The Annals of Statistics
2020, Vol. 48, No. 3, 1615–1647
https://doi.org/10.1214/19-AOS1861

SEGMENTATION AND ESTIMATION OF CHANGE-POINT MODELS: FALSE POSITIVE CONTROL AND CONFIDENCE REGIONS

BY XIAO FANG[1], JIAN LI[2] AND DAVID SIEGMUND[3]

[1]*Department of Statistics, The Chinese University of Hong Kong, xfang@sta.cuhk.edu.hk*
[2]*Adobe Systems, jianli0124@gmail.com*
[3]*Department of Statistics, Stanford University, siegmund@stanford.edu*

To segment a sequence of independent random variables at an unknown number of change-points, we introduce new procedures that are based on thresholding the likelihood ratio statistic, and give approximations for the probability of a false positive error when there are no change-points. We also study confidence regions based on the likelihood ratio statistic for the change-points and joint confidence regions for the change-points and the parameter values. Applications to segment array CGH data are discussed.

1. Introduction. Diverse scientific applications have led to recent interest in segmentation of models involving multiple change-points. A model having some direct applicability and additional theoretical interest for the insights it provides is as follows. Let $X_1, X_2, \ldots, X_m$ be independent and normally distributed with variances equal to 1. Assume that there exist $M \geq 0$ and integers $0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = m$ such that the mean μ_i of X_i, $1 \leq i \leq m$ is a step function with constant values on each of the intervals $(\tau_{k-1}, \tau_k]$, $1 \leq k \leq M+1$, but different values on adjacent intervals. Segmentation amounts to determining the value of M, the τ_k and perhaps also the μ_i. Because of the computational difficulty of sorting through all possible partitions of $[1, m]$ to find the change-points when m is large, there have often been different algorithms for suggesting a set of candidate change-points τ_k and for determining which of those possible sets is "correct." For example, one might use a dynamic programming algorithm to propose a relatively small set of possible M and τ_k, $1 \leq k \leq M$, then use a statistical procedure to determine a final choice from those suggested in the first stage of analysis. For recent reviews imbedded in otherwise original research articles, see Frick, Munk and Sieling (2014) and Fryzlewicz (2014). Recent consistency results under essentially minimal conditions on the spacing and amplitude of the change-points are given in Chan and Chen (2017).

Substantial motivation for recent research has been copy number variation (CNV) in genetics (e.g., Olshen et al. (2004), Pollack et al. (1999), Picard et al. (2005), Lai et al. (2005), Snijders et al. (2003), Zhao et al. (2004), Zhang and Siegmund (2007), Niu and Zhang (2012), Frick, Munk and Sieling (2014), Zhang et al. (2016)). CNV can occur as somatic mutations, especially in cancer cells, where they can involve a substantial portion of a chromosome and show no particular pattern, or as germline mutation, which typically involves a short interval exhibiting an increase or decrease in the mean followed by a decrease or increase that returns the mean to a baseline value. Like other genetic polymorphisms, inherited CNV can be used to track relatedness of different individuals in populations or may be of interest because of a possible relation to particular inherited diseases. Data in the literature can help us determine interesting sample sizes and values for parameters in our numerical examples. The sample size m is typically moderately large to large, while M can be small or large in an absolute

Received March 2018; revised February 2019.
MSC2010 subject classifications. 62G05, 62G15.
Key words and phrases. Array CGH analysis, change-points, confidence regions, exponential families, likelihood ratio statistics.

sense, while still small compared to m; and consecutive change-points can be quite close together.

Another genomic application involves sequences of Bernoulli variables, which equal 0 or 1 according as the DNA letter at that location is A or T, or is C or G. Since a CG "rich" region is an indication of the presence of a gene or genes, it may be useful to segment a genome or part of a genome into regions of relatively low or high CG content. See, for example, Churchill (1989) who used a hidden Markov model, or Elhaik, Graur and Josić (2010). A variety of other examples motivated by particular scientific experiments is given by Du, Kao and Kou (2016). In particular, they describe examples where several consecutive changes are expected to have the same sign and where the pattern of change-points may arise from a hidden Markov model.

Scientific focus may emphasize detection and estimation of the change-points, estimation of the step function of mean values, or a combination of the two. Our primary focus is on the change-points themselves, which in a genomic context indicate the existence and location of a signal of interest.

To this end, we study iterative thresholding methods that allow one (with varying degrees of success, discussed below) to control the global false positive error rate; and subject to successful control, to understand the relative strengths and weaknesses of different methods. We also provide approximations to the local power (defined below) and large sample joint confidence regions for the change-points or for the change-points and mean values.

It bears emphasizing that we have *not* considered a large class of other methods, in particular, dynamic programming to compute a penalized likelihood function, or two stage methods, where a list of candidate change-points obtained in the first stage is followed by a model selection method. (Three of the four methods used below for comparative purposes have been originally proposed as two stage methods, but we have adapted them to be single stage thresholding methods.) To the best of our knowledge, these methods all involve selection of arbitrary parameters that may have no simple statistical interpretation. In contrast, we have produced a set of tools to help us understand a restricted set of procedures, each depending on a single thresholding parameter, in terms of classical statistical criteria of false positive and false negative error rates, perhaps supplemented by joint confidence regions.

To motivate the methods introduced below, let $S_0 = 0$, $S_j = \sum_1^j X_i$ for $j \geq 1$ and consider the generalized likelihood ratio statistic for testing the hypothesis $M = 0$ against $M = 1$: $\max_{0<j<m} |S_j - jS_m/m|/[j(1-j/m)]^{1/2}$. This statistic is the basis of the binary segmentation suggestion of Vostrikova (1981), which is a "top down" procedure, in the sense that one tests all the data to determine if there is *at least* one change-point and iterates the procedure in the intervals immediately to the "left" and "right" of the most recently detected change-point. We discuss below the weaknesses of this method compared to a number of other thresholding procedures.

Here, we consider "bottom up" procedures motivated by the observation that in the presence of multiple change-points or to mitigate the effects of inadequately controlled drift in the "baseline" mean value (see below), it may be useful to compare a candidate change-point at j to an appropriate "local" background (i, k), where $i < j < k$. Similar approaches are the Wild Binary Segmentation (WBS) of Fryzlewicz (2014), which uses a random set of possible backgrounds and an apparently empirically determined threshold, and the method of Niu and Zhang (2012), who use a limited number of symmetric backgrounds, to suggest several sets of candidate change-points followed by model selection to make the final choice. (Our suggested procedures could also form part of a two-stage procedure, but here we consider in detail only a single stage, which controls the rate of false positives.)

To that end, consider

$$\max_{0\leq i<j<k\leq m} |Z_{i,j,k}|, \tag{1.1}$$

where for $i < j < k$:

$$Z_{i,j,k} = \big[S_j - S_i - (j-i)(S_k - S_i)/(k-i)\big] \Big/ \big[(j-i)\big(1-(j-i)/(k-i)\big)\big]^{1/2}. \tag{1.2}$$

Our first theoretical result in Theorem 2.1 below is an approximation for the tail probability of (1.1) when there is no change. This approximation gives strong control of the probability of a false positive result in the sense that if there are M true change-points, say at $0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = m$, the maximum of (1.2) over $n = 0, \ldots, M$ and $\tau_n \le i_0 < j_0 < k_0 \le \tau_{n+1}$ is stochastically smaller than (1.1) when there are no change-points. Hence, except for an event of the probability evaluated asymptotically in Theorem 2.1, any background interval (i_1, k_1) where $\max_j |Z_{i_1,j,k_1}|$ exceeds the threshold must contain a τ_ℓ for some $1 \le \ell \le M$.

Our second principal result is an approximate likelihood ratio confidence region jointly for the change-points $\{\tau_k, k = 1, \ldots, M\}$ or for the change-points and the mean values. A related result allows one to approximate the local power (cf. Section 3.2), which we find useful in helping us understand which change-points are relatively easily detected and which may be missed.

