n \log n)$.

4. STOPPING RULE

In every step of the BWD, two small groups are merged into a bigger group and we want to test whether this merging removes a real change point. In such a standard case, it is natural to use the t-statistic. Since the unknown variance is assumed to be homogeneous across all the observations, a global estimate of the noise variance is used at every step. At the mth iteration the following statistic

$$S_{(m)} = \frac{\left|\bar{Y}_j^{(m)} - \bar{Y}_{j+1}^{(m)}\right|}{\hat{\sigma}_n \sqrt{\left|\mathcal{G}_j^{(m)}\right|^{-1} + \left|\mathcal{G}_{j+1}^{(m)}\right|^{-1}}}, \tag{5}$$

is used to determine when to stop. The backward procedure is terminated if $S_{(m)}$ is too large. Here $\hat{\sigma}_n^2$ denotes an estimate of the unknown noise variance based on all the observation. If the true signals are very short and sparse, the sample variance of Y_i can also be used in practice as a simple alternative. The use of a global estimate brings an additional saving in computation since $R_j = \hat{\sigma}_n^2 S_{(m)}^2$. We use $\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_i^{(h)}\right)^2$ with $\bar{Y}_i^{(h)} = (2h+1)^{-1} \sum_{j=i-h}^{i+h} Y_j$ for a given window $h > 0$ in the upcoming analysis as used in Niu & Zhang (2012). An alternative estimate

is the median absolute deviation estimator, as pointed out in Jeng, Cai & Li (2010).

Similarly to the usual t-statistic, $S(m)$ in (5) may cause a false alarm when both of the two groups have small size. To avoid the possible false alarm caused by small group sizes, we can set

$S_{(m)} = 0$ if both of the two consecutive segments are shorter than a minimum number M. The M can be chosen to be, say, 3 or 5 depending on the application. Such a modification is acceptable in CNV applications since it is very unlikely that any of two CNVs are closely located.

Now, the question is how large the critical value should be to attain a desired target level α where α denotes the familywise error rate (FER) of the proposed BWD. We remark that the $(1 - \alpha/2)$th quantile of the t-statistic with the associated degrees of freedom will fail to attain the nominal level since $S_{(m)}$ is correlated with other group means via the maximizer index j. We propose the following numerical procedure to select a cutoff value that controls FER being at most α.

1. Repeat the steps (a)–(c) below B times: for each iteration b, $b = 1, \ldots, B$,

   1-(a) Randomly generate a sequence of size n from the null distribution that there exists no change point.
   1-(b) Apply the backward procedure until merging the whole sequence into one group.
   1-(c) $u_b = \max_{m=1,\ldots,n-1} S_{(m)}$.

2. The $(1 - \alpha)$th sample quantile of $u_1, \ldots, u_B$ would be the cutoff value which attains a given level α.

We remark that the cutoff value is chosen from the null distribution of $\max_{m=1,\ldots,n-1} S_{(m)}$, not $S_{(m)}$, thus α controls the FER. The very first step 1-(a) that simulates samples from the null distribution is crucial in the proposed numerical procedure. Toward this we consider two scenarios: (i) Normality is assumed to be true while a noise variance σ^2 is still unknown. (ii) Neither normality nor σ^2 are known. In the first scenario, we can generate samples form the standard normal distribution. Note also that this can be easily extended to any distribution other than normal distribution, whenever it is known. In the second scenario when the underlying distribution is not known, the null distribution can be obtained by randomly permuting or bootstrapping residuals $r_i = Y_i - \bar{Y}_i^{(h)}, i = 1, \ldots, n$.

The proposed numerical procedure becomes computationally too intensive especially when the sample size is very large, for instance over a million, which is not uncommon in CNV applications. Under the normality assumption, we numerically investigate cutoff values for different $\alpha = 0.01, 0.05$ and 0.10 as functions of sample size n. Figure 2 depicts estimated cutoffs for different sample sizes from 1,000 to 100,000 by 1,000 and it shows a clear log-linear relationship between the estimated cutoffs and the sample size n. Thus desired cutoffs for large n can be approximated from the fitted regression line.

5. SIMULATED EXAMPLES

We evaluated the performance of the proposed backward procedure via numerical comparison against existing methods. The target levels considered were $\alpha = 0.01$ and 0.05. We considered both the original BWD (BWD1) and the modified BWD (BWD2) for epidemic change-points under the assumption that the baseline mean μ_0 is known. As described in Section 4, there are two ways to obtain the cutoff values depending on how to simulate null samples. We can obtain a cutoff either from standard normal samples under the normality assumption (cutoff1) or from the permuted residuals if the normality assumption is not valid (cutoff2). We used the former in Section 5.1 with Gaussian error and the latter in Section 5.2 with non-Gaussian error. We considered CBS, WBS, LRS and TVP as competing methods. CBS is one of the most widely

used methods in the literature and shares principals similar to a typical stepwise approach with the proposed method. WBS is a recent development based on binary segmentation (i.e., forward detection) that overcomes its shortcoming when detecting a short signal. LRS is a carefully

