Multiple Change-Point Detection: A Selective Overview

Yue S. Niu, Ning Hao and Heping Zhang

Abstract. Very long and noisy sequence data arise from biological sciences to social science including high throughput data in genomics and stock prices in econometrics. Often such data are collected in order to identify and understand shifts in trends, for example, from a bull market to a bear market in finance or from a normal number of chromosome copies to an excessive number of chromosome copies in genetics. Thus, identifying multiple change points in a long, possibly very long, sequence is an important problem. In this article, we review both classical and new multiple change-point detection strategies. Considering the long history and the extensive literature on the change-point detection, we provide an in-depth discussion on a normal mean change-point model from aspects of regression analysis, hypothesis testing, consistency and inference. In particular, we present a strategy to gather and aggregate local information for change-point detection that has become the cornerstone of several emerging methods because of its attractiveness in both computational and theoretical properties.

Key words and phrases: Binary segmentation, consistency, multiple testing, normal mean change-point model, regression, screening and ranking algorithm.

1. INTRODUCTION

Studies of change-point detection problem date back to 1950s (Page, 1954, 1955, 1957). Since then, this topic has been of interest to statisticians and researchers in many other fields such as engineering, economics, climatology, biosciences, genomics and linguistics, just to name a few. In many applications, observed are an ordered sequence of random quantities, from which the change points, that is, positions of structural change are inferred. Examples of such sequences include the daily average temperatures of a specific location over the years, the quantity of some

Yue S. Niu is Assistant Professor and Ning Hao is Assistant Professor, Department of Mathematics, University of Arizona, 617 N. Santa Rita Ave., Tucson, Arizona 85721, USA (e-mail: yueniu@math.arizona.edu; nhao@math.arizona.edu). Heping Zhang is Susan Dwight Bliss Professor of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, Connecticut 06520-8034, USA (e-mail: heping.zhang@yale.edu).

harmful elements, for example, antimony, in the drinking water, stock prices at some time points over a period, and recently, sequencing data in genomics. Depending on the goal of the data that are collected, detection of change points can be crucial for decision making or necessary for understanding certain scientific issues. Many methods have been introduced to detect the change of the mean, variance, slope of regression line, hazard rate, or nonparametric distribution for various models. We refer to the books by Brodsky and Darkhovsky (1993, 2000), Carlstein, Müller and Siegmund (1994), Csörgő and Horváth (1997), Chen and Gupta (2000) for various aspects for classical change-point analysis, and to a recent article Lee (2010) for a list of comprehensive bibliography of books and research papers on this topic. While this body of work constitutes a rich literature, it mainly deals with the inference of a single change in a short or moderate sized sequence. Detecting multiple change points in a very long sequence has emerged as an important problem that has attracted more and more attention recently.

We review both classical and new multiple change-point detection strategies and discuss their strengths

and limitations by examining the general strategies, assessing the computational complexity, and establishing the asymptotic theory. In particular, we present the distinctive characteristic of multiple change-point model from the single change-point model, and give insights on the strategies employed by state-of-the-art change-point detection methods. Some interesting research directions are also discussed. We should point out that there exist a massive number of research papers on change-point detection or closely related topics. We will concentrate mainly on the frequentist approach and narrow our review and discussion to the so-called "posteriori change-point detection" problem according to the terminology of Fryzlewicz (2014). Even with this focus, it is impossible to do justice for all related work, and hence we can offer only a selective overview.

The rest of paper is organized as follows. Section 2 introduces several formulations of multiple change-point model and its distinct features. Section 3 reviews a variety of multiple change-point detection techniques. Section 4 focuses on theoretical properties of these methods. Section 5 summarizes new strategies used by some recent methods. We end this paper with some concluding remarks in Section 6.

2. MULTIPLE CHANGE-POINT PROBLEM

2.1 From Classical to Modern Data

Two datasets, Nile river data (Cobb, 1978) and British coal mine disaster data (Carlin, Gelfand and Smith, 1992), plotted in Figure 1(a) and (b), respectively, have been commonly used in classical change-point analyses. The main focus of those analyses was to test whether there was a change point along the sequence. Work on detection for multiple change points began in the 1980s (e.g., Vostrikova, 1981, Yin, 1988, Yao, 1988, Yao and Au, 1989), although the applications seemed limited during that decade. The advent of high-throughput technologies has produced high dimensional data that are of great interest in statistical sciences, and not surprisingly, revitalized the change-point analysis. In Figure 2, we plot a time series data set and a SNP array data set, which represent typical modern applications of multiple change-point model. A key difference between Figures 1 and 2 is the number of data points.

2.2 A Variety of Multiple Change-Point Model Formulations

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ be a sequence of independent random variables (or vectors) with probability distribution functions $F_1, \ldots, F_n$, respectively. A change point
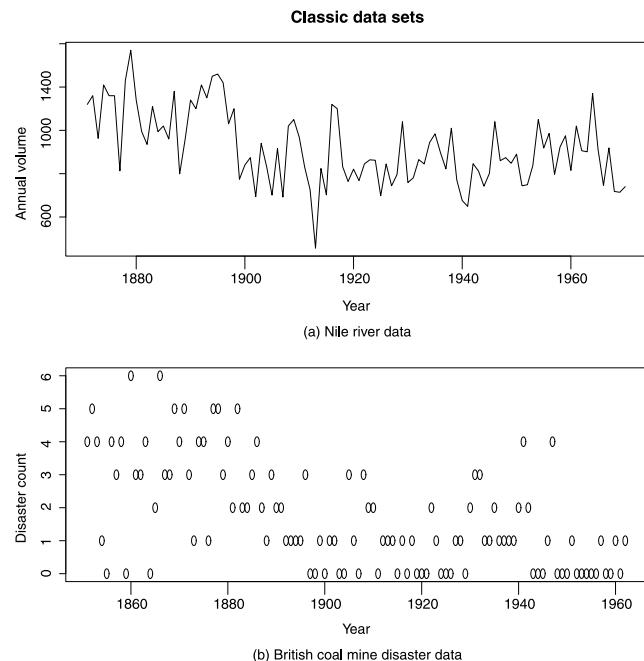

FIG. 1. *Two classical data sets.* (a) *Measurements of the annual flow of the river Nile at Ashwan* 1871–1970. (b) *Yearly counts of British coal-mining disasters between* 1851–1962.
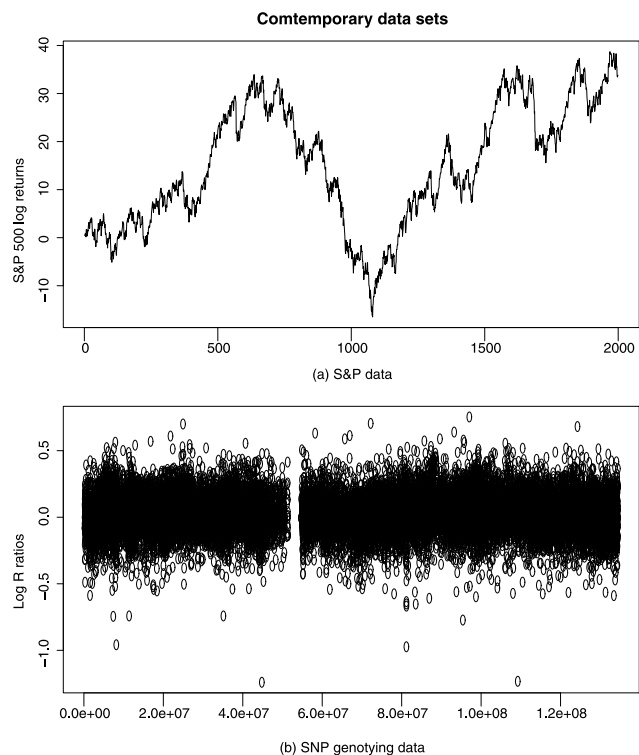

FIG. 2. *Two contemporary data sets.* (a) *Log-returns on the daily closing values of S&P* 500 *of length* 2000, *ending* 26 *October* 2012. (b) *Log R ratios of length* 27272 *along Chromosome* 11 *of the subject father from a SNP genotying data set for a father-mother-offspring trio produced by Illumina* 550K *platform, downloaded from* http://www.openbioinformatics.org/penncnv/.

occurs at τ when $F_\tau \neq F_{\tau+1}$. The goal is to estimate the number of change points, J, and location vector of change points $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)^\top$. We use the convention $\tau_0 = 0$ and $\tau_{J+1} = n$ throughout this paper. Depending on applications, we may assume that the distributions $\{F_i\}_{i=1}^n$ belong to a common parametric family or keep the setting nonparametric.