In more detail, our segmentation procedure based on (1.2), which we call the Local Likelihood Ratio (LLR) is as follows. (In the following, we use the same acronym to refer to both the statistic and associated segmentation procedures.) Because of local correlations between different $Z_{i,j,k}$, thresholding (the absolute value of) (1.2) produces a frequently large list of candidate change-points j, each one against multiple backgrounds (i, k). Since our goal is to detect individual changes against the appropriate background, we find it convenient in searching the list of candidates to require that the background for one candidate change-point j not overlap another candidate change-point j' in the sense that if $j < j'$, the corresponding backgrounds should satisfy $k \le j'$ and $i' \ge j$. This can be accomplished by sequentially reevaluating candidate change-points until they satisfy the constraint. Hence, when a new change-point is identified, the existing putative change-points to its left and to its right may need to be removed or altered. An approach that requires very little and usually no reevaluation of candidate change-points is to select the shortest of the possible backgrounds (i, k) from among those for which $Z_{i,j,k}$ exceeds the required threshold, which is similar to the method recommended in Baranowski, Chen and Fryzlewicz (2019). If there is a tie for the shortest value of $k - i$, we choose the one with the largest value of $|Z_{i,j,k}|$. Another possible algorithm is selection based on the largest value of the statistic, which rarely leads to significant differences from selection based on the shortest background, although now iteration to enforce the no overlap condition is common.

For various scientific reasons and in particular for determination of the confidence regions discussed below, an important consideration is the size, as well as the location of the change. Although it seems natural to conjecture that the largest $|Z|$-value, subject to no overlap, provides the most accurate estimate since it is based on a longer background, from simulations we can see that this is by no means always the case. Since we have "paid up front" for protection against false positive errors, we can also choose to look at a number of candidate change-point-background combinations to find one that is subjectively appealing. Simple possibilities that seem to present themselves frequently are to choose from among the (i, j, k) combinations having, say, the five shortest backgrounds the value of j that appears most frequently, or has the largest Z-score, or also heads the list generated by the algorithm that is focused on the largest Z-score.

We also study a pseudo-sequential procedure (SLLR) where we initially set $i = 0$, find the smallest $k > i + 1$ such that $\max_{i<j<k} |Z_{i,j,k}|$ is above an appropriate threshold, set j_1 equal the largest such j or the value of j that maximizes $|Z_{i,j,k}|$, then set $i = j_1$ and iterate

the process. In simulations not reported here, we have found that there is rarely a significant difference between these alternatives for defining j_1 and have not found a preference for one method over the other. This procedure has lower computational complexity than LLR, although it is still a bottom up procedure in the sense that each detected change-point is compared to a local background that ideally contains no other change-points that might introduce potential biases into the estimated location or size of the change. Because SLLR has a lower threshold than LLR, it also usually has larger power to detect change-points. However, as explained below Theorem 2.2, we do not have as strong a theoretical guarantee that the false positive error probabilities are controlled. See the unpublished Stanford Ph.D. thesis of E. S. Venkatraman for an early discussion of a similar idea.

The paper is organized as follows. In Section 2, we give approximations to control the false positive probabilities of our proposed and other segmentation methods. Approximate joint confidence regions are discussed in Section 3. Using simulations and analysis of some real data, we compare our methods in Section 4 with other thresholding methods that control false positive probabilities with varying degrees of success, and we give some numerical examples for confidence regions. Section 5 contains extensions to exponential families, and Section 6 some additional discussion. In an online supplement (Fang, Li and Siegmund (2019)), we prove the theorems stated in Sections 2 and 3.

REMARKS. (i) For some applications, for example, for inherited CNV, the signal to be detected extends over a relatively short range in the form of a departure from a baseline value where one change is followed by a second, nearby change in the opposite direction. For problems of this form, it seems reasonable to use statistics adapted to the expected shape of the signals, (e.g., Olshen et al. (2004), Frick, Munk and Sieling (2014), Zhang et al. (2010)). We give appropriate approximations for false positive control in Section 2 and consider such procedures in more detail in Section 4, where we show that they perform very well even when there is no particular pattern to the change-points. Inherited CNV also provide motivation for studying multivariate observations (i.e., multiple DNA sequences), since the change-points may be difficult to detect in individual sequences and their occurrence in several sequences indicate possible relationships among those sharing the same change-points (Zhang et al. (2010)).

(ii) For theoretical calculations, we have assumed the variance of the observations is known. The sequences are usually long enough that, under our assumption of independence, the variance can be accurately estimated by one-half the average of the squared differences of consecutive observations. This estimator avoids the substantial upward bias of the empirical variance of the data, which arises when there are multiple change-points and the changes themselves show no particular pattern. We have used this estimator in our simulations and applied examples. Alternatives are to use a function of the order statistics of adjacent observations, for example, the median of $|X_i - X_{i-1}|$ or the interquartile range of $X_{i+1} - X_i$, multiplied by suitable constants. Still another estimator that may have some value is the average of squared second-order differences of consecutive observations, which has the virtue of nullifying the effect of linear drift and perhaps also the relatively slow oscillations in the baseline distribution that plague some genomic applications (Olshen et al. (2004)). The estimator based on pairwise differences is inappropriate when the data are autocorrelated, a problem we expect to study in the future. An example where the empirical variance is itself satisfactory occurs when the data are of the form envisioned in Remark (i), where change-points occur in a relatively sparse set of departures from a baseline value followed by a return to the baseline a few observations later.

(iii) In recent research, some authors recommend a multiscale modification of the likelihood ratio statistic. One possibility is to modify (1.2) by subtracting $\{2\kappa \log[3m/\min(j -$

$i, k-j)]\}^{1/2}$, $\kappa > 0$, in order to obtain greater power to detect relatively small changes that persist over longer intervals at the cost of less power to detect large changes that come relatively close together; see, for example, Dümbgen and Spokoiny (2001) and Frick, Munk and Sieling (2014). Our methods can be adapted to study these modifications, and in Section 4 we investigate a procedure based on the statistic recommended in Frick, Munk and Sieling (2014). However, these methods are not central to our studies for the following reasons. (a) For problems of CNV detection, difficult detections in the synthetic data suggested in the applied literature and in the real data in Section 4 often involve short intervals and relatively large changes. (b) What appear to be small, relatively isolated changes may arise from technical artifacts and are not scientifically interesting (cf. Olshen et al. (2004), Zhang et al. (2010), and Table 5). (c) Multiscale modifications are not uniquely defined; and in different problems different statistics may have slight advantages and disadvantages. (d) Multiscale methods do not appear to adapt as naturally for the determination of confidence regions as the likelihood ratio statistic.

2. Approximate *p*-values. In what follows, we write $A \asymp B$ to mean that $0 < c_1 \le A/B \le c_2 < \infty$ for two absolute constants c_1 and c_2, and $A(b) \sim B(b)$ means $A(b)/B(b) \to 1$ as $b \to \infty$; also φ and Φ are the standard normal probability density function and distribution function, respectively.

We have the following p-value approximation for $\max_{0\le i<j<k\le m} |Z_{i,j,k}|$.

THEOREM 2.1. *Let $(X_1, \dots, X_m)$ be an independent sequence of normally distributed random variables with mean μ and variance* 1. *Then for $Z_{i,j,k}$ as defined by* (1.2), *we have for $b \to \infty$ and $m \asymp b^2$,*

$$\begin{aligned}
&\mathbb{P}\Big\{\max_{0\le i<j<k\le m} |Z_{i,j,k}| \ge b\Big\} \\
&\quad\sim \frac{b^5\varphi(b)}{4} \sum_{\substack{u,v\in\{1,\dots,m\}:\\ u+v\le m}} \frac{(m-u-v)}{uv(u+v)} \nu\Big[b\Big(\frac{u}{v(u+v)}\Big)^{1/2}\Big] \\
&\qquad\times \nu\Big[b\Big(\frac{v}{u(u+v)}\Big)^{1/2}\Big]\nu\Big[b\Big(\frac{u+v}{uv}\Big)^{1/2}\Big].
\end{aligned} \tag{2.1}$$

The function ν is defined, for example, in Siegmund and Yakir (2007) *page* 112 *and given to a simple approximation* (*which we use below*) *by the equation*

$$\nu(x) = \big(\Phi(y) - 1/2\big)/\big[y\big(y\Phi(y) + \varphi(y)\big)\big],$$

where $y = x/2$.