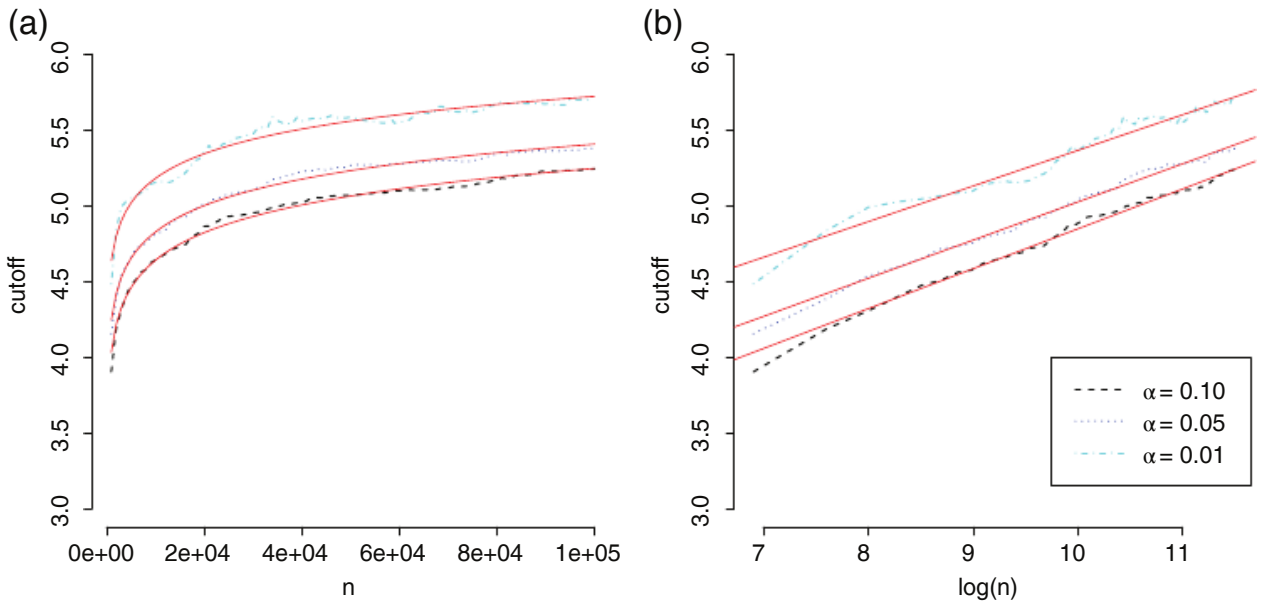

FIGURE 2: Log-linear relationship between estimated cutoffs ($\alpha = 0.01, 0.05, 0.10$) and sample size n under normality assumption. The (red) solid lines in (a) and (b) are fitted regression lines of cutoffs on sample size n and $\log n$, respectively.

designed method for detecting sparse and short signals and is known to be optimal under some required model assumptions that include normality, and shortness and sparsity of the signals. In addition, we compared our method to a recently proposed bottom-up method, called TGUH (Fryzlewicz, 2018). TGUH is designed for a general change-point detection problem and our simulation shows that TBUH is in pain when it detects a short signal.

We considered the following mean change-point model

$$y_i = \sum_{k=1}^{\kappa} \delta_k \mathbf{1}_{\{i \in I_k\}} + \epsilon_i,$$

where $I_k \subset \{1, \ldots, n\}, k = 1, \ldots, \kappa$ denote index sets corresponding to true signals with $I_k \cap I_{k'} = 0$ for any $k \neq k'$, κ is the number of signal segments (i.e., CNVs), and δ_k, are unknown means of true signals. We set $|I_k| = L$ and $\delta_k = \delta, k = 1, \ldots, \kappa$, and hence the strength of true signals is controlled by L and δ/σ. We considered two different noise distributions of ϵ including the normal distribution and the t-distribution with degrees of freedom (df). We set $(n, L, \delta) = \{1{,}000, 3{,}000, 5{,}000\} \times \{5, 10\} \times \{1.5, 2.0, 2.5\}$ with $\sigma = 1$ for the normal model, and $(L, df) = \{5, 10\} \times \{10, 5\}$ with $n = 1{,}000$ and $\delta = 3$ for the t-distribution model. The number of true segments was given by $\kappa = n/1{,}000$ and minimum distance between two true segments was set to 200. Numerical performance was evaluated based on 1,000 independent repetitions.

We claim that the signal segment I_k is correctly detected by $\hat{I}_k$ if $I_k \cap \hat{I}_k \neq \phi$ and $|\hat{I}_k| < 2L$. To measure performance of the methods the following two measures are considered.

- Sensitivity: (# of correctly detected signals) / (# of true signals, κ)
- Precision : (# of correctly detected signals) / (# of detected signals)

Sensitivity relates to the ability to identify true signals and precision measures reliability of the detected signals. Note that both measures lie between zero and one (by setting $0/0 = 0$) and a method is perfect if both measures have a value of one.

TABLE 2: Performances under normal error—BWD and LRS outperform CBS and WBS. The BWD with $\alpha = 0.05$ (0.01) shows higher (lower) sensitivity but lower (higher) precision compared to LRS.