The simplest case where $J \leq 1$ has been studied extensively. It is usually formulated as a hypothesis testing problem:

$$H_0 : F_1 = F_2 = \cdots = F_n \quad \text{versus}$$

$$(1) \qquad H_1 : F_1 = \cdots = F_\tau \neq F_{\tau+1} = \cdots$$

$$= F_n \quad \text{with unknown } \tau.$$

The first step is to test hypotheses $J = 0$ versus $J = 1$, and if we reject $J = 0$, the next step is to make inference on the location parameter τ (Hawkins, 1977).

Both steps turned out to be much more complicated than they appeared. For example, even if $\{F_i\}_{i=1}^n$ are Gaussian with equal variance, the exact distribution of the likelihood ratio test statistic under H_0 is formidable, and usually approximated by asymptotic or numerical methods (Sen and Srivastava, 1975, Hawkins, 1977, Csörgő and Horváth, 1997). The inference on τ has been regarded as a difficult problem, and studied in Hinkley (1970), Worsley (1986), Siegmund (1988) for normal mean model and one-parameter exponential families.

Concerning $J > 1$, a normal mean change-point model is typically employed, and here we focus on this model. First of all, it is the most basic multiple change-point model (MCM). In fact, even under this basic model, there are many challenging and open problems. Second, despite the simplicity, it is useful in many applications including genomics and econometrics; see Braun, Braun and Müller (2000), Olshen et al. (2004), Zhang and Siegmund (2007), Niu and Zhang (2012), Frick, Munk and Sieling (2014), Fryzlewicz (2014), among many others. Third, detecting changes in a sequence can often be reduced to the detection of mean changes in another derived sequence (Carlstein, Müller and Siegmund, 1994, Brodsky and Darkhovsky, 2000). Last but not least, it serves as a benchmark model for comparing different methods of detecting change points (Fryzlewicz, 2014). Next, we discuss a few formulations of the normal mean MCM.

2.2.1 *Normal mean multiple change-point model.*
Consider a sequence of independent random variables $Y_1, \ldots, Y_n$ with $Y_i \sim \mathcal{N}(\theta_i, \sigma^2)$ and the mean parameter shifts at J unknown locations as introduced above.

The normal mean MCM can be presented as

$$Y_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), 1 \leq i \leq n, \text{ with}$$

$$(2) \qquad \theta_1 = \theta_2 = \cdots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \cdots = \theta_{\tau_2} \neq \theta_{\tau_2+1}$$

$$= \cdots = \theta_{\tau_J} \neq \theta_{\tau_J+1} = \cdots = \theta_n.$$

In other words, we assume that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ is piecewise constant with jumps or drops at $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)^\top$.

Besides the piecewise structure, the change points are usually assumed to be sparsely located along the sequence. Roughly speaking, it means that J is much smaller than n and change points are not too close to each other. Therefore, we may view (2) as a high dimensional sparse model. Moreover, its natural sequential structure is a distinct feature and crucial in developing efficient algorithms.

2.2.2 *A regression model.* Let

$$\beta_0 = \theta_1, \quad \beta_j = \theta_{j+1} - \theta_j; \quad j = 1, \ldots, n-1.$$

Then model (2) is transformed to a linear model,

$$(3) \qquad Y_i = \sum_{j=0}^{i-1} \beta_j + \varepsilon_i, \quad i = 1, \ldots, n,$$

or in matrix form

$$(4) \qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{n-1})^\top$ is coefficient vector with J nonzero entries besides the intercept β_0 and the design matrix $\mathbf{X} = (\mathbf{X}_0, \ldots, \mathbf{X}_{n-1})$, with $\mathbf{X}_j = (\mathbf{0}_j^\top, \mathbf{1}_{n-j}^\top)^\top$. Here, $\mathbf{1}_k$ and $\mathbf{0}_k$ are k-dimensional column vectors with equal entries 1 and 0, respectively. We see that $\mathbf{X}$ is simply a lower triangular matrix with all nonzero elements equal to 1. Model (3) is sparse when J is much smaller than n.

We can observe from

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

that the columns of $\mathbf{X}$ spans the space of all piecewise constant vectors of length n. Note that each column, say, $(\mathbf{0}_j^\top, \mathbf{1}_{n-j}^\top)^\top$, has only one discontinuity at j. Therefore, the piecewise constant $\boldsymbol{\theta}$ is a linear combination of piecewise constant vectors with a single discontinuity.

With this formulation, modern high dimensional sparse regression techniques can be applied to solve change-point problem directly; see, for example, Huang et al. (2005), Tibshirani and Wang (2008), Rinaldo (2009), Zhang et al. (2010), Harchaoui and Lévy-Leduc (2010), Qian and Jia (2012).

2.2.3 *Multiple hypothesis testing.* We can generalize the hypothesis testing problem for a single change in (1) to the multiple change-point setting as

(5)
$$H_0 : \text{there is no change point} \quad \text{versus}$$
$$H_1 : \text{at least one change point.}$$

Unfortunately, the alternative hypothesis above is too broad to be useful in practice. To be specific about each location as a candidate of change, it becomes a multiple testing problem which tests whether or not each of individual positions is a change point. In other words, we test

(6)
$$H_0(j) : j \text{ is a not change point} \quad \text{versus}$$
$$H_1(j) : j \text{ is a change point,}$$

where $j = 1, 2, \ldots, n-1$. In particular, for the normal mean model, $H_0(j)$ and $H_1(j)$ correspond to $\beta_j = 0$ and $\beta_j \neq 0$ in model (3), respectively. This clearly involves a large number of tests. The existing methods and theory on the control of false discover rate (FDR, Benjamini and Hochberg, 1995) can be applied if we ignore the sequential structure in the data. However, if we want to take advantage of the sequential structure, this is not a typical multiple testing problem. Hao, Niu and Zhang (2013) attempted to address this problem by generalizing the concept of FDR and testing a series of window-shifting hypotheses:

(7)
$$H_0(j) : F_{j+1-h} = \cdots = F_{j+h} \quad \text{versus}$$
$$H_1(j) : F_{j+1-h} = \cdots = F_j \neq F_{j+1} = \cdots$$
$$= F_{j+h},$$

where h is a fixed integer with the assumption that any adjacent change points must be least h points apart. Based on our theoretical analysis, the optimal choice of h depends on the minimal distance between adjacent change points and typically grows at a slower rate than n. In practice, we may determine h when prior knowledge on the distances among the change points is available. For the normal mean change-point model, (7) is reduced to

(8)
$$H_0(j) : \theta_{j+1-h} = \theta_{j+2-h} = \cdots = \theta_{j+h}; \quad \text{versus}$$
$$H_1(j) : \theta_{j+1-h} = \cdots = \theta_j \neq \theta_{j+1} = \cdots = \theta_{j+h}.$$

Note that the alternative hypothesis in (7) is more specific than that in (6).