For the proof of Theorem 2.1, we use a new method beginning from an observation of Zhang and Liu (2011), which was used there as the basis for Monte Carlo simulation with a one (time)-dimensional random field and which we have used for an analytic approximation involving maxima of certain three (or higher dimensional) random fields. The starting point is a number of simple observations, which require a large number of detailed calculations for complete justification.

Denote the right-hand side of (2.1) by p, which asymptotically

$$\begin{aligned}
&\sim \frac{\varphi(b)}{4b} \sum_{\substack{u,v\in\{1,\dots,m\}:\\ u+v\le m}} (m-u-v)\Big\{\frac{b^6}{uv(u+v)} \nu\Big[b\Big(\frac{u}{v(u+v)}\Big)^{1/2}\Big] \\
&\qquad\times \nu\Big[b\Big(\frac{v}{u(u+v)}\Big)^{1/2}\Big]\nu\Big[b\Big(\frac{u+v}{uv}\Big)^{1/2}\Big]\Big\}.
\end{aligned}$$

It was shown in Siegmund (1985) that $\nu(x) = \exp(-cx) + o(x^2)$ as $x \to 0$ for $c \approx 0.583$, while $x^2\nu(x)/2 \to 1$ as $x \to \infty$. Therefore, the term inside the curly brackets above is bounded. Hence

$$p \asymp b^5\varphi(b) \to 0, \tag{2.2}$$

where we used the assumption that $m \asymp b^2$.

Fix a sufficiently small constant c_0. We will prove first that

$$\begin{aligned}
&\mathbb{P}\Big\{\max_{\substack{0\le i<j<k\le m:\\ j-i,k-j\ge c_0b^2}} Z_{i,j,k} \ge b\Big\} \\
&\quad\sim \frac{1}{8}b^5\varphi(b) \sum_{\substack{u,v\in\{1,\dots,m\}:\\ u+v\le m;u,v\ge c_0b^2}} \frac{(m-u-v)}{uv(u+v)}\nu\Big[b\Big(\frac{u}{v(u+v)}\Big)^{1/2}\Big] \\
&\qquad\times \nu\Big[b\Big(\frac{v}{u(u+v)}\Big)^{1/2}\Big]\nu\Big[b\Big(\frac{u+v}{uv}\Big)^{1/2}\Big].
\end{aligned} \tag{2.3}$$

We write

$$\begin{aligned}
&\mathbb{P}\Big(\max_{\substack{0\le i<j<k\le m\\ j-i,k-j\ge c_0b^2}} Z_{i,j,k} \ge b\Big) \\
&\quad= \sum_{\substack{0\le i<j<k\le m\\ j-i,k-j\ge c_0b^2}} \mathbb{P}\Big(Z_{i,j,k} \ge b, Z_{i,j,k} = \sum_{\substack{0\le r<s<t\le m\\ s-r,t-s\ge c_0b^2}} Z_{r,s,t}\Big) \\
&\quad= \sum_{\substack{0\le i<j<k\le m\\ j-i,k-j\ge c_0b^2}} \int_0^\infty \mathbb{P}\Big(\max_{\substack{0\le r<s<t\le m\\ s-r,t-s\ge c_0b^2}} Z_{r,s,t} \le b+x \,|\, Z_{i,j,k} = b+x\Big) \\
&\qquad\times \mathbb{P}(Z_{i,j,k} \in b+dx) \\
&\quad= \sum_{\substack{C\log b\le i<j<k\le m-C\log b\\ j-i,k-j\ge c_0b^2}} \int_b^{b+1} \mathbb{P}\Big(\max_{\substack{0\le r<s<t\le m\\ s-r,t-s\ge c_0b^2}} Z_{r,s,t} \le x \,|\, Z_{i,j,k} = x\Big)\, d\mathbb{P} + R,
\end{aligned}$$

where C is a positive constant to be chosen. The rest of the proof involves a detailed analysis of the conditional probability, to show that R and various other terms that have been ignored are indeed negligible. These technical details are facilitated by the assumption that $b^2 \asymp m$, which guarantees that the range of perturbations of i, j and k that must be considered is not too large and in this range the conditional probability is well behaved. Details are given in Appendix A of the online supplement.

REMARKS. (i) Other methods that appear to be adaptable to prove Theorem 2.1, albeit with more, less intuitive, computation are those of Siegmund (1988a) and of Yakir (2013).

(ii) Usually we are interested in small probabilities, and then we can use (2.1) as given. Occasionally, we may be interested in cases where m is so large that the probability is not small. In those cases, we can supplement our large deviation approximation with a "Poisson" approximation in the form $1 - \exp[-\text{RHS}(2.1)]$, which reduces to our approximation when the probability is small; see Siegmund and Yakir (2000) for a proof in a related case.

(iii) Based on other, related calculations (see, e.g., (5) in Zhang et al. (2010)), it seems clear that similar results apply to multivariate data with a few changes. This case is particularly

interesting for detection of inherited CNV, which are short and sometimes difficult to detect in single DNA sequences (e.g., Zhang et al. (2010)). It is also interesting to infer which subsets of the distributions have changed at the various change-points. The required modifications of the approximation given in the theorem are (a) replace $\varphi(b)$ by $bf_d(b^2)$ where f_d is the χ^2 probability density function with d degrees of freedom, and (b) multiply the entire expression by q^5 and the arguments of the functions ν by q, where $q = 1 - (d-1)/b^2$ (to account for the curvature of the sphere when the dimension d is large); see (2.9) below.

(iv) For a multiscale statistic along the lines suggested in Remark (iii) at the end of Section 1, where we subtract, for example, $\{2\kappa \log[3m/\min(j-i, k-j)]\}^{1/2}$ from (1.2), a similar approximation holds, with the right-hand side modified by replacing b by $b(u,v) = b + \{2\kappa \log[3m/\min(u,v)]\}^{1/2}$ and moving the expressions involving $b(u,v)$ inside the summation; see (2.9) below for a similar approximation involving the (Frick, Munk and Sieling (2014)) recommended statistic.

(v) In applications, we may wish to put a lower and/or an upper bound on the length of the background, for example, $m_0 \le j-i, k-j \le m_1$. The appropriate change to (2.1) is to restrict the summation on the right-hand side to $m_0 \le u, v \le m_1$. For applications where very short intervals between change-points can occur, we may want to take $m_0 = 1$. Values of $m_1 \ll m$ can be used to minimize detection of small jumps, which may themselves reflect experimental artifacts that lead to drift in the underlying distributions (cf. Olshen et al. (2004), Zhang et al. (2010)); and they speed up what may otherwise be time consuming computations for large values of m; see Section 4 for other possible speed-ups.

(vi) If there is a large number of change-points to be detected, one might prefer to control the rate of false positive errors via the false discovery rate (FDR). In change-point problems, it is important to distinguish between discoveries and "tests," since in our context many correlated test statistics may refer to relatively few change-points. Efforts to clarify and deal with this distinction that seem applicable in principle to our segmentation problem are found in Schwartzman, Gavrilov and Adler (2011), Hao, Niu and Zhang (2013) and Siegmund, Zhang and Yakir (2011); but since the number of change-points in our motivating examples is typically not large, we do not consider this possibility in detail.

(vii) One can choose to approximate the probability in the theorem by simulation; and to do this once for a particular study does not seem to pose difficulties. Simulation may also be useful to study variations of our problem under different models, and for WBS we do not know any alternative. But the analytic approximations are much faster to evaluate, and hence we find it useful to perform a relatively limited set of simulations to gain confidence that our approximations are reasonably accurate, then use the approximations. In addition to determining suitable thresholds, we may want to compute p-values, defined by the probability that a statistic would exceed its observed value under specified conditions, or to determine the confidence regions suggested in Section 3, which involve iterative computation of probabilities; for these and other problems rapid analytic computation is useful. We will, nevertheless, see in Section 4 that simulations play an indispensable role when we deal with difficult to approximate probabilities.