	L			5						10					
	δ			1.5		2.0		2.5		1.5		2.0		2.5	
n	Methods			Sen.	Pre.	Sen.	Pre.	Sen.	Pre.	Sen.	Pre.	Sen.	Pre.	Sen.	Pre.
1,000	BWD	.01	cutoff1	.195	.924	.581	.975	.919	.985	.646	.983	.960	.993	.998	.994
			cutoff2	.229	.909	.640	.965	.925	.982	.696	.971	.970	.983	.999	.985
		.05	cutoff1	.335	.819	.727	.910	.952	.933	.777	.914	.983	.939	.999	.945
			cutoff2	.335	.819	.726	.913	.950	.943	.773	.920	.982	.944	.999	.948
	CBS			.165	.948	.555	.975	.900	.979	.648	.972	.972	.976	.999	.977
	WBS			.176	.884	.570	.942	.909	.954	.634	.948	.971	.956	.999	.957
	LRS			.216	.911	.638	.973	.942	.983	.701	.979	.984	.988	1.000	.989
3,000	BWD	.01	cutoff1	.123	.966	.493	.990	.856	.995	.543	.992	.954	.997	.999	.999
			cutoff2	.161	.947	.578	.985	.902	.992	.610	.988	.965	.995	.999	.997
		.05	cutoff1	.229	.911	.653	.972	.926	.982	.700	.972	.980	.982	.999	.988
			cutoff2	.250	.900	.676	.964	.931	.976	.718	.968	.982	.978	1.000	.981
	CBS			.103	.960	.490	.984	.882	.985	.572	.983	.974	.982	.999	.983
	WBS			.079	.967	.399	.991	.819	.990	.470	.991	.938	.992	.999	.990
	LRS			.152	.948	.553	.986	.898	.993	.606	.988	.972	.994	.999	.996
5,000	BWD	.01	cutoff1	.116	.967	.482	.994	.863	.997	.532	.992	.945	.996	.999	.996
			cutoff2	.112	.971	.472	.995	.852	.998	.517	.992	.941	.997	.999	.997
		.05	cutoff1	.212	.936	.632	.981	.924	.991	.667	.981	.974	.989	.999	.991
			cutoff2	.213	.937	.623	.983	.917	.993	.664	.982	.974	.989	.999	.991
	CBS			.088	.973	.468	.989	.890	.987	.551	.983	.966	.985	1.000	.987
	WBS			.052	.985	.327	.994	.730	.995	.383	.994	.891	.995	.997	.995
	LRS			.130	.962	.512	.992	.881	.996	.563	.992	.957	.997	.999	.997

5.1. Gaussian Error

In many applications including CNV detection, the normality assumption is often used. Although our backward procedure does not strongly require the normality assumption, it performs best under normal noise because of the use of squared error loss. Table 2 reports the performance of different methods considered in various scenarios under normality. Both the original and modified versions of BWD outperform the others except LRS in most scenarios considered. This is because the true signal was very short ($L = 5$ and 10). LRS performed quite well. This is not surprising since the designed simulation model perfectly satisfies the assumptions required for LRS. The modification for epidemic change points is useful when the true signals are not strong. TVP was very fast, but gave too many false positives. TGUH showed very low

sensitivity indicating that it cannot detect short signals well. It is interesting to observe that BWD still performed comparably well in the sense that it outperformed LRS in terms of sensitivity with $\alpha = 0.05$ and in terms of precision with $\alpha = 0.01$. It is another benefit of BWD that it

TABLE 3: Performance under t-distributed error—BWD outperforms all other methods for detecting short signals.

L		5				10			
df		10		5		10		5	
Methods		Sen.	Pre.	Sen.	Pre.	Sen.	Pre.	Sen.	Pre.
BWD1	.01	.932	.990	.589	.982	1.000	.993	.968	.983
	.05	.965	.949	.879	.879	1.000	.952	.994	.942
BWD2	.01	.936	.993	.606	.985	1.000	.996	.967	.986
	.05	.969	.952	.876	.893	1.000	.958	.993	.954
CBS		.853	.987	.287	.986	.998	.986	.864	.990
WBS		.981	.867	.939	.644	1.000	.865	.999	.651
LRS		.983	.711	.907	.338	1.000	.741	.999	.375
TVP		.997	.183	.974	.212	1.000	.204	.999	.230
TGUH		.273	.720	.545	.619	.391	.752	.699	.666

controls the relative importance between sensitivity and precision through the target level α. We also remark that BWD is simple and does not require stringent model assumptions such as sparsity.

5.2. Non-Gaussian Error

Performance of the change-point detection methods was evaluated under t-distributed noise. BWD does not require the normality assumption and hence is not overly sensitive to the violation of normality, whereas LRS does. Before applying LRS, we standardized the observations first by using the sample mean and sample standard deviation. We remark that such naive estimates should work fairly well for standardizing observation since the signals are very short compared to the entire sequence of data. Table 3 displays the numerical performance of the methods under consideration. The advantages of BWD are much clearer than in the previous setting with normality. CBS failed to detect true signals when the signal strength is not very strong while the backward procedure performed well in all the scenarios considered. Both WBS and LRS were good in terms of sensitivity but they detect too many false signals in this case. Again, BWD2 outperformed BWD1 when the true signals were not strong.