2.3 Locality and Symmetry

Relative to the single change point problem, MCM involves the unknown J which creates additional difficulties to the case when J is known. MCM is also distinct from the single change point problem in two other aspects.

The first one is the *local nature* of the MCM. Consider a simple normal mean model with three change points at $\tau_1 < \tau_2 < \tau_3$ and assume that the error distribution is known. In order to infer τ_2, the best way is to use only the data between τ_1 and τ_3 as other data are irrelevant. In practice, the error distribution is unknown, and the data outside the interval $(\tau_1, \tau_3]$ may be helpful only to learn the error distribution and find a threshold or stopping rule for some detection procedure. On the other hand, misuse of the information outside of the interval $(\tau_1, \tau_3]$ may bias the inference on τ_2. Therefore, ideally, an oracle procedure would divide multiple change-point problem (2) into J simple and local subproblems, which are single change-point problems over the segments $(0, \tau_2], (\tau_1, \tau_3], \ldots, (\tau_{J-1}, n]$, respectively. Although such a division is not feasible in practice, it is still helpful to mimic the oracle by using local information.

The second one is *symmetry*. As in Olshen et al. (2004), we can connect the head and tail of the sequence to make it a circle. Now, the MCM becomes symmetric. As a result, there is no boundary effect and all locations play the same role. We should be cautious with this technique when J is small, such as the single change point problem. The reason is that we might have a priori knowledge on whether there is a change point where the head and tail connect. We may lose such information using this circular model. Nevertheless, in many applications it is harmless to consider this symmetric model. With this symmetry, it is interesting to study equivariant detection tools.

Both the difficulties and uniqueness of MCM make it a more interesting and challenging problem to investigate than the single change-point problem.

2.4 Signal Strength Levels of MCM

Change itself is a relative concept. Therefore, it is necessary for us to impose certain reasonable assumptions to avoid nonidentifiability. Intuitively, there are two situations where it is difficult, if not impossible to detect a change point. The first case is when the mean shifts so small that we cannot distinguish between a real change or an effect of noise. The second scenario is when there are two (or more) change points

so close to each other that we do not have enough data to draw inference. Therefore, it is reasonable to measure the overall signal strength of normal mean MCM by $S = \delta^2 L$, where $L = \min_{0 \leq \kappa \leq J}\{\tau_{\kappa+1} - \tau_\kappa\}$ is the minimal distance between change points and $\delta = \min_{1 \leq i < n}\{|\theta_i - \theta_{i+1}|\}/\sigma$ is the ratio of the minimal shift to the noise level. We present in Table 1 some commonly used technical conditions. As we will see, the existing methods work best when these assumptions are satisfied.

Early work such as Yao (1988), Yao and Au (1989) usually assumed that $\boldsymbol{\tau}/n \to \mathbf{t}$, a constant vector, and the mean shifts at change points and the noise variance σ^2 remain constants as $n \to \infty$. Under these assumptions, $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\|_\infty = O_P(1)$, which is the optimal result (unless $\delta \to \infty$). An equivalent formulation is to assume that there is a fixed piecewise constant function on the interval $[0, 1]$ to be estimated as the number of observations $n \to \infty$. This result extended the theory from the single change-point case to the multiple change-point case, but may not be applicable to some modern applications when the sequence is long and some change points are close to each other. Rinaldo (2009), Qian and Jia (2012) used regression approach and obtained the consistency result $\mathbf{P}(\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}) \to 1$. But they required $\delta^2 \gg \log n$, which may be unrealistic in practice. On the other hand, Arias-Castro, Donoho and Huo (2005) showed that $\delta^2 L \geq 2 \log n$ is a necessary condition to recover all change points even for an MCM with two change points. Hao, Niu and Zhang (2013) further showed that $\delta^2\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\|_\infty = O_P(1)$ under a slightly relaxed condition $\delta^2 L \geq 32 \log n$. It remains an open problem to find weakest possible sufficient and necessary conditions on $\delta^2 L$ for full recovery of all change points. When the signal is not strong enough for us to recover all change points, it is desirable to control the false positives. As discussed by Hao, Niu and Zhang (2013), it is challenging to establish a reasonable framework for this atypical multiple testing problem.

3. MULTIPLE CHANGE-POINT DETECTION TOOLS

For the multiple change-point problem, the main goal is to estimate the number of change points and their locations. In this section, we review some classical and new approaches to identifying change points. From now on, we use J^* and $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_J^*)^\top$ to denote the true number of change points and their location (in ascending order) vector of a specific data generating process.

3.1 Exhaustive Search

Ignoring its computational complexity, an exhaustive search among all possibilities $0 \leq J \leq n - 1$ and $0 < \tau_1 < \cdots < \tau_J < n$ can be applied. Define $\Gamma_J = \{\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J) : 0 < \tau_1 < \cdots < \tau_J < n\}$ the set of all possible ordered J-dimensional vectors representing locations of J change points. For any J, define

$$\hat{\sigma}_J^2 = \min_{\boldsymbol{\tau} \in \Gamma_J} \hat{\sigma}_{\boldsymbol{\tau}}^2,$$

where $\hat{\sigma}_{\boldsymbol{\tau}}^2$ is the maximum likelihood estimator for variance conditional on the change-point location $\boldsymbol{\tau}$. Yao (1988) showed the consistency property of the estimator $\hat{J}$ determined by Bayesian Information Criterion (BIC)

$$(9) \qquad \hat{J} = \operatorname*{argmin}_J \frac{n}{2} \log \hat{\sigma}_J^2 + J \log n.$$

Furthermore, Yao and Au (1989) showed the convergence rate of the location estimator

$$(10) \qquad \hat{\tau}_\kappa - \tau_\kappa^* = O_P(1), \quad 1 \leq \kappa \leq J^*,$$

where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_{J^*})^\top = \operatorname{argmin}_{\boldsymbol{\tau} \in \Gamma_{J^*}} \hat{\sigma}_{\boldsymbol{\tau}}^2$, and J^* is the true number of change points, which can be estimated consistently by (9).

Exhaustive search among all possible subsets is not efficient from a computational point of view, and too intensive when n is large. Making use of the sequential structure, dynamic programming techniques can be applied to reduce the computational burden down to

TABLE 1
Different signal strength levels of MCM (2) *used in the literature. Here, $a_n \asymp b_n$ means $0 < \frac{1}{M_1} < \frac{a_n}{b_n} < M_2 < \infty$ as $n \to \infty$ for some constants M_1, M_2. $\| \cdot \|_\infty$ is infinity norm*

Signal strength level	Typical assumption	Typical conclusion
S1. very strong on L	(A1) $L \asymp n$, $\delta \asymp 1$	(C1) $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\|_\infty = O_P(1)$
S2. very strong on δ	(A2) $\delta^2 \gg \log n$	(C2) $\mathbf{P}(\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}) \to 1$
S3. strong	(A3) $\delta^2 L \asymp \log n$	(C3) $\delta^2\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\|_\infty = O_P(1)$
S4. weak	(A4) N/A	(C4) FDR control

the order of $O(n^2)$ (Braun, Braun and Müller, 2000, Jackson et al., 2005). Killick, Fearnhead and Eckley (2012) further reduced the complexity to $O(n)$, under, however, an assumption which may not be practical in applications (Fryzlewicz, 2014).