(viii) The natural setting for these approximations is the likelihood ratio statistic in exponential families, as discussed briefly in Section 5. A more robust, although often quite conservative approximation is to replace the discrete time random walk of the theorem by continuous time Brownian motion. The resulting approximation would look the same, but the functions ν would be replaced by 1, and the sums would be integrals (perhaps still evaluated as sums). While this would in principle allow the theorem to be applied to a wide variety of statistics, in specific cases the approximation may be quite conservative.

For the SLLR, we have a similar approximation to the probability of a false positive detection.

THEOREM 2.2. *Let $(X_1, \dots, X_m)$ be a sequence of independent normally distributed random variables with mean μ and variance* 1. *Let $Z_{i,j,k}$ be defined as in* (1.2). *We have for $b \to \infty$ and $m \asymp b^2$,*

$$\begin{aligned}&\mathbb{P}\Big\{\max_{0<j<k\le m} |Z_{0,j,k}| \ge b\Big\}\\ &\quad\sim \frac{1}{2}b^3\varphi(b)\sum_{1<k\le m}\sum_{0<j<k} j^{-2}\nu\big\{b\big[(k-j)/(jk)\big]^{1/2}\big\}\\ &\qquad\times \nu\big\{b\big[k/(j(k-j))\big]^{1/2}\big\}.\end{aligned} \tag{2.4}$$

Compared to the segmentation procedure based on LLR, the pseudo-sequential procedure has the advantage that it is easier and faster to implement. In the case of no discovery, it searches over two indices j, k, as opposed to three indices i, j, k in LLR. It does not, however, have as strong a theoretical guarantee that the false positive error probabilities are controlled, since each restart creates an independent opportunity for a false positive error. For example, if there is a large change (or even no change) at τ, a detection may occur at $t < \tau$, and process restarted at t may detect the same change a second time. This situation is usually easy to recognize and correct *ad hoc*. It could be prevented by delaying the restart, but this would hurt the power to detect a genuine second change that occurs near the first, since SLLR must have enough data to estimate the current mean before it can detect a change. Finally, since SLLR does not use the maximum available background, the location of the change-point and the magnitude of a change may not be as accurately estimated as with LLR. However, as simulations and examples below suggest, SLLR is quite stable and efficient.

The statistic suggested by Niu and Zhang (2012) is similar to LLR, but uses a background that is symmetric around a putative change-point. Consider the local maxima with respect to j of

$$Z_{j,h} = \big|\big[(S_{j+h} - S_j) - (S_j - S_{j-h})\big]\big|/(2h)^{1/2}, \tag{2.5}$$

where h is a parameter to be chosen. Since there is no obvious choice for h, Niu and Zhang suggest maximizing (2.5) over a finite number of values of h. For their applications to copy number variation, they suggest 3 values, 10, 20 and 30. To complete their method, which they call SaRa, they use a model selection procedure following their use of (2.5), a step that we omit.

Our implementation of SaRa below takes maxima over all j and h. The methods of proof of Theorems 2.1 and 2.2 (and (2.9) below) use the fact that local perturbations of the processes around a high maximum, for example, the difference between a large value of $Z_{i,j,k}$ and values $Z_{i',j',k'}$ as a function of $i' \approx i$, $j' \approx j$ and $k' \approx k$ involves a sum of three approximately independent random walks. Since the local random walks obtained from perturbations of j and h for (2.5) are not independent, we cannot apply the same methods to obtain a theoretical approximation to the false positive error probability. Since the local increments are weakly positively dependent, it seems natural to conjecture that treating them as if they were independent would produce a slightly conservative approximation.

An approximation to the the maximum over j and h of (2.5), calculated on the assumption that the local increments obtained from perturbations of j and h are independent is given by

$$\begin{aligned}&\mathbb{P}\Big\{\max_{0<t<m,0<h<\min(t,m-t)} |Z_{t,h}| \ge b\Big\}\\ &\quad\sim 1.5mb^3\varphi(b)\sum_h \nu\big[b(3/h)^{1/2}\big]\nu\big[b(1/h)^{1/2}\big]/h^2.\end{aligned} \tag{2.6}$$

TABLE 1
Approximation (2.1). Simulated values based on $N = 10{,}000$ repetitions in the first three rows, 1000 in the last two rows and 2000 otherwise

b	m	m_0	m_1	p_{Approx}	Monte Carlo
3.64	25	1	24	0.050	0.052
4.00	50	1	49	0.050	0.049
4.30	100	1	99	0.049	0.046
4.54	200	1	199	0.052	0.049
4.68	300	1	299	0.050	0.049
4.76	400	1	399	0.051	0.048
4.83	500	1	499	0.050	0.047
4.83	500	1	100	0.043	0.042
4.83	500	1	50	0.035	0.035
4.71	500	1	50	0.059	0.053
4.60	500	1	100	0.11	0.10
4.77	500	1	100	0.056	0.054
4.71	500	3	100	0.056	0.043
4.45	500	3	50	0.12	0.11
4.62	500	1	499	0.11	0.11
5.17	2000	1	1000	0.055	0.042
4.99	1000	1	300	0.055	0.046
4.40	1000	1	300	0.46	0.41
4.30	1000	1	300	0.59	0.51

Simulations indicate that (2.6) is slightly conservative, as expected, so we have used simulated thresholds in comparing SaRa to other procedures in Section 4. Because of the restriction to a symmetric background, SaRa is, on one hand very easy to simulate, but on the other hand can suffer a serious loss of power when change-points are spaced irregularly, with some being close to others.

Table 1 compares simulated values for the maximum of LLR with the approximation given in Theorem 2.1 (with the simple approximation to the function ν) for various values of m (i) when the maximum is constrained by $m_0 \le j - i, k - j \le m_1$ and (ii) for the related Poisson approximation described above when m is sufficiently large compared to b^2 that the tail probability approximation exceeds 0.10. Some of the thresholds are used in comparing different methods in Section 4.

Some numerical experimentation, not reported here in detail, suggests that the approximation of Theorem 2.2 is also reasonably accurate. For example, for $m = 500$, the threshold $b = 4.34$ yields the probability 0.051, while simulations (2500) repetitions give the probability 0.045.

In Section 4, we compare the methods described above with (a) the widely used method of Olshen et al. (2004), which the authors called CBS (for Circular Binary Segmentation) and developed specifically for applications to copy number data, and (b) a threshold based implementation of the multiscale modification of CBS suggested by Frick, Munk and Sieling (2014) (called Multi below). For completeness, we give in (2.9) appropriate approximations for their false positive control. Consider

$$\max_{0 \le j < j+n \le m} Z_{j,n}, \tag{2.7}$$

where

$$Z_{j,n} = \frac{|S_{j+n} - S_j - nS_m/m|}{[n(1-n/m)]^{1/2}} - \{2\kappa \log[3m/n(1-n/m)]\}^{1/2}. \tag{2.8}$$

CBS, which has $\kappa = 0$, is the likelihood ratio statistic for testing the hypothesis of no change-points against the alternative that there exists a pair of changes, where the second change is equal in magnitude but opposite in sign to the first change. The case $\kappa = 1$ is the multiscale statistic of Frick, Munk and Sieling (2014), who argued that CBS puts relatively too much power into the detection of short intervals of large amplitude, but has much less power to detect relatively long intervals of small amplitude.

An approximation for the false positive probability of (2.8) stated here for the case of d-dimensional X_i with covariance matrix Σ (cf. Remark (iii) following the statement of Theorem 2.1), is given by

$$
\begin{aligned}
&\mathbb{P}\Bigg\{\max_{\substack{0\le j<j+n\le m\\ m_0\le n\le m_1}}\Bigg\{\frac{\|\Sigma^{-1/2}(S_{j+n}-S_j-nS_m/m)\|}{[n(1-n/m)]^{1/2}}\\
&\qquad -\Bigg\{2\kappa\log\Bigg[\frac{3m}{n(1-n/m)}\Bigg]\Bigg\}^{1/2}\Bigg\}\ge b\Bigg\}\\
&\quad\sim 2\sum_{n=m_0}^{m_1}(m-n)\,f_d\big(b_n^2\big)\Bigg(\frac{b_n^4q_n^3}{(2n(1-n/m))^2}\Bigg)\\
&\qquad\times\nu^2\Bigg(\frac{b_nq_n}{[n(1-n/m)]^{1/2}}\Bigg),
\end{aligned}
\tag{2.9}
$$

where f_d denotes the chi-square probability density function with d degrees of freedom, $b_n = b + \{2\kappa\log[3m/n(1-n/m)]\}^{1/2}$, and $q_n = 1-(d-1)/b_n^2$. The derivation of (2.9) is similar to that of Theorem 2.1 for $d = 1$, modified as suggested in the proof of (5) of Zhang et al. (2010) for $d > 1$.