5.3. Empirical Test Level

We numerically checked whether the backward procedure actually attains a target nominal level α under the null hypothesis that there exists no signal. Since the two versions of BWD show similar results we report the results of only the original version to avoid redundancy. We generated samples under the null hypothesis by letting $\delta = 0$ and reported the proportion of cases for which any signal is detected by each of the methods (Table 4). Recall that we have two scenarios. The first scenario assumes normality and uses "cutoff1" for the cutoff value. Hence the levels are correct if the data are indeed from a normal model but cannot satisfy the nominal

level if the data are from t-distribution. In this case, the 'cutoff2' can be used as an alternative cutoff value and the results seem good enough to be used in practice. Note that there are a couple of cases in which 'cutoff2' failed to produce the nominal level, which was partially caused by

TABLE 4: Estimated level—cutoff1 performs well when the normality assumption is true and cutoff2 can be used if the normality assumption is suspect.

Methods		Normal				t(10)	t(5)
		α	$n = 1{,}000$	$n = 3{,}000$	$n = 5{,}000$	$n = 1{,}000$	$n = 1{,}000$
BWD	cutoff1	0.01	.011	.013	.009	.032	.030
		0.05	.051	.046	.058	.077	.065
	cutoff2	0.01	.014	.018	.010	.007	.002
		0.05	.065	.070	.053	.052	.022
CBS		N/A	.007	.011	.010	.003	.003
WBS		N/A	.016	.005	.004	.112	.395
LRS		N/A	.022	.028	.028	.388	.905
TVP		N/A	.966	.609	.600	.968	.925
TGUH		N/A	.030	.023	.019	.098	.210

the uncertainty about the null distribution. CBS seems very conservative about detecting signal and both LRS and TGUH break down when the normality assumption is not valid. TVP failed again to control the type I error. As mentioned before, it is another distinguishing advantage of the proposed BWD to be able to control type I error. This is practically attractive since the relative importance of sensitivity and precision varies depending on the application.

6. REAL DATA ILLUSTRATION

6.1. Trio Data from an SNP array

The BWD is demonstrated for the SNP array data collected from AGRE (Bucan et al., 2009). The data set contains three parallel sequences of LRR for 547,458 SNPs over 23 chromosomes of a father--mother–offspring trio.

All methods considered in Section 5 were applied except TVP and TGUH. For LRS, the data were standardized by the sample mean and variance. We set $\alpha = 0.05$ and the corresponding cutoff value was approximated from the log-linear relation between cutoff and sample size under the normality assumption as described in Section 4. We applied each of the methods to each chromosome. Figure 3 shows the results for the first two chromosomes (chromosomes 1 and 2) of the offspring. The detect signals are marked as vertical lines. The parameter μ_t estimated by three methods, CBS, LRS and BWD, is indicated by different line types and colours. Note that LRS detected only very short or sparse signals. We would like to point out that although all of CBS, WBS and BWD are developed under a similar framework, the results are quite different. For example in chromosome 2 (Figure 3b), CBS detects no change point after around 54,000th SNP position, while both the BWD and WBS detected several.

The complete CNV detection results for the trio data are summarized by a Venn diagram in Figure 4 which reports the number of CNVs detected by different methods for each of trio (father/mother/offspring) as well as the collapsed data. We consider CNVs to be detected segments whose lengths in terms of the SNP index are between 2 and 200 bp. First, LRS

detected a much larger number of CNVs than other competing methods, while the majority (237/356 = 66.5%) were unique calls which are likely to be suspect as false signals. BWD called 121 CNVs which was more than those from CBS (84) and WBS (100), while the number of