3.2 Stepwise Selection

Stepwise selection is a popular substitute for the exhaustive search with reduced complexity and simple implementation. Forward and backward selection methods are well-known stepwise procedures in model selection. Both of them can be applied to the change-point problem. In particular, the forward selection method, which is called binary segmentation (BS), has been studied for a long time (Scott and Knott, 1974, Vostrikova, 1981). Recently, circular binary segmentation (CBS, Olshen et al., 2004) and wild binary segmentation (WBS, Fryzlewicz, 2014) have been proposed enhance the power of BS in identifying short segments. In contrast, the backward selection procedure has not been widely used for change point detection with the exception of Shin, Wu and Hao (2014).

3.2.1 *Forward selection approach.* We now introduce the binary segmentation algorithm (Scott and Knott, 1974, Vostrikova, 1981). It is a forward stepwise method with the following steps:

Step 1: Test for no change point versus one change point (1). If H_0 is not rejected, stop. Otherwise, estimate the first change point $\hat{\tau}$, which divides the whole sequence into two segments;

Step 2: Test these two segments, respectively, for further segmentation;

Step 3: Repeat the procedure for each segment until no one can be segmented further.

We illustrate here a special BS algorithm for solving model (2) with known variance $\sigma^2 = 1$ based on the likelihood ratio method. BS tests the hypotheses recursively in each segment. The following is for the initial segment and the rest is similar:

$$H_0 : \theta_1 = \cdots = \theta_n, \quad \text{versus}$$

$$(11) \qquad H_1 : \theta_1 = \cdots = \theta_j \neq \theta_{j+1} = \cdots$$

$$= \theta_n \quad \text{for some } 1 \leq j < n.$$

A likelihood ratio type statistics

$$(12) \qquad T_1 = \max_{1 \leq j \leq n-1} (-2 \log \Lambda_j)$$

can be used (Sen and Srivastava, 1975), where

$$(13) \qquad \begin{aligned} -2 \log \Lambda_j &= (\bar{Y}_{j+} - \bar{Y}_{j-})^2 \\ &\quad /[j^{-1} + (n-j)^{-1}], \end{aligned}$$

and $\bar{Y}_{j-} = \sum_{k=1}^{j} Y_k / j$ and $\bar{Y}_{j+} = \sum_{k=j+1}^{n} Y_k / (n-j)$. Moreover, when the alternative is supported, the segment can be divided into two parts by $\hat{j}$, obtained from

$$(14) \qquad \hat{j} = \underset{1 \leq j \leq n-1}{\operatorname{argmax}} (-2 \log \Lambda_j).$$

The distribution of T_1 under H_0 can be approximated numerically or asymptotically. We refer to Csörgő and Horváth (1997) for asymptotic properties of T_1 and $\hat{j}$.

The main drawback of the BS algorithm is that it can rarely detect short segments embedded in the middle of long segments (Olshen et al., 2004, Fryzlewicz, 2014, Shin, Wu and Hao, 2014). To increase the power for the BS algorithm in detecting short segments, Olshen et al. (2004) proposed the Circular Binary Segmentation (CBS) algorithm. The main idea is to splice two ends of the segment to make it a circle and check whether there exists a segment, say between l and r, such that its mean is significantly different from the mean of the remaining part. In particular, they considered and applied recursively a test of no change against a so-called epidemic alternative, as described below.

$$H_0 : \theta_1 = \cdots = \theta_n, \quad \text{versus}$$

$$(15) \quad H_1 : \theta_1 = \cdots = \theta_l = \theta_{r+1} = \cdots = \theta_n$$

$$\neq \theta_{l+1} = \cdots = \theta_r \quad \text{for a pair } l < r.$$

Here, the alternative hypothesis is called epidemic in analogy to an epidemic running from l through r after which the normal state is restored (Levin and Kline, 1985, Yao, 1993). Similar to (13), we can calculate the log likelihood ratio for a fixed pair (l, r),

$$(16) \qquad \begin{aligned} -2 \log \Lambda_{l,r} &= (\bar{Y}_I - \bar{Y}_O)^2 \\ &\quad /[(r-l)^{-1} + (n-r+l)^{-1}], \end{aligned}$$

where the outside mean $\bar{Y}_O = \sum_{k \leq l \text{ or } k > r} Y_k / (n - r + l)$ and inside mean $\bar{Y}_I = \sum_{k=l+1}^{r} Y_k / (r - l)$.

Therefore, one may use

$$T_2 = \max_{1 \leq l < r \leq n} (-2 \log \Lambda_{l,r})$$

as a test statistic for problem (15).

Since the CBS algorithm considers all pairs of points when calculating the test statistic, it is more powerful in detecting the short segments and has become one of the most popular algorithms in some applications

such as copy number variation detection. However, it is computationally more expensive than BS.

Recently, Fryzlewicz (2014) proposed the wild binary segmentation (WBS) to improve BS algorithm. Instead of the global test (11), WBS considers the same test on a random segment $[s, e]$, that is,

$$H_0 : \theta_s = \cdots = \theta_e, \quad \text{versus}$$

$$(17) \qquad H_1 : \theta_s = \cdots = \theta_j \neq \theta_{j+1} = \cdots$$

$$= \theta_e \quad \text{for some } s \leq j < e.$$

The test statistic $T_1^{[s,e]}$ and $\hat{j}^{[s,e]}$ are defined similar to T_1 and $\hat{j}$ in (12) and (14). WBS first draws random intervals $[s_m, e_m] \subset [1, n]$, $m = 1, \ldots, M$ and defines the first change point as

$$\hat{j}_{\text{WBS}} = \hat{j}^{[s_{\hat{m}}, e_{\hat{m}}]}, \quad \text{where } \hat{m} = \underset{1 \leq m \leq M}{\text{argmax}} \, T_1^{[s_m, e_m]}.$$

Then WBS repeats the same procedure on each segment until some stopping rule is met. WBS offers one way to localize the BS procedure using local and random intervals instead of the whole segment in segmentation.

3.2.2 *Backward selection approach*. Shin, Wu and Hao (2014) considered a backward detection (BWD) procedure. It starts with n groups, each of which contains only one data point, and then merges groups based on some criterion until a stopping time. This is similar to the bottom-up clustering analysis where small clusters in the lower levels are joined to form larger ones in the upper levels. Specifically:

1. Start with a superset $\mathbb{G}^{(0)} = \{\mathcal{G}_1^{(0)}, \mathcal{G}_2^{(0)}, \ldots, \mathcal{G}_n^{(0)}\}$ where $\mathcal{G}_i^{(0)} = \{i\}$.

2. At step k given $\mathbb{G}^{(k-1)} = \{\mathcal{G}_1^{(k-1)}, \mathcal{G}_2^{(k-1)}, \ldots, \mathcal{G}_{n-k+1}^{(k-1)}\}$, a decision is made to merge two consecutive groups in $\mathbb{G}^{(k-1)}$. Say, if $\mathcal{G}_i^{(k-1)}$ and $\mathcal{G}_{i+1}^{(k-1)}$ are merged to one group, then in $\mathbb{G}^{(k)}$,

$$\mathcal{G}_j^{(k)} = \begin{cases} \mathcal{G}_j^{(k-1)}, & \text{if } j < i; \\ \mathcal{G}_j^{(k-1)} \cup \mathcal{G}_{j+1}^{(k-1)}, & \text{if } j = i; \\ \mathcal{G}_{j+1}^{(k-1)}, & \text{if } j > i. \end{cases}$$

3. Either stop if a stopping criterion is met; or complete the whole process and then determine a best model along the sequence $\mathbb{G}^{(k)}$, $0 \leq k \leq n$.