The method of proof of Theorem 2.1 also appears to be applicable to some sparse interval systems considered in the literature, although theoretical and/or numerical justification for the approximations that we can formally obtain requires investigation. For example, consider the sparse interval system (2.3) of Chan and Chen (2017) with $T = m$, $h = m$, $r = 1+\epsilon$ with small $\epsilon > 0$. A modification of the calculations used in the proof of our Theorem 2.1 (which is given in the online Supplementary Material) produces the approximation

$$
\begin{aligned}
&\mathbb{P}\Bigg\{\max_{\substack{0\le i<j<k\le m\\ j-i,k-j\in\{\lfloor(1+\epsilon)^a\rfloor:a\in\mathbb{Z}^+\}}}|Z_{i,j,k}|\ge b\Bigg\}\\
&\quad\sim\frac{b^5\varphi(b)}{4}\sum_{\substack{u,v\in\{1,\dots,m\},u+v\le m\\ u,v\in\{\lfloor(1+\epsilon)^a\rfloor:a\in\mathbb{Z}^+\}}}(m-u-v)\prod_{i=1}^{3}\big(4d_i^2\big)\nu(2d_i),
\end{aligned}
$$

where

$$
d_1=\frac{b}{2}\Bigg[\frac{(\lfloor\epsilon u\rfloor\vee 1)v}{u(u+v)}\Bigg]^{1/2},
$$

$$
d_2=\begin{cases}\dfrac{b}{2}\Bigg[\dfrac{1}{u}+\dfrac{1}{v}\Bigg]^{1/2} & \text{if } \big(\lfloor\epsilon v\rfloor\vee 1\big)=\big(\lfloor\epsilon u\rfloor\vee 1\big)=1,\\[2ex] \dfrac{b}{2}\Bigg[\dfrac{2(u^2+v^2+uv)}{uv(u+v)}\Bigg]^{1/2} & \text{otherwise,}\end{cases}
$$

$$
d_3=\frac{b}{2}\Bigg[\frac{(\lfloor\epsilon v\rfloor\vee 1)u}{v(u+v)}\Bigg]^{1/2}.
$$

Here, d_1 and d_3 can be interpreted as the standardized drifts of the local random walks obtained by perturbing i and k, respectively, by the amounts of $(\lfloor \epsilon u \rfloor \vee 1) \times \mathbb{Z}^+$, while d_2 corresponds to perturbing j in one case, and shifting (i, j, k) in the other case. For $m = 500$, $b = 4.83$, $\epsilon = 0.1$, this approximation gives 0.043 and simulation based on 2000 repetitions gives 0.040. For $m = 1000$, $b = 5.1$, $\epsilon = 0.1$, the approximation gives 0.029 and simulation based on 2000 repetitions give 0.023. Similar approximations can be obtained if we allow the threshold b to depend on i, j, k as in Chan and Chen (2017).

3. Confidence regions and local power. We continue to assume independent normal observations $X_1, \ldots, X_m$ with mean values forming a step function with jumps at τ_k, $1 \le k \le M$ and variance equal to one. For a given value of M, we can use the likelihood ratio statistic to construct a joint confidence region for the change-points $\tau = (\tau_1, \ldots, \tau_M)$ or for the change-points and mean values $\mu = (\mu_1, \ldots, \mu_{M+1})$.

We use the inverse relation between confidence intervals and hypothesis tests. For testing a putative value of the positions of change-points and the corresponding mean values, the maximum log likelihood ratio statistic is

$$\begin{aligned} T_{\tau,\mu} &= \max_{0<t_1<\cdots<t_M<m} \sum_{k=1}^{M+1} \frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} \\ &\quad - \sum_{k=1}^{M+1} \left[\mu_k (S_{\tau_k} - S_{\tau_{k-1}}) - \frac{\mu_k^2}{2} (\tau_k - \tau_{k-1}) \right] \\ &=: \max_{0<t_1<\cdots<t_M<m} U_{t,\tau,\mu}, \end{aligned} \tag{3.1}$$

$t = (t_1, \ldots, t_M)$, $t_0 = \tau_0 = 0$, $t_{M+1} = \tau_{M+1} = m$ and $S_i = \sum_{j=1}^{i} X_j$ for $0 \le i \le m$. The $1 - \alpha$ confidence region consists of those τ and μ such that $T_{\tau,\mu} \le a_{\tau,\mu}$ where

$$\mathbb{P}_{\tau,\mu}(T_{\tau,\mu} > a_{\tau,\mu}) = \alpha. \tag{3.2}$$

If we are only interested in the confidence region of τ and treat μ as a nuisance parameter, the maximum log likelihood ratio statistic is

$$T_\tau = \max_{t_1,\ldots,t_M} \sum_{k=1}^{M+1} \frac{(S_{t_k} - S_{t_{k-1}})^2}{2(t_k - t_{k-1})} - \sum_{k=1}^{M+1} \frac{(S_{\tau_k} - S_{\tau_{k-1}})^2}{2(\tau_k - \tau_{k-1})}. \tag{3.3}$$

By sufficiency the conditional distribution of T_τ given $\{S_{\tau_k} : 1 \le k \le M+1\}$ does not depend on μ. Therefore, a $1 - \alpha$ confidence set for the change-points is the set of τ such that $T_\tau \le a_{\tau, S_{\tau_1}, \ldots, S_{\tau_M}}$ where

$$\mathbb{P}_\tau(T_\tau > a_{\tau, S_{\tau_1}, \ldots, S_{\tau_M}} | \tau, S_{\tau_1}, \ldots, S_{\tau_M}) = \alpha. \tag{3.4}$$

In the case there is known to be only one change-point, that is, $M = 1$, for exponentially distributed random variables, the exact value of the left-hand side of (3.4) was given by Worsley (1986). For $M = 1$, asymptotic approximations for the left-hand side of both (3.2) and (3.4) were given by Siegmund (1988b) for distributions from exponential families. Since the asymptotic approximations in Siegmund (1988b) seem difficult to generalize to the case where $M \ge 2$, here we use a different approach to obtain asymptotic approximations for the left-hand side of both (3.2) and (3.4) for $M \ge 1$.

3.1. *Tail approximations*. To construct the joint confidence region for the change-points and the corresponding parameters, for each τ and μ, we need to find $a_{\tau,\mu}$ such that

$$\mathbb{P}_{\tau,\mu}(T_{\tau,\mu} > a_{\tau,\mu}) = \alpha,$$

where $T_{\tau,\mu}$ is defined in (3.1). The following theorem, the proof of which is deferred to the online Supplementary Material (Appendix B), gives an approximation to

$$\mathbb{P}_{\tau,\mu}(T_{\tau,\mu} > a)$$

for large a. We assume that the putative change-points are close enough to the true change-points that the maximum can be taken over relatively small neighborhoods ($|t_k - \tau_k| \le n_k$) of the putative change-points, that is,

$$\mathbb{P}_{\tau,\mu}(T_{\tau,\mu} > a) \sim \mathbb{P}_{\tau,\mu}\Big(\max_{t:|t_k-\tau_k|\le n_k} U_{t,\tau,\mu} > a\Big).$$

The assumption (3.5) imposed in the theorem also ensure that the change-points are reasonably well separated from one another. Despite these technical assumptions, simulation shows that our approximation is reasonably accurate (cf. Tables 2 and 7).