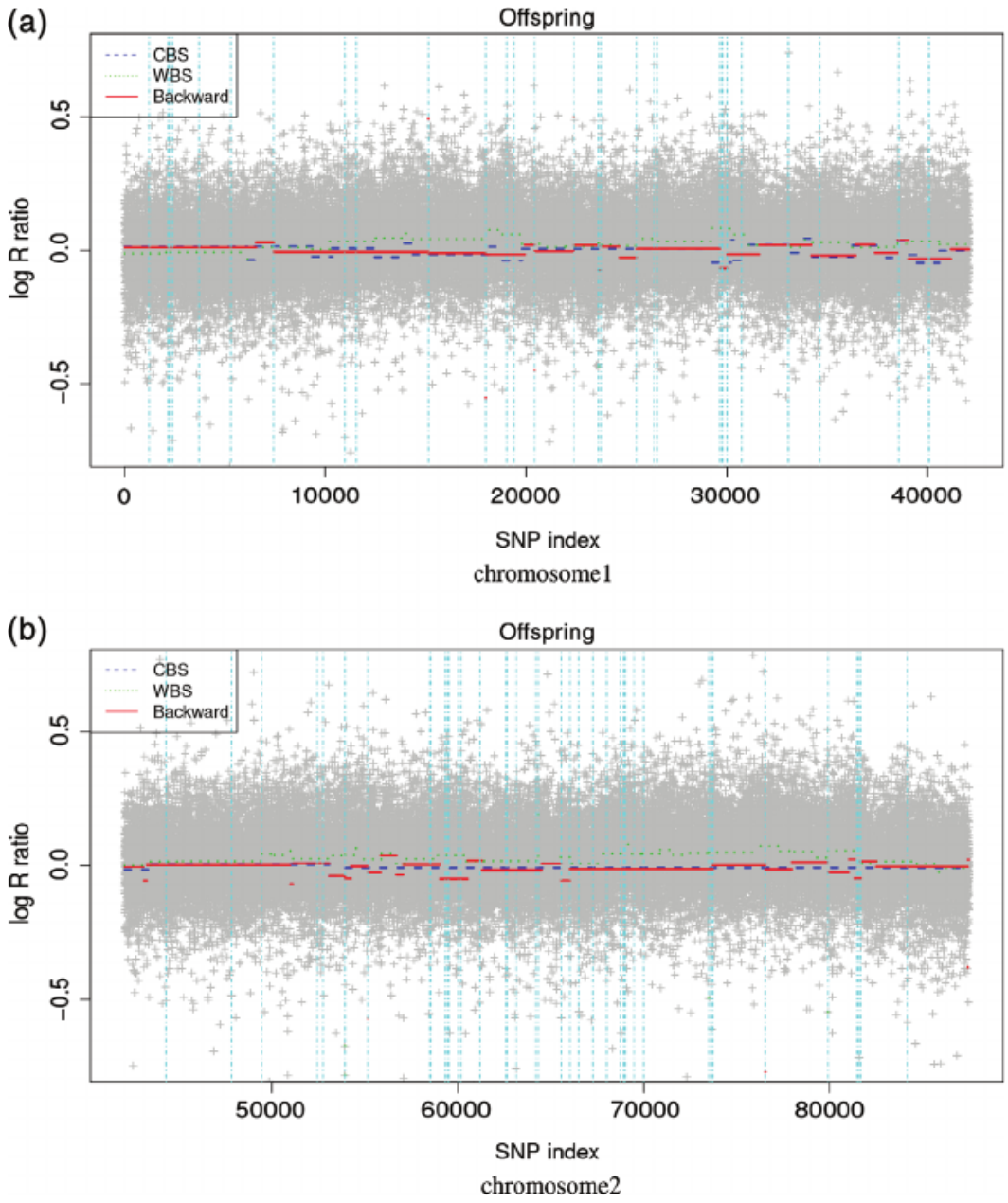

FIGURE 3: CNV detection results for chromosome 1 and 2 of offspring—All of CBS, WBS, and BWD show quite different results. The vertical lines represent the signals detected by LRS.

unique calls by BWD was only 12.4% (15/121), fewer than any of the others (CBS: 34/82 = 41%; WBS: 17/100 = 17%). This can be interpreted that BWD showed the best precision if we assume that most CNVs uniquely called by a single method are false positives. Next, BWD missed only 2 CNVs that were identified by all other methods, while CBS, WBS and LRS missed 24, 8 and 6 such CNVs, respectively, meaning that BWD outperformed others in terms of sensitivity as well. Finally, 25 CNVs identified by all the methods can be regarded as true CNVs and used in Figure 1 in order to show that short CNVs are indeed common in real data.

The genetic information is inherited from parents to offsprings and can be utilized for validation of the detected CNVs. Table 5 lists all the offspring's CNVs that were detected from one of both parents as well. All the CNVs in Table 5 are nearly, if not exactly, identical to the corresponding ones detected from the parents and thus those CNVs are considered as true. We

would like to emphasize that most true CNVs are quite short and both of CBS and WBS miss many of them while LRS and BWD missed only one and three, respectively. We claim that some jointly detected CNVs from (at least one of) parents and offspring are still suspected to be false