For the normal mean model (2), Shin, Wu and Hao (2014) employed a dissimilarity index to decide which pair of segments to merge in each step k. The dissimilarity index between two groups is defined as

$$(18) \quad \begin{aligned} \text{DI}_j^{(k-1)} &= \text{DI}(\mathcal{G}_j^{(k-1)}, \mathcal{G}_{j+1}^{(k-1)}) \\ &= \frac{|\bar{Y}_j^{(k-1)} - \bar{Y}_{j+1}^{(k-1)}|}{\sqrt{|\mathcal{G}_j^{(k-1)}|^{-1} + |\mathcal{G}_{j+1}^{(k-1)}|^{-1}}}, \end{aligned}$$

where $\bar{Y}_j^{(k-1)} = \frac{1}{|\mathcal{G}_j^{(k-1)}|} \sum_{m \in \mathcal{G}_j^{(k-1)}} Y_m$, and $|\mathcal{G}_j^{(k-1)}|$ is the cardinality (i.e., size) of the set $\mathcal{G}_j^{(k-1)}$. Thus, for $i = \text{argmin}_j \, \text{DI}_j^{(k-1)}$, $\mathcal{G}_i^{(k-1)}$ and $\mathcal{G}_{i+1}^{(k-1)}$ are merged at step k. In other words, at each step, BWD merges two consecutive groups with the minimal dissimilarity index. In fact, (18) is essentially the log likelihood ratio test statistic for

$$H_0 : \mathcal{G}_j^{(k-1)} \text{ and } \mathcal{G}_{j+1}^{(k-1)} \text{ have the same mean,}$$

$$(19) \qquad \text{versus}$$

$$H_1 : \mathcal{G}_j^{(k-1)} \text{ and } \mathcal{G}_{j+1}^{(k-1)} \text{ have different mean.}$$

3.2.3 *Summary*. The pros and cons of forward and backward stepwise procedures are analogous to those in model selection. In practice, there is little harm in using both and then comparing results. The BS algorithm takes $O(n)$ operations for each step, and $O(\log n)$ steps in the worst scenario. So the total complexity is $O(n \log n)$. The CBS optimizes over all pairs of points so the complexity is increased by an order of magnitude. In general, BS and CBS are easy to implement, and the early stopping rule can be applied to accelerate the procedures (Venkatraman and Olshen, 2007). BWD is also easy to implement and of complexity $O(n \log n)$ (Shin, Wu and Hao, 2014).

3.3 ℓ_1-Penalization

We discussed in Section 2.2.2 that detecting change points is essentially a regression problem with sparsity. Therefore, it is not surprising that there have been attempts to use the methods from penalized regression to the detection of change points. Huang et al. (2005), Harchaoui and Lévy-Leduc (2010) studied the following optimization problem:

$$\text{minimize } \|\mathbf{Y} - \boldsymbol{\theta}\|^2$$

$$(20) \qquad \text{subject to } \sum_{j=1}^{n-1} |\theta_j - \theta_{j+1}| \leq s.$$

Through reparametrization $\beta_j = \theta_{j+1} - \theta_j$, the above optimization problem is equivalent to

$$(21) \quad \text{minimize } \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{subject to } \sum_{j=1}^{n-1} |\beta_j| \leq s,$$

which gives the LASSO solution for regression model (4). (20) and (21) are usually solved through their dual forms

$$\text{minimize } \|\mathbf{Y} - \boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{n-1} |\theta_j - \theta_{j+1}|$$

$$\text{or} \quad \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{n-1} |\beta_j|,$$

where $\lambda \sum_{j=1}^{n-1} |\theta_j - \theta_{j+1}|$ is called the total variation penalty, $\lambda \sum_{j=1}^{n-1} |\beta_j|$ is called ℓ_1 penalty or LASSO penalty.

Furthermore, Tibshirani and Wang (2008) applied the fused LASSO (Tibshirani et al., 2005) for change-point detection. In particular, they estimated $\boldsymbol{\theta}$ through the following constrained optimization:

$$(22) \quad \begin{aligned} &\text{minimize } \|\mathbf{Y} - \boldsymbol{\theta}\|^2 \\ &\text{subject to } \|\boldsymbol{\theta}\|_{\ell_1} \le s_1, \sum_j |\theta_j - \theta_{j+1}| \le s_2, \end{aligned}$$

or equivalently,

$$\text{minimize } \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\text{subject to } \sum_{j=1}^{n} \left| \sum_{k=0}^{j-1} \beta_k \right| \le s_1, \sum_{j=1}^{n-1} |\beta_j| \le s_2.$$

The difference between (20) and (22) lies in whether $\boldsymbol{\theta}$ is also assumed to be sparse. In the context of detecting chromosome copy number variations, it is reasonable to assume that $\boldsymbol{\theta}$ is sparse because of the discrete nature of the copy numbers. From the methodological and computational perspectives, the solution for (22) can be obtained by simply thresholding the solution of (20) (Friedman et al., 2007). Hence, there is little loss of generality by focusing on (20) and (21).

For the ℓ_1 penalization methods, standard regression solvers such as the LARS (Efron et al., 2004) and coordinate decent algorithm (Friedman et al., 2007) can be used to solve optimization problem (21). The complexity of some related change-point detection algorithms can achieve $O(K_{\max} n \log n)$, where $K_{\max}$ is an upper bound for the number of change points (Harchaoui and Lévy-Leduc, 2010).

It is crucial to determine the values of tuning parameters in these procedures. Some tuning parameter selection methods have been used in the literature. In particular, Huang et al. (2005) proposed to choose s empirically. They examined the solutions of (20) for an increasing sequence of s. As more change points are

added into the model with larger values of s, they chose to stop increasing s when the resulted mean difference at the new change point is not big enough. Tibshirani and Wang (2008) estimated s_1 and s_2 for (22) based on heavily and moderately smoothed versions of $\mathbf{Y}$, respectively. Harchaoui and Lévy-Leduc (2010) gave an asymptotic order of the tuning parameter λ in the dual optimization problem. Friedman et al. (2007) recommended cross-validation to select tuning parameters for the fused LASSO. As stated in Rinaldo (2009), it is an important open problem to find the optimal values for the parameters.

3.4 Screening and Ranking Algorithm

A Screening and Ranking algorithm (SaRa) has been studied in Niu and Zhang (2012), Hao, Niu and Zhang (2013) to detect change points. For the normal mean model (2), they considered the locally defined statistic at each position $h \le j \le n - h$,

$$D_h(j) = \left(\sum_{k=j+1}^{j+h} Y_k - \sum_{k=j-h+1}^{j} Y_k \right) \Big/ h,$$

where h is a fixed bandwidth. Essentially, $D_h(j)$ is the likelihood ratio test statistic for the local testing problem (8). Hence, the sequence $\{D_h(\cdot)\}$ roughly measures the relative likelihood for each position to be a change point. The SaRa proceeds as follows. First, it calculates $D_h(\cdot)$ and finds all local maximizers of $|D_h(\cdot)|$. Here, j is a local maximizer if $|D_h(j)| \ge |D_h(k)|$ for all $k \in (j - h, j + h)$. Second, the SaRa estimator is obtained by applying a thresholding rule $|D_h(\cdot)| > \lambda$ to all local maximizers. Consequently, the estimated change point locations are a set of ordered positions

$$\begin{aligned} \mathcal{J}_{h,\lambda} = \{ \hat{\tau}_\kappa : &\hat{\tau}_\kappa \text{ is a local maximizer of } |D_h(\cdot)|, \\ &\text{and } |D_h(\hat{\tau}_\kappa)| > \lambda \}. \end{aligned}$$

Let $\hat{J} = |\mathcal{J}_{h,\lambda}|$. We let $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{J}})^\top$ denote the SaRa estimator.