THEOREM 3.1. *Let $\tau = \{\tau_1, \dots, \tau_M\}$ and $\mu = \{\mu_1, \dots, \mu_{M+1}\}$ be defined as above. Define $\delta_k = \mu_{k+1} - \mu_k$ for $1 \le k \le M$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \le k \le M+1$. Suppose that $|\delta_k| \asymp 1$ and*

$$1 \ll a \ll n_k \ll (m_k \wedge m_{k+1})/a, \tag{3.5}$$

where $A \ll B$ means $A/B \to 0$ as $a \to \infty$. We have

$$\mathbb{P}_{\tau,\mu}\Big(\max_{t:|t_k-\tau_k|\le n_k} U_{t,\tau,\mu} > a\Big) \sim \mathbb{P}\bigg(\sum_{k=1}^{M} W_k + \frac{1}{2}\chi^2_{M+1} > a\bigg), \tag{3.6}$$

where $U_{t,\tau,\mu}$ was defined in (3.1), $W_1, \dots, W_M, \chi^2_{M+1}$ are independent, χ^2_{M+1} is a chi-squared random variable with $M+1$ degrees of freedom, and for $1 \le k \le M$ the distribution of W_k is given by

$$\mathbb{P}(W_k > x) = 2\nu(|\delta_k|)e^{-x} - \nu^2(|\delta_k|)e^{-2x}, \quad \forall x \ge 0 \tag{3.7}$$

for $1 \le k \le M$.

We have a similar approximation for the left-hand side of (3.4). Table 3 shows that the approximation is reasonably accurate.

THEOREM 3.2. *Let T'_τ be defined as in (3.3) with the maximum taken over $|t_k - \tau_k| \le n_k$ for $1 \le k \le M$. Define $\hat\delta_k = \hat\mu_{k+1} - \hat\mu_k$ for $1 \le k \le M$, $\hat\mu_k = (s_{\tau_k} - s_{\tau_{k-1}})/(\tau_k - \tau_{k-1})$ and $m_k = \tau_k - \tau_{k-1}$ for $1 \le k \le M+1$. Suppose that $|\hat\delta_k| \asymp 1$ and*

$$1 \ll a \ll n_k \ll (m_k \wedge m_{k+1})/a.$$

We have

$$\mathbb{P}_\tau\big(T'_\tau > a | S_{\tau_1} = s_{\tau_1}, \dots, S_{\tau_M} = s_{\tau_M}\big) \sim \mathbb{P}\bigg(\sum_{k=1}^{M} W_k > a\bigg), \tag{3.8}$$

where $W_1, \dots, W_M$ are independent and have the same distributions as in Theorem 3.1 with δ_k replaced by $\hat\delta_k$.

TABLE 2
Approximation (3.6) for $M = 2$. Simulated values based on $N = 10{,}000$ repetitions. The values of $\hat{p}$ in parentheses correspond to those δ_2 in parentheses

$m/\tau_1/\tau_2$	δ_1	δ_2	a	$\hat{p}$ (Monte Carlo)
105/35/70	1.5	1.5 (−1.5)	6.55	0.052 (0.055)
	2	2 (−2)	6.11	0.048 (0.043)
	2.25	2.25 (−2.25)	5.92	0.058 (0.042)
	1.5	0.75 (−0.75)	6.94	0.060 (0.062)
210/70/140	0.75	0.75 (−0.75)	7.33	0.059 (0.067)
	1.5	1.5 (−1.5)	6.55	0.051 (0.048)
	2	2 (−2)	6.11	0.044 (0.046)
	1.5	0.75 (−0.75)	6.94	0.055 (0.060)

It is easy to evaluate the distributions of $\sum W_k$ and $\sum W_k + \chi^2_{M+1}$ by Fourier inversion, for values of M up to about 100, and by asymptotic methods in the rare case that still larger values of M are of interest. We start from the standard inversion formula for a probability density function f with an integrable characteristic function $\hat{f}$:

$$f(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-\sqrt{-1}\lambda x)\, \hat{f}(\lambda)\, d\lambda.$$

For a distribution on the nonnegative numbers, we integrate this from 0 to b to find that probability to the left of a equals

$$\pi^{-1} \int_0^{\infty} \mathbf{Re}\{[1 - \exp(-\sqrt{-1}\lambda a)]\hat{f}(\lambda)\}\, d\lambda/\lambda.$$

For our special case, for simplicity assume that $\delta_k = \delta$ for all k. Let $\nu = \nu(\delta)$ and $\hat{f}(\lambda) = (1-\nu)^2 + 2\nu/(1+\sqrt{-1}\lambda) - 2\nu^2/(2-\sqrt{-1}\lambda)$ denote the characteristic function of W_k. Let $\hat{g}(\lambda)$ be the characteristic function of a χ^2_{M+1} random variable. Finally, let $h(\lambda) = \hat{f}^m(\lambda) * g(\lambda)[1 - \exp(\sqrt{-1}\lambda a)]/(1+\sqrt{-1}\lambda)$. Then the probability on the right-hand side of (3.6) equals $1 - \int_0^{\infty} \mathbf{Re}[h(\lambda)]\, d\lambda/\pi$. For Theorem 3.2, a similar expression without the factor h provides a numerical value for the approximation.

In Tables 2 and 3, we use simulations to check the accuracy of the approximations of Theorems 3.1 and 3.2, respectively. The number of change-points is $M = 2$. The other parameters are indicated in the tables. For different values of δ_1 and δ_2, we compute the threshold a such that our approximation of the relevant probability equals 0.05. The values $\hat{p}$ denotes a Monte

TABLE 3
Approximation (3.8) for $M = 2$. Simulated values based on $N = 10{,}000$ repetitions. The values of $\hat{p}$ in parentheses correspond to those $\hat{\delta}_2$ in parentheses

$m/\tau_1/\tau_2$	$\hat{\delta}_1$	$\hat{\delta}_2$	a	$\hat{p}$ (Monte Carlo)
105/35/70	1.5	1.5 (−1.5)	4.28	0.054 (0.047)
	2	2 (−2)	3.80	0.045 (0.047)
	2.25	2.25 (−2.25)	3.59	0.043 (0.042)
	1.5	0.75 (−0.75)	4.68	0.053 (0.056)
210/70/140	0.75	0.75 (−0.75)	5.09	0.055 (0.057)
	1.5	1.5 (−1.5)	4.28	0.049 (0.047)
	2	2 (−2)	3.80	0.049 (0.047)
	1.5	0.75 (−0.75)	4.68	0.056 (0.055)

Carlo estimate of the appropriate probability with $n_1 = n_2 = m$, based on 10,000 repetitions each. We see that the approximations are reasonably accurate for the range $1 < |\delta| < 2$.

For a simple example of a confidence region for the change-points, we simulated $m = 161$ observations with changes in the mean value of size ± 2 at observations 51, 91 and 121. In the first simulation, $\hat{\delta} \approx 2$ for all three change-points. This value gave a threshold of 4.95 for a 95% conditional confidence region. The joint confidence region consisted of the point estimators 51, 91 and either of 121 or 122. In a second simulation with the same parameters, the smallest estimate of $\hat{\delta}$ was 1.5, which if used for all three change-points would lead to a conservative threshold, in this case equal to 5.6. The joint confidence region based on this threshold was substantially larger. The union of the three regions was 50, 51, 91, 92, 93, 121, 122. The joint confidence region consisted of 7 of the $2 \times 3 \times 2 = 12$ possible combinations of these values; we omit the details. When the size of the changes was decreased to ± 1.5, we again used the smallest value of $\hat{\delta}$, which again gave a threshold of 5.6, and the 95% joint confidence region extended up to 5 observations away from the change-points at 51 and 91, and a couple of observations away from 121. As a reflection of the fluctuations in the sample paths of the random walk, the regions around the individual change-points were neither symmetric nor connected.

For applications to copy number variation, see Section 4.2.