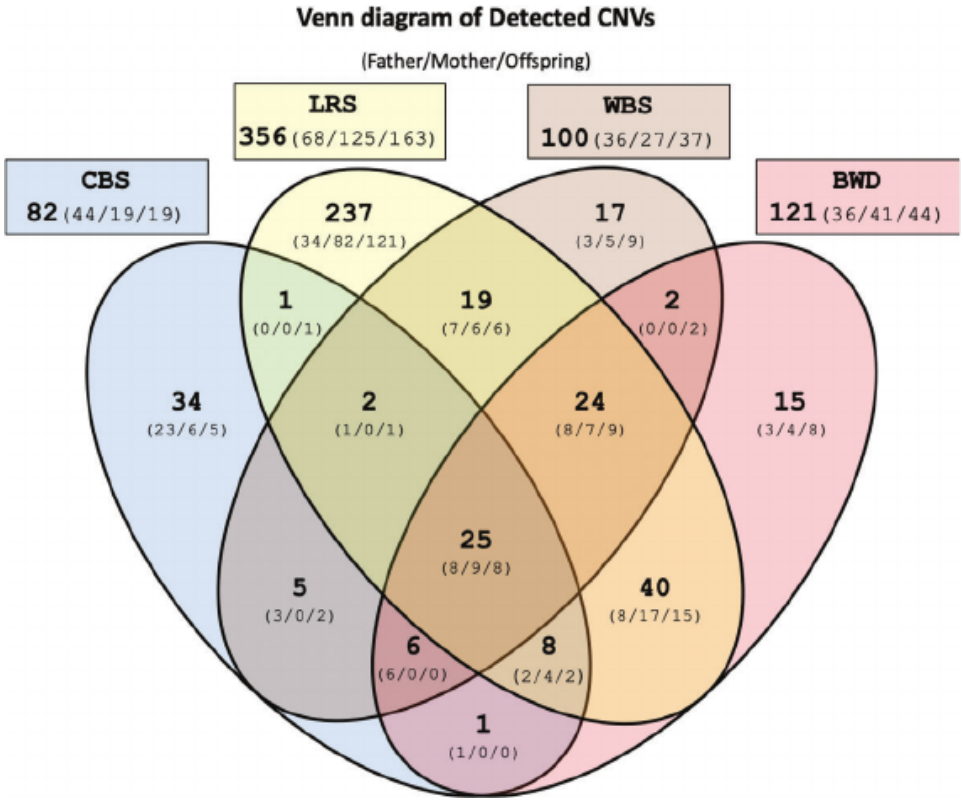

FIGURE 4: Venn diagram of detected CNVs (father/mother/offspring)—BWD identifies the least number of unique CNVs most of which are likely to be false positives, while it missed only 2 CNVs that are identified by all the other methods, while CBS, WBS, and LRS missed 24, 8, and 6 such CNVs, respectively.

if only a very minor portion of the detected CNVs overlapped compared to their entire length. LRS detected nine such suspect CNVs while CBS WBS, and BWD detected one, one and two, respectively.

In summary judging, from the real-data analysis for the trio SNP array, LRS tended to call too many CNVs that included many false positives while CBS and WBS missed some true short CNVs. We can conclude that the proposed BWD outperformed all the others. This is concordant to the findings in the simulation studies in Section 5.

6.2. RD from NGS Sequencing Data

We further illustrate the BWD on the RD data from high-throughtput sequencing data on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1,000 Genomes Project. The RD y_i of the ith locus ($i = 1, \ldots, 54\,361\,060$) was adjusted by the guanine-cytosine content. Although the data can be used to analyze genomic variants in higher resolution with the raw measure, as mentioned earlier the observed values highly fluctuate due to complicated sequencing process and require a proper normalization/transformation. To handle these difficulties, we considered a local-median transformation as motivated by Cai, Jeng & Li (2012). In particular, we first partitioned the RD data into small bins of size M, and then applied BWD to the sequence of the medians of observations in each bin. The transformed data sequence is then well-approximated by a normal distribution regardless of the underlying distribution of the original data. If M is large, the data are more accurately approximated by the normal model,

but a CNV shorter than M bp cannot be accurately identified. (i.e., M is a minimal resolution). As shown in Section 5, BWD is not overly sensitive to violation of the normal assumption and we set a relatively small number of $M = 100$ in the analysis.

TABLE 5: Offspring's CNVs detected from one/both of parents—for each of CNVs, starting SNP indices are presented along with the size in parentheses.

I. With father

Index	CBS		WBS		LRS		BWD	
1			22,369	(7)	22,369	(7)	22,369	(7)
2					76529	(3)	76529	(3)
3	163,327	(18)			163,331	(14)	163,331	(15)
4	252,509	(2)	252,509	(2)	252,509	(2)		
5					325625	(6)		
6	359,377	(11)	359,372	(14)				
7					379,916	(2)	379,916	(2)
8	392,433	(5)			392,433	(5)	392,433	(4)
9	507,037	(6)	507,039	(4)	507,038	(5)	507,038	(122)
10					561,443	(2)	561,443	(2)

II. With mother

Index	CBS		WBS		LRS		BWD	
1					130,814	(10)	130,814	(7)
2					228,119	(6)	228,119	(6)
3					277,328	(2)	277,328	(2)
4	363,744	(9)	363,744	(9)	363,744	(9)	363,744	(9)
5			414,247	(3)	414,247	(3)	414,247	(3)
6					457,597	(3)	457,597	(3)
7			519,555	(10)	519,555	(10)	519,555	(10)
8	532,211	(7)	532,210	(8)	532,211	(7)	532,210	(8)

III. With both father and mother

Index	LRS		BWD		Comments
1	53,949	(2)	53,949	(2)	
2	152,827	(2)	152,827	(2)	
3	359,377	(11)	359,377	(9)	CBS missed father.
4	442,247	(4)	442,243	(8)	WBS missed mother.
5	547,470	(146)	547,459	(157)	WBS missed mother.

For BWD we set $\alpha = 0.05$ and the cutoff value was computed under the normal assumption.

BWD called 15 CNVs. Figure 5 provides zoom-in plots of some of the CNVs identified by BWD. The proposed method worked reasonably well for the NGS read-depth data as well as for the data after a simple transformation.