In fact, the SaRa procedure can be described as a modified version of binary segmentation. At the first step, $|D_h(\cdot)|$ is calculated and its global maximizer divides the sequence into two segments. Then in each segment, we can repeat the same procedure until the maximum of $|D_h(\cdot)|$ is below some threshold for each segment. Therefore, the difference between the SaRa and BS is the use of a local or global likelihood ratio test statistic as the basis of optimization. The SaRa has several advantages. First, the test statistic does not

have to be recalculated every step. Second, one does not have to conduct the SaRa stagewisely, and a unified threshold can be applied simultaneously. Moreover, with the framework of the SaRa, it is possible to solve change-point problem as a multiple testing one (8).

The computational complexity of the SaRa is only at $O(n)$ because the sequence $D_h(\cdot)$ can be calculated by a recursive formula (Niu and Zhang, 2012).

3.5 SMUCE

Frick, Munk and Sieling (2014) introduced a new change-point detection tool called simultaneous multiscale change-point estimator (SMUCE). It estimates the mean vector $\boldsymbol{\theta}$ by solving an optimization problem

$$(23) \quad \text{minimize} |J(\boldsymbol{\theta})| \quad \text{subject to } T_n(\mathbf{Y}, \boldsymbol{\theta}) \leq q,$$

where $|J(\boldsymbol{\theta})|$ is the number of change points along piecewise constant vector $\boldsymbol{\theta}$, $T_n(\mathbf{Y}, \boldsymbol{\theta})$ is a multiscale statistic defined below and q is a threshold. For a fixed $\boldsymbol{\theta}$,

$$(24) \quad \begin{aligned} T_n(\mathbf{Y}, \boldsymbol{\theta}) = \max_{1 \leq i \leq j \leq n; \theta_i = \theta_{i+1} = \cdots = \theta_j} & \left[\sqrt{2T_i^j(\mathbf{Y}, \theta_i)} \right. \\ & \left. - \sqrt{2 \log \frac{en}{j-i+1}} \right], \end{aligned}$$

where $T_i^j(\mathbf{Y}, \theta_i)$ is a local log-likelihood ratio test statistic for testing $H_0 : \theta = \theta_i$ versus $H_0 : \theta \neq \theta_i$ on the interval $[i, j]$. In the context of normal mean MCM,

$$T_i^j(\mathbf{Y}, \theta_i) = \frac{j - i + 1}{2} (\bar{Y}_i^j - \theta_i)^2,$$

where $\bar{Y}_i^j = \frac{1}{j-i+1} \sum_{\ell=i}^j Y_\ell$. Note that T_i^j is defined only on the interval $[i, j]$ where $\boldsymbol{\theta}$ is constant and it reflects the local discrepancy of the model $\boldsymbol{\theta}$ and the data. That is, $T_n(\mathbf{Y}, \boldsymbol{\theta})$ is an aggregation of local statistics T_i^j. Nevertheless, (23) is a global optimization problem. Frick, Munk and Sieling (2014) employs dynamic programming technique to solve SMUCE. The general complexity is $O(n^2)$, and may be reduced under certain conditions. In a recent work, Pein, Sieling and Munk (2015) further extended SMUCE to the heterogeneous case.

4. CONSISTENCY, CONVERGENCE RATE AND SIMULTANEOUS INFERENCE

In this section, we discuss theoretical results on change-point analysis. Recall that for normal mean MCM (2), we define $L = \min_{0 \leq \kappa \leq J}\{\tau_{\kappa+1} - \tau_\kappa\}$, $\delta =$ $\min_{1 \leq i < n}\{|\theta_i - \theta_{i+1}|\}/\sigma$, and $S = \delta^2 L$ which represents the overall signal strength. For asymptotic theory, we typically assume that all model parameters and these important quantities depends on n and study the asymptotic behavior of an estimator $\hat{\boldsymbol{\tau}}$ as $n \to \infty$.

4.1 Simple Cases

Models (11) and (15) are the two simplest and best understood change-point models. We start with the single change-point model (11), studied in, for example, Hinkley (1970), Csörgő and Horváth (1997). If H_1 is true, and the change-point location τ and mean jump at the change point $\delta = |\theta_{\tau+1} - \theta_\tau|/\sigma$ satisfy either of the following two conditions:

$$(25) \quad \begin{aligned} 0 < \frac{\tau}{n} & \quad \to \quad t < 1, \quad \delta \to 0, \\ & \text{with } \lim_{n \to \infty} \frac{n\delta^2}{\log \log n} = \infty; \end{aligned}$$

$$(26) \quad \begin{aligned} \frac{\tau}{n} & \to 0, \quad \delta \to 0, \\ & \text{with } \lim_{n \to \infty} \frac{\tau\delta^2}{\log \log n} = \infty, \end{aligned}$$

then the maximum likelihood estimator $\hat{\tau} = \hat{j}$ in (14) satisfies

$$(27) \quad \delta^2 |\hat{\tau} - \tau| = O_P(1).$$

For a less challenging setting $\delta \to c > 0$, it implies

$$(28) \quad |\hat{\tau} - \tau| = O_P(1).$$

It was regarded as a result of inconsistency in Hinkley (1970) because there is no way to recover exact change-point location with an overwhelming probability. But at the same time, it was interpreted as a consistency result in the literature since it indicates

$$(29) \quad \left| \frac{\hat{\tau}}{n} - \frac{\tau}{n} \right| = O_P\left(\frac{1}{n}\right).$$

That is, if we embed the sequence into the unit interval, the change-point location $\frac{\tau}{n}$ can be estimated consistently with the convergence rate $\frac{1}{n}$. This is an optimal result and the conditions (25) and (26) cannot be relaxed further (Csörgő and Horváth, 1997).

Arias-Castro, Donoho and Huo (2005) studied the epidemic change-point model (15) and concluded that no method can detect the mean shift reliably when $\delta^2(r - l) < 2 \log n$. Moreover, they proposed a near-optimal procedure which can efficiently and reliably detect the mean shift segment if $\delta^2(r - l)$ is slightly above $2 \log n$. Their work suggests that $2 \log n$ be the optimal detection threshold for (15) and offers a benchmark for the necessary conditions to solve the general change-point problem.

4.2 General Cases

For general cases, the consistency result includes two parts, on the number J^* and location τ^* of change points. That is,

$$\lim_{n \to \infty} P(\hat{J} = J^*) = 1, \tag{30}$$

$$\delta^2 \|\hat{\tau} - \tau^*\|_\infty = O_P(1). \tag{31}$$

Such results have been obtained by Yao (1988), Yao and Au (1989) for the exhaustive search method under an asymptotic setting that J^* and δ are fixed, and $\tau^*/n \to \mathbf{t}$ as $n \to \infty$, where $\mathbf{t} = (t_1, \dots, t_{J^*})^\top$ with $0 < t_1 < \cdots < t_J < 1$ is a constant vector. This asymptotic framework corresponds to category S1 listed in Table 1 of Section 2.4.

Using regression approach and the LASSO techniques, Rinaldo (2009), Qian and Jia (2012) obtained consistency result $\mathbf{P}(\hat{\tau} = \tau^*) \to 1$. However, a strong condition $\delta^2 \gg \log n$ is typically required in most of the related theory with an exception of Harchaoui and Lévy-Leduc (2010). We listed this framework as category S2 in Table 1. The result is less useful because it relies only on big mean changes and not on the distance between the change points.