REMARK. As one sees from an examination of the conditions of Theorem 3.1 and Theorem 3.2, the methods discussed in this section work well if the sizes of the changes and the distances between them are reasonably large. If there is a mixture of large and small changes, or if it is unclear whether a putative change is real or not, the procedure can be adapted appropriately. For example, suppose we are interested in a joint confidence region for the change-points when there are clear changes close to $\tau_1 < \tau_3$, with what may or may not be a change at $\tau_2 \in (\tau_1, \tau_3)$. In taking the maximum indicated above, one can fix the value $t_2 = \tau_2$ and maximize only over t_1 and t_3, while evaluating the conditional probabilities given by all three τ_i. Whether there is a change at τ_2 or not, the conditional probability adapted from Theorem 3.2 now involves the sum of two conditionally independent maxima, not three. To be more conservative in protecting against a change-point near, but not exactly at, τ_2, we can bracket τ_2 by, say $\tau_{20} < \tau_2 < \tau_{21}$, and proceed from there. The confidence coefficient is still asymptotically as given in Theorem 3.2, but the confidence region itself may have changed due to the change in the statistics used for conditioning. Presumably unnecessary conditioning leads to less accurate estimation.

3.2. *Power.* To help our intuition concerning the relation between background and size of a change that makes a particular change-point either easy or difficult to detect and to compare different procedures under hypothesized conditions, it is helpful to have an approximation for the power to detect a change.

When the size of a change in the mean value is $\delta > 0$ and the (largest possible) background is (i^*, k^*) for a change-point at j^*, we define the *marginal power* to be

$$1 - \Phi\bigl(b - \delta\bigl[h_1 h_2/(h_1 + h_2)\bigr]^{1/2}\bigr), \tag{3.9}$$

where $h_1 = j^* - i^*$, $h_2 = k^* - i^*$. This is just the marginal probability that the statistic $Z_{i,j,k}$ evaluated at the true change-point $j = j^*$ with the largest possible background $i = i^*$, $k = k^*$ exceeds the threshold b. A detection may fail to occur at i^*, j^*, k^*, but occur at i', j', k' which is a local perturbation of the values i^*, j^*, k^* in the sense that $i^* \le i' < j' < k' \le k^*$. Using a similar argument as in the derivation of (3.6) (cf. Appendix B in the online supplement), we can approximate the probability of such a detection by conditioning on $Z^2_{i^*,j^*,k^*}$ to obtain

$$2\int_0^{b^2/2} \mathbb{P}\Bigl\{\sum_{i\in\{0,1,2\}} W_i > b^2/2 - x\Bigr\} f(2x; 1, \lambda)\,dx, \tag{3.10}$$

where $f(\cdot; 1, \lambda)$ is the probability density function of a χ^2 distribution with one degree of freedom and noncentrality parameter $\lambda = \delta^2 h_1 h_2/(h_1 + h_2)$, W_0, W_1, W_2 are independent, W_0 is nonnegative and has the probability distribution $\mathbb{P}\{W_0 > x\} = 2\nu(\Delta)\exp(-x) - \nu^2(\Delta)\exp(-2x)$ for $x \geq 0$ with $\Delta = b\sqrt{1/h_1 + 1/h_2}$, and for $i = 1, 2$, W_i is nonnegative and has the distribution given by $\mathbb{P}\{W_i > x\} = \nu(\Delta_i)\exp(-x)$ for $x \geq 0$ with $\Delta_1 = \Delta/(1 + h_1/h_2)$ and $\Delta_2 = \Delta/(1 + h_2/h_1)$. We use the term *local power* to denote the sum of the marginal power (3.9) and the perturbation (3.10). Similar approximations can be obtained for the pseudo-sequential procedure, for multidimensional statistics and for multiscale statistics. We omit the details.

4. Simulations and applications. In this section, we report the result of numerical exercises involving simulated and real data to compare a number of different segmentation procedures, with emphasis on their efficiency to detect change-points without an excessive number of false positive errors. We consider only thresholding algorithms that control the false positive error rate under the global null hypothesis that there are no change-points. As we see below on the basis of simulations that control is compromised to varying degrees when iteration to find multiple change-points is required.

In contrast to LLR and SaRa, both CBS and Multi are "top down" procedures, where we begin by searching the entire interval of observations. When one change-point (resp., a pair of change-points) is detected, the interval searched is divided into two (resp., three) subintervals, and those subintervals are searched for additional change-points. Since the methods are designed to detect change-points occurring in pairs, under various conditions, for example, when there is only one change-point to be detected in a search interval, or when consecutive changes are both positive or both negative, one of the paired "detections" often suggests a change-point very near to one end-point of the search interval. This is usually a false detection that is easy to recognize and disregard, although the decision to disregard it has an element of subjectivity. To minimize this subjectivity in our simulations, after some experimentation we usually discard any detection having a distance to an end-point of the interval searched that is within 5% of the length of that interval. If both detections are within this distance, the one closer to an end-point is discarded. If they are equally distant from an endpoint, the one to be discarded is chosen at random. While objective, this rule can in some cases lead to errors, so in practice we recommend making a subjective decision based on a careful examination of the data.

Although the top down iterations of CBS and Multi make it natural to suspect that their false positive error control may be inadequate, in most cases this does not appear to be a major problem. If a large interval is partitioned into smaller intervals by correctly detected change-points, the false positive probability for CBS for the initial interval is numerically very close to the sum of the probabilities for the subintervals, so the sum of the false positive probabilities for the small intervals is roughly the same as that of the initial search. For Multi, this sum is much less than the false positive probability of the initial search (provided the value of m is used for all searches, not changed to reflect the lengths of the different subintervals). It appears that for both of these statistics the main source of false positive errors arises, fortunately not often, when a correct detection is paired with a false detection that is not close enough to an endpoint to be excluded.

It is also possible to give approximations for the local power for these two statistics, at least under the simplifying conditions that we are at a stage of the search where there is only one or a pair of change-points to be detected in the interval searched. For simplicity, we consider only the CBS statistic when either (i) the mean before the first change-point at τ_1 equals μ_1, between the first and second change-point at τ_2 equals μ_2, and returns to the value μ_1 after τ_2, or (ii) there is only one change-point at τ_1 and $\tau_2 = m$. Denote the magnitude of the

change by δ and let $n_0 = \tau_2 - \tau_1$ denote the length of the changed interval. Approximations and some calculus similar to that given in Section 3.2 lead to

$$\begin{aligned}
&\mathbb{P}_\delta\left\{\max_{0\le i<j\le m}\frac{|S_j - S_i - (j-i)S_m/m|}{[(j-i)(m-j+i)/m]^{1/2}} \ge b\right\}\\
&\quad\approx \Phi\big(\delta\big[n_0(1-n_0/m)\big]^{1/2} - b\big)\\
&\qquad + 2\int_0^{b^2/2}\mathbb{P}\big(W_3 + W_4 \ge b^2/2 - x\big) f\big(2x; 1, \delta^2 n_0(1-n_0/m)\big)\,dx,
\end{aligned} \tag{4.1}$$

where $f(\cdot; 1, \lambda)$ is the density function of the chi-squared distribution with 1 degree of freedom and noncentrality parameter λ, and W_3, W_4 are independent nonnegative random variables similar to those appearing in the approximation for the local power of LLR (cf. (3.10)). If $\tau_2 < m$, both have the distribution given by $\mathbb{P}(W_3 > x) = 2\nu(\Delta)\exp(-x) - \nu^2(\Delta)\exp(-2x)$ for $x \ge 0$, where $\Delta = b/\sqrt{n_0(1-n_0/m)}$. If $\tau_2 = m$, the right-hand tail of the distribution of W_4 equals $\Delta\exp(-x)$. Similar results hold for the local power of Multi and of SLLR.

Some other top down iterative thresholding procedures appear to have poorly controlled false positive error rates if iterated with the same threshold and poor power under easily understood conditions. In particular, we mention here but do not consider the thresholding procedure suggested by Aston and Kirch (2012), which is similar to CBS and Multi in the sense that it searches for a complementary pair of change-points; but the statistic $S_j - S_i - (j-i)S_m/m$ to be maximized over $i < j$ in searching for change-points is not standardized to have a marginal distribution with unit variance when there are no change-points. Its false positive probability can be approximated by a modification of the arguments of this paper and is found to be poorly controlled when the procedure is iterated with a fixed threshold to find multiple change-points. It also has very poor power to detect nearby change-points that move in opposite directions. For example, assume that in 300 observations there are only two changes: an increase of 3 standard deviations at observation 100 followed by a decrease of the same magnitude 7 observations later. The Aston and Kirch (2012) procedure with a 0.05 false positive error rate when there are no changes, either detects neither of the two indicated changes or detects changes with a value of the statistic that is maximized at $i \ll 100$ and $j \gg 107$, which give an incorrect impression of the size of the interval. This latter possibility is a reflection of the bias of the statistic in favor of up-down (or down-up) changes that occur relatively much farther apart than the example given here. A procedure that has similar weaknesses is the classical binary segmentation (BS) procedure (Vostrikova (1981)), but in view of its apparent popularity, we do include it in Table 6 below. Siegmund (1985), Chapter XI, among others, has given an approximation for the false positive probability, which turns out to be approximately logarithmic in the length of the sequence. Hence, when the statistic is applied iteratively over short intervals of data to search for multiple change-points, the false positives increase quickly with the number of iterations. For the up-down pair of changes described above, it is easy to see by computing mean values that BS has almost no power. By way of contrast, the other procedures we have discussed would detect both change-points with probability almost 1, as can be seen by simulations or by the local power approximations discussed above.