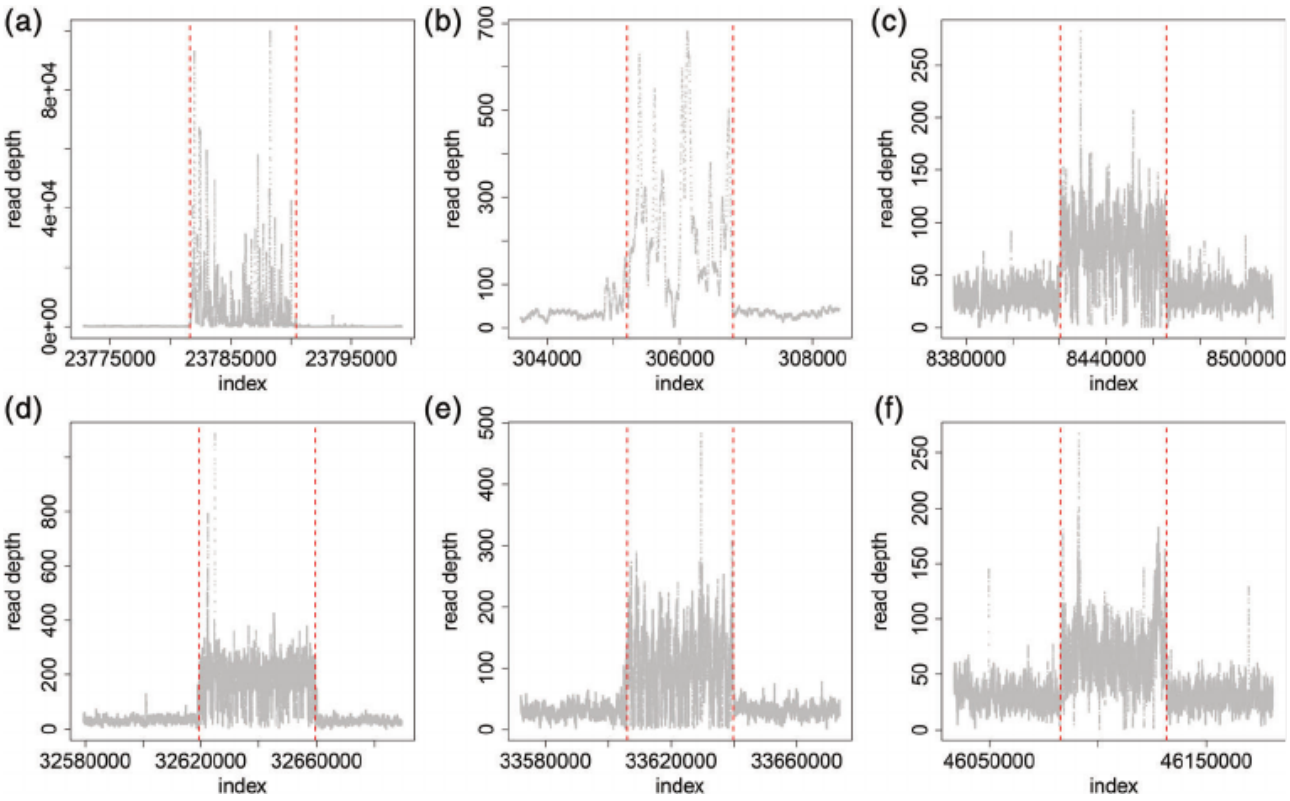

FIGURE 5: Zoom-in plots of the CNVs identified by the backward detection from the NGS read-depth data: The CNVs can be detected directly from the NGS read-depth after simple local median transformation.

Many existing CNV analysis tools for high-throughput NGS data employ CBS as a primary tool for identifying CNVs. We would like to remark that BWD can be a desirable alternative under the presence of short CNVs hardly detected by CBS.

7. DISCUSSION

We propose a BWD procedure for change-point detection and apply it to CNV detection. The proposed BWD is a simple procedure that can be readily employed for high-dimensional data, but it still performs very well, as illustrated with both simulated and real data, especially when the true signals of interest are short, which is often the case in CNV detection problems. Similar to CBS, BWD is a general approach for change-point detection problems that can be used in various applications besides CNV detection from which it was originally motivated, since it does not depend on any application-specific assumptions.

The simple idea of the proposed BWD provides a possibility for further extension in various ways. First, the gain of a backward procedure compared to forward detection, including CBS, is obvious for short signal detection. However, forward detection also has a clear benefit when the true signal is long and the mean change is minor. Thus we can select either of the two depending on application. Moreover, we can develop a method that hybridizes between forward and BWD analogous to stepwise variable selection in a regression context. The idea is straightforward but requires additional effort to improve computational efficiency, especially for CNV applications. Next, we can extend BWD to loss functions other than squared L_2 loss. For example, the absolute deviance error can be used as a reasonable alternative in the presence of outliers. It is also possible to generalize the idea to more complex structures such as a graph (Chen et al., 2015) by

introducing a proper loss function defined on the space of the complex data object. Finally, as motivated by the trio data, the backward idea can be extended to detect common signals shared by multiple sequences of observations.

ACKNOWLEDGEMENTS

We thank two reviewers, an associate editor, and the editor for their most helpful comments that led to substantial improvements in the article. Shin is supported by grants from the National Research Foundation of Korea and Korea University. Wu and Hao are supported by the U.S. National Science Foundation. Hao is also supported by the Simons Foundation.

BIBLIOGRAPHY

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, 21, 974–984.

Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., & Hurles, M. E. (2008). A robust statistical method for case--control association testing with copy number variation. *Nature Genetics*, 40, 1245–1252.

Braun, J. V., Braun, R. K., & Muller, H. G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87, 301–314.