We should note that a naive approach can also achieve the consistency.

PROPOSITION 1. *Let $z_j = Y_{j+1} - Y_j$ for $j = 1, \dots, n-1$. Consider a naive estimator $\hat{\tau}_{\text{naive}}(c)$ that is the vector of ordered elements in set $\{j : |z_j| > c, 1 \le j \le n-1\}$. With $\delta = \min_{j \in \tau^*} |\theta_{j+1} - \theta_j|/\sigma > 4\sqrt{\log n}$, we have $\mathbf{P}(\hat{\tau}_{\text{naive}}(c) = \tau^*) \to 1$ for $c = 2\sigma\sqrt{\log n}$.*

PROOF. Consider two events $\mathcal{A} = \{|z_j| > c$ for all $j \in \tau^*\}$ and $\mathcal{B} = \{|z_j| < c$ for all $j \notin \tau^*\}$. It is sufficient to show $\mathbf{P}(\mathcal{A} \cap \mathcal{B}) \to 1$ as $n \to \infty$. By definition, we have $z_j \sim \mathcal{N}(\theta_{j+1} - \theta_j, 2\sigma^2)$ for all $1 \le j \le n-1$. Define $\tilde{z}_j = (z_j - (\theta_{j+1} - \theta_j))/(\sqrt{2}\sigma) \sim \mathcal{N}(0,1)$. Then

$$\mathbf{P}\Big(\max_{1 \le j \le n-1} \{\tilde{z}_j^2\} < 2\log n\Big) \to 1, \quad n \to \infty,$$

which implies $\mathbf{P}(\mathcal{A} \cap \mathcal{B}) \to 1$, provided $c = 2\sigma\sqrt{\log n}$ and $\delta > 4\sqrt{\log n}$. It is easy to see that the result holds for $\delta > 2(\sqrt{\log n} + \sqrt{\log J^*})$ if the number of change points $J^* \ll n$. □

Following the work reviewed in Section 4.1, there is a series of results that relax the conditions such as $S = \delta^2 L > C \log n$ or $S \gg \log n$. In particular, Harchaoui and Lévy-Leduc (2010) studied ℓ_1 penalization approach and showed consistency result with a rate slightly slower than $O_P(\log n/\delta^2)$ for change-point locations, under the condition $\delta^2 L \gg \log n$. Hao, Niu and Zhang (2013) obtained optimal $O_P(1/\delta^2)$ convergence with condition $\delta^2 L > 32 \log n$ for the SaRa. Fryzlewicz (2014) proposed WBS which requires $\delta^2 L \gg C \log n$ for sufficiently large C to achieve convergence rate $O_P(\log n/\delta^2)$. SMUCE also requires $\delta^2 L \gg \log n$ (Frick, Munk and Sieling, 2014). In spite of these results, the optimal condition to achieve consistency remains to be an open issue.

We presented above sufficient conditions for different methods to obtain consistency result. However, the necessary conditions are rarely discussed in the literature. Loosely speaking, $S \ge 2 \log n$ appears necessary as we discussed in Section 4.1. Nonetheless, finding necessary conditions to assure consistency is an important line of inquiry.

4.3 Simultaneous Inference

Hinkley (1970), Worsley (1986), Siegmund (1988) studied confidence interval construction for single change point models. For MCM, it is natural to ask how to:

1. construct simultaneous confidence intervals for change-point locations and

2. assign significance simultaneously for all detected change points.

These two questions are distinct but related. The first one concerns with assessing the accuracy and uncertainty of a point estimator for change-point location. When the signal strength is relatively weak, it may not be practical to recover and construct confidence intervals for all change points. However, it is desirable to assign significance level for each detected change point, which will help to identify as many change points as possible while controlling for false positives.

As far as we know, not much work has been done on these important topics until very recently. Frick, Munk and Sieling (2014) proposed SMUCE to construct confidence intervals for multiple change points as well as the stepwise mean function. In addition, the familywise error rate (FWER) for estimated change points can be controlled through a tuning parameter. Hao, Niu and Zhang (2013) discussed the second question and established a framework to control FDR via the SaRa. In particular, they considered the multiple testing formulation introduced in Section 2.2.3. A key point is that, for each change point detected by the SaRa that is a false positive, the distribution of p value for the local test (8) can be obtained. Hence, the significance

levels of all detected change points can be assigned by a simple transformation of the local test statistic. Moreover, the detected change-point locations are well separated so the FDR can be assessed easily by standard Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995).

For any error control procedure, one important and subtle point is to define "true positive" for a detected change-point location. In a standard multiple testing framework without a sequential structure among the tests, it is straightforward to define a true positive. However, this is overly restrictive for change-point problems. Recall that in (6)–(8) $H_1(j)$ is true when j is a change point. In the classical results such as the one presented in (28), the detected change point location is not expected to be exactly the same as the true one with overwhelming probability. That is, it is very likely that $H_0(j)$ is rejected when the position j is close to a true change point. Obviously, it is reasonable to treat it as true positive if $H_0(j)$ is the only rejected null hypothesis in a small neighborhood of a true change-point location τ. This strategy was used in Hao, Niu and Zhang (2013). In particular, they treated an estimate $\hat{\tau}$ as revealing a true change point τ if they are close enough, say, $|\hat{\tau} - \tau| < h$, where h is the same window size as in (8).

Except SaRa and SMUCE, the methods for MCM have not been well understood theoretically, and it remains to establish an inferential framework for those methods.

5. AGGREGATING LOCAL INFORMATION

For some of aforementioned algorithms including exhaustive search, dynamic programming, ℓ_1-penalization, BS and CBS, all the data are involved directly and simultaneously in an optimization procedure or the initial step of a sequential test procedure. We may call them *global* methods, which have been playing a dominant role in change-point analysis. On the other hand, a variety of *local* methods have been developed recently. Those local methods share a common strategy, that is, to first gather and then aggregate local information. Among the methods reviewed in this paper, we may consider the SaRa, WBS and BWD as examples of local methods. In addition, SMUCE also shares a similar spirit.

Change-point inference is often stated in a hypothesis testing framework of the form (5). Exhaustive search aims to solve (5) directly by searching for the strongest evidence against the null hypothesis over all possible location combinations. Forward stepwise methods BS and CBS solve testing problems of types (11) and (15) recursively. In contrast, local methods start from various localized versions of (11). For example, the SaRa focuses on a sequence of local tests (8), and makes decision based on local maxima of the test statistic sequence and a global thresholding rule; WBS considers test (17) defined on a random interval $[s, e]$, and then combines the test statistics over a set of random intervals; BWD considers a sequence of local test problems (19) at each step, and retain a single null hypothesis by comparing local test statistics. We see that the SaRa, WBS and BWD focus first tests on small local segments then make decisions based on an aggregation of the local test statistics. Note that for BS and CBS, the alternative hypothesis is never truly correct except the very last step. Therefore, the signal brought by each change point may be weakened by other change points, leading to reduced power. To elaborate, consider a sequence of Gaussian random variables of length $n = 1000$ with mean $\theta_i = 1$ when $500 < i \leq 510$ and 0 otherwise. That is, the mean jumps up 1 at position 500 and drops down 1 at position 510. Without the second change point at 510 (i.e., $\theta_i = 1$ for all $i > 500$), the statistic $-2 \log \Lambda_{500}$ as defined in (13) would be X^2 for $X \sim \mathcal{N}(\sqrt{250}, 1)$. With the change point at 510, $-2 \log \Lambda_{500} = X^2$ for $X \sim \mathcal{N}(1/\sqrt{10}, 1)$, which takes a much smaller value. The negative log likelihood ratios at other positions behave similarly as $-2 \log \Lambda_{500}$. Therefore, when BS is applied, the test statistic (12) may not be significant as the signals brought by the two change points are neutralized. Thus, aggregating local tests should avoid signal cancellation caused by multiple change points. As illustrated in Section 2.3, a good method may mimic the oracle to perform single change-point test on small segments, each of which has at most one change point. To this end, WBS draws a set of random segments in each step and hopes that some of these segments are close to the ones in the oracle division; SaRa checks a small neighborhood of each position; BWD builds up segments adaptively from the data beginning with single point segments. SMUCE estimates change points by solving a global optimization problem. Nevertheless, the employed multiscale statistic (24) can be considered as an aggregation of local statistics.