Bucan, M., Abrahams, B. S., Wang, K., Glessner, J. T., Herman, E. I., Sonnenblick, L. I., Retuerto, A. I. A., Imielinski, M., Hadley, D., Bradfield, J. P., Kim, C., Gidaya N. B., Lindquist, I., Hutman, T., Sigman, M., Kustanovich, V., Lajonchere, C. M., Singleton, A., Kim, J., Wassink, T. H., McMahon, W. M., Owley, T., Sweeney, J. A., Coon, H., Nurnberger, J. r., Li, M., Cantor, R. M., Minshew, N. J., Sutcliffe. J.S., Cook, E. H., Dawson, G., Buxbaum, J. D., Grant, S. F. A., Schellenberg, G. D., Geschwind, D. H., & Hakonarson, H. (2009). Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genetics*, 5, e1000536.

Cai, T., Jeng, J., & Li, H. (2012). Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 773–797.

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, 39, S16–S21.

Chen, H., Jiang, Y., Maxwell, K. N., Nathanson, K. L., & Zhang, N. (2017). Allele-specific copy number estimation by whole exome sequencing. *The Annals of Applied Statistics*, 11, 1169.

Chen, H. & Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics*, 43, 139–176.

Chernoff, H. & Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35, 999–1018.

Duan, J., Zhang, J. -G., Deng, H. -W., & Wang, Y. -P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS one*, 8, e59128.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews Genetics*, 7, 85–97.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., Aburatani, H., Jones, K. W., Tyler-Smith, C., Hurles, M. E., Carter. N.P., Scherer, S. W., & Lee, C. (2006). Copy number variation: New insights in genome diversity. *Genome Research*, 16, 949–961.

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 495–580.

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G., & Jain, A. N. (2004). Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis*, 90, 132–153.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42, 2243–2281.

Fryzlewicz, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46, 3390–3421.

Gardner, L. A. (1969). On detecting changes in the mean of normal variates. *The Annals of Mathematical Statistics*, 40, 116–126.

Hao, N., Niu, Y. S., & Zhang, H. (2013). Multiple change-point detection via a screening and ranking

algorithm. *Statistica Sinica*, 23, 1553–1572.
Harchaoui, Z. & Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105, 1480–1493.

Huang, T., Wu, B., Lizardi, P., & Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, 21, 3811–3817.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., & Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12, 105–108.

Jeng, X. J., Cai, T. T., & Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105, 1056–1066.

Killick, R., Fearnhead, P., & Eckley, I. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107, 1590–1598.

International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52–58.

McCarroll, S. A. & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature Genetics*, 39, S37–S42.

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, KW., Rava, R., Daly, MJ., Gabriel, SB., & Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, 40, 1166–1174.

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6, S13–S20.

Niu, Y. S. & Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Annals of Applied Statistics*, 6, 1306–1326.

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5, 557–572.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42, 523–527.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, 44, 248–252.

Peiffer, D., Le, J., Steemers, F., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C., Belmont, J., Cheung, S., Shen, R., Barker, D., & Gunderson, K. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, 16, 1136–1148.

Roy, S. & Reif, A. M. (2013). Evaluation of calling algorithms for array-CGH. *Frontiers in Genetics*, 4, 217.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks J., Ye, K., Reiner, A., Gilliam, TC., Trask, B., Patterson, N., Zetterberg, A., & Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305, 525–528.

Sen, A. & Srivastava, M. S. (1975). On tests for detecting change in mean. *The Annals of Statistics*, 3, 98–108.

Sun, W., Wright, F. A., Tang, Z., Nordgard, S. H., Van Loo, P., Yu, T., Kristensen, V. N., & Perou, C. M. (2009). Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Research*, 37, 5365–5377.

Szatkiewicz, J. P., Wang, W., Sullivan, P. F., Wang, W., & Sun, W. (2013). Improving detection of copy-number variation by simultaneous bias correction and read-depth segmentation. *Nucleic Acids Research*, 41, 1519–1532.

Tibshirani, R. & Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9, 18–29.

Urban, A. E., Korbel, J. O., Selzer, R., Richmond, T., Hacker, A., Popescu, G. V., Cubells, J. F., Green, R., Emanuel, B. S., Gerstein, M. B., Weissman, S. M., & Snyder M. (2006). High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays.

Proceedings of the National Academy of Sciences of the United States of America, 103, 4534–4539.
Venkatraman, E. S. & Olshen, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23, 657–663.

Vostrikova, L. Y. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics Doklady*, 24, 55–59.

Wald, J. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, 17, 1665–1674.

Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., MacAulay, C., Ng, R. T., Brown, C. J., Eichler, E. E., & Lam, W. L. (2007). A comprehensive analysis of common copy-number variations in the human genome. *The American Journal of Human Genetics*, 80, 91–104.

Yao, Q. (1993). Tests for change-points with epidemic alternatives. *Biometrika*, 80, 179–191.

Yao, Y. -C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6, 181–189.

Yao, Y. -C. & Au, S. T. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, 51, 370–381.

Yu, T., Ye, H., Sun, W., Li, K. -C., Chen, Z., Jacobs, S., Bailey, D. K., Wong, D. T., & Zhou, X. (2007). A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics*, 8, 145.

Zhang, N. R. (2010). DNA copy number profiling in normal and tumor genomes. *Frontiers in Computational and Systems Biology*, 259–281.

Zhang, Z., Lange, K., Ophoff, R., & Sabatti, C. (2010). Reconstructing DNA copy number by penalized estimation and imputation. *Annals of Applied Statistics*, 4, 1749–1773.

Received 09 January 2019
Accepted 09 August 2019