In spite of these recent developments, researchers are still looking for better methods to aggregate local information, especially, when little information on the distances among change points is available. One difficulty is how to determine a local neighborhood when

calculating a local statistic. Intuitively, it may involve two or more change points if too large neighborhood is used. On the other hand, it may reduce the power if a small neighborhood is used. Moreover, it may not be straightforward to compare and aggregate local statistics if they are calculated based on neighborhoods having different sizes.

6. CONCLUDING REMARKS

In spite of recent rapid developments on this topic, there are many interesting and open research directions to explore such as optimal detection and simultaneous inference for multiple change points. In contrast to classical approaches which solve global optimization to estimate change points, there are emerging new methods which decompose MCM into a set of local problems, and then gather and aggregate local information to solve MCM. There are several advantages to this strategy. First, the local information is usually summarized by a single statistic for each local problem. A typical approach to aggregating these statistics is to find (local or global) extreme values. Therefore, the computation is less expensive as no complex optimization is involved. Second, these local statistics together bring more information than a point estimator so inference is possible. For example, in the SaRa, local maxima are extracted to represent the likelihood for the presence of change points, and the local statistics around local maxima are ignored but potentially useful for inference. So far, besides recent works (Hao, Niu and Zhang, 2013, Frick, Munk and Sieling, 2014), a solid framework for inference on multiple change points based on local methods is lacking. It is important to find the best way to gather and aggregate the local information.

Without prior knowledge, a change point may be located anywhere along the sequence. It is well known that there is no optimal test for (1) because of the indeterminacy of the change-point location (Sen and Srivastava, 1975). We may have one procedure that is more powerful when the change point is in the middle and the other one more powerful when it is close to the boundary. As illustrated in Section 2.3, we may consider the circular model to make the change-point problem symmetric. The symmetry comes from the finite group action $\mathbb{Z}_n$ to the change-point locations, where $\mathbb{Z}_n$ is a cyclic group of order n. Under this new formulation, we may study equivariant change-point detection procedures. Among the aforementioned algorithms, CBS is designed to solve the circular model,

while exhaustive search, SMUCE, SaRa and BWD are able to solve the circular model after slight modifications. Studying theoretical properties of equivariance procedures for the symmetric model remains an open topic.

It seems that two striking features of MCM, that is, locality and symmetry, have not been emphasized in the literature. We hope that this paper can stimulate more research on these aspects.

ACKNOWLEDGMENTS

REFERENCES

ARIAS-CASTRO, E., DONOHO, D. L. and HUO, X. (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* **51** 2402–2425. MR2246369

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BRAUN, J. V., BRAUN, R. K. and MÜLLER, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. MR1782480

BRODSKY, E. and DARKHOVSKY, B. S. (1993). *Nonparametric Methods in Change Point Problems*. Kluwer, Dordrecht. MR1228205

BRODSKY, B. E. and DARKHOVSKY, B. S. (2000). *Nonparametric Statistical Diagnosis*: *Problems and Methods*. *Mathematics and Its Applications* **509**. Kluwer Academic, Dordrecht. MR1862475

CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Stat.* **41** 389–405.

CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D., eds. (1994). *Change-Point Problems*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23**. IMS, Hayward, CA. MR1477909

CHEN, J. and GUPTA, A. K. (2000). *Parametric Statistical Change Point Analysis*. Birkhäuser, Inc., Boston, MA. MR1761850

COBB, G. W. (1978). The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika* **65** 243–251. MR0513930

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2743035

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979

HAO, N., NIU, Y. S. and ZHANG, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica* **23** 1553–1572. MR3222810

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

HAWKINS, D. M. (1977). Testing a sequence of observations for a shift in location. *J. Amer. Statist. Assoc.* **72** 180–186. MR0451496

HINKLEY, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17. MR0273727

HUANG, T., WU, B., LIZARDI, P. and ZHAO, H. (2005). Detection of DNA copy number alterations using penalized least squares regression. *Bioinform.* **21** 3811–3817.

JACKSON, B., SCARGLE, J. D., BARNES, D., ARABHI, S., ALT, A., GIOUMOUSIS, P., GWIN, E., SANGTRAKULCHAROEN, P., TAN, L. and TSAI, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12** 105–108.

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

LEE, T.-S. (2010). Change-point problems: Bibliography and review. *J. Stat. Theory Pract.* **4** 643–662. MR2758751

LEVIN, B. and KLINE, J. (1985). The cusum test of homogeneity with an application in spontaneous abortion epidemiology. *Stat. Med.* **4** 469–488.

NIU, Y. S. and ZHANG, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6** 1306–1326. MR3012531

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostat.* **5** 557–572.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850

PAGE, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* **42** 523–527. MR0072412

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

PEIN, F., SIELING, H. and MUNK, A. (2015). Heterogeneous change point inference. Preprint. Available at arXiv:1505.04898.

QIAN, J. and JIA, J. (2012). On pattern recovery of the fused Lasso. Preprint. Available at arXiv:1211.5194.

RINALDO, A. (2009). Properties and refinements of the fused lasso. *Ann. Statist.* **37** 2922–2952. MR2541451

SCOTT, A. J. and KNOTT, M. (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* **30** 507–512.

SEN, A. and SRIVASTAVA, M. S. (1975). On tests for detecting change in mean. *Ann. Statist.* **3** 98–108. MR0362649

SHIN, S. J., WU, Y. and HAO, N. (2014). A backward procedure for change-point detection with applications to copy number variation detection. Unpublished manuscript.

SIEGMUND, D. (1988). Confidence sets in change-point problems. *Int. Stat. Rev.* **56** 31–48. MR0963139

TIBSHIRANI, R. and WANG, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostat.* **9** 18–29.

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

VENKATRAMAN, E. S. and OLSHEN, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinform.* **23** 657–663.

VOSTRIKOVA, L. Y. (1981). Detecting "disorder" in multidimensional random processes. *Sov. Math., Dokl.* **24** 55–59.

WORSLEY, K. J. (1986). Confidence regions and test for a change-point in a sequence of exponential family random variables. *Biometrika* **73** 91–104. MR0836437

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373

YAO, Q. W. (1993). Tests for change-points with epidemic alternatives. *Biometrika* **80** 179–191. MR1225223

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhya, Ser. A* **51** 370–381. MR1175613

YIN, Y. Q. (1988). Detection of the number, locations and magnitudes of jumps. *Commun. Stat., Stoch. Models* **4** 445–455. MR0971600

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

ZHANG, Z., LANGE, K., OPHOFF, R. and SABATTI, C. (2010). Reconstructing DNA copy number by penalized estimation and imputation. *Ann. Appl. Stat.* **4** 1749–1773. MR2829935