We would like to emphasize that our goal in this paper is to develop methods to study change-point problems. The variety of problems is such that the method of choice in a particular problem may well depend on what can be expected by way of the number and configuration of change-points. The remarks of the preceding paragraph notwithstanding, procedures that appear inadequate for some problems may be the method of choice for others.

Tables 4 and 5 do not include WBS, but the results in Table 6, where it is included, suggest that it would perform similar to LLR and SLLR, with the major difference being that we cannot without additional simulations give a threshold bringing WBS close to the others in regard to false positive errors.

4.1. *Simulations*. For the numerical examples of this section, our implementation of both LLR and SaRa was to choose the values j by minimizing the associated length of the background $k - i$ from among those values of $|Z_{i,j,k}|$ exceeding the threshold. If necessary, we enforced the condition mentioned above that the backgrounds not overlap. For SLLR initialized at i, we choose the maximum value of j associated with the smallest k such that $\max_{i<j<k} |Z_{i,j,k}|$ exceeds the relevant threshold. The other possibilities mentioned above, to choose the value j giving the largest Z-value (subject to the no overlap condition) rarely leads to a substantially different segmentation. Our preference here is simply to have a definite algorithm, although it may not always make the best choice.

The first example in Table 4 is a modified version of a suggestion of (Olshen et al. (2004)), which those authors said was typical of the copy number data that motivated their study. There are three hundred observations and four change-points at 138, 199, 208 and 232, with mean values 0.0, 0.75, 2.5, 0.25 and 1.5 in the five gaps between change-points. According to the local power approximation of the preceding section, LLR has local power 0.77, 0.73, 0.91 and 0.85, respectively, to detect these change-points, so its expected number of change-points detected is 3.3. As a reflection of its lower threshold, SLLR has an expected number of 3.45 detections, although, as we remarked above and see empirically in Table 6 below, it also has a larger rate of false positives. Simulations not reported here indicate that these approximate expected values are quite accurate.

The second example in the table has changes of the same magnitude in the same locations, but with all changes in a positive direction. The results are similar in spite of the fact that both CBS and Multi are not designed with this case in mind. The third case is qualitatively similar to the first one, but it contains one very short up-down pair of change-points. In this case, the expected number of change-points detected by LLR is predicted by our local power approximation to be 3.4.

Failure to detect a change-point is marked in the table by a zero (0), and false positives by an asterisk (*).

Although the table contains only a few examples, several entries reinforce our intuition. The procedure SaRa lacks power to detect both of two nearby changes by virtue of its requirement to use a symmetric background. The procedure Multi fails to detect a short interval that CBS detects—not surprising since its justification involved an increase in power to detect longer intervals paid for by a decrease in power to detect very short intervals.

In Table 4, we see only a few false positive errors. For CBS and Multi there is a false positive error that occurred when searching an interval where there is only one true change-point to be detected. The statistics detect two, and the incorrect detection is not eliminated by the 5% rule described above. Other simulations suggest that this is the most commonly occurring false positive error of those statistics.

As mentioned in Section 1, in studying CNV various authors starting with Olshen et al. (2004) have found in their data technical artifacts in the form of local trends that tend to disrupt the idealized model of a step function mean value. The local trends appear to be affected primarily by CG content, which oscillates in a roughly sinusoidal fashion. To test robustness against these perturbations Olshen et al. (2004) suggest adding a low frequency sinusoid, which produces some degradation of performance. In Table 5, we report a very small simulation comparing only LLR and CBS in the presence of a sinusoidal perturbation of the mean values. In the first four rows, the amplitude and frequency are slightly larger than

TABLE 4

Examples of segmentations: $m = 300$, $b_{\mathrm{LLR}} = 4.68$, $b_{\mathrm{SLLR}} = 4.21$, $b_{\mathrm{SaRa}} = 4.27$, $b_{\mathrm{CBS}} = 4.23$ *and* $b_{\mathrm{Multi}} = 1.51$. *The initial mean value is* 0. *Locations of change-points and mean values after the change are as indicated*

Procedure/Parameters	138, 0.75	199, 2.5	208, 0.25	232, 1.5
LLR	164	198	206	248
SLLR	134	197	206	248
SaRa	48*, 140	198	207	249
CBS	149	199	207	249
Multi	149	199	208	249
LLR	127	198	211	230
SLLR	127	199	209	230
SaRa	130	0	211	230
CBS	135	199	212	231
Multi	135	199	212	231

Procedure/Parameters	138, 0.75	199, 2.5	208, 4.75	232, 6.0
LLR	145	198	207	234
SLLR	134	197	206	232
SaRa	134	0	207	234
CBS	140	199	208	235
Multi	140	199	208	235
LLR	137	0	207	231
SLLR	136	197	206	235
SaRa	137	199	208	235
CBS	138	0	207	236
Multi	138	0	207	236
LLR	159	198	207	231
SLLR	129	198	209	231
SaRa	131	198	0	231
CBS	130	199	208	232
Multi	130	199	208	232

Procedure/Parameters	100, 3.0	103, −0.5	120, 1.8	200, 2.5
LLR	97	102	119	199
SLLR	97	101	117	199
SaRa	0	0	119	200
CBS	98	103	120	200
Multi	0	0	120	200
LLR	98	102	119	200
SLLR	96	101	118	199
SaRa	0	104	119	200
CBS	100	103	120	201
Multi	0	0	120	207
LLR	99	101	122	214
SLLR	98	0	117	214
SaRa	99	0	121	214
CBS	100	102	122	140*, 215
Multi	100	102	122	140*, 215

TABLE 5

Examples with sinusoidal local trends: $m = 200$, $b_{\mathrm{LLR}} = 4.54$, $b_{\mathrm{CBS}} = 4.13$. *Format as in Table 4, but to simulate local trends, in the first four rows* $0.2\sin(0.1k + U)$ *is added to the kth mean value, where U is a uniformly distributed random phase. For the next four rows, the addition to the kth mean value is* $0.4\sin(0.05k + U)$, *and in the last four rows it is* $0.7\sin(0.03k + U)$

Procedure/Parameters	60, 3.0	63, −0.2	83, 1.1	153, 2.0
LLR	59	63	83	152
CBS	60	63	83	152
LLR	60	63	78	155
CBS	60	63	83, 141*	155
LLR	60	63	99	0
CBS	60	63	99	0
LLR	0	0	83	146
CBS	0	0	83	146
Procedure/Parameters	50, 1.5	65, −0.1	125, 1.2	145, 2.6
LLR	42	66	127	144
CBS	42	65, 73*	126	144
LLR	43	0	130	0
CBS	25*, 43	0	0	143
LLR	50	65	122	145
CBS	50	65	0	144
LLR	49	65	125	145
CBS	49	65	125	145
Procedure/Parameters	50, 2.0	60, 0.0	135, 3.0	140, 0.0
LLR	50	60	135	140
CBS	50	60	121*, 135	140, 181*
LLR	50	60	135	140
CBS	47	60, 82*	135	140, 184*
LLR	50	60	135	140
CBS	50	60	135	140
LLR	49	60	135	140
CBS	49	63	135	140

those suggested by Olshen et al. (2004). In the next four rows the amplitude is still larger and the frequency is relatively small. These and other simulations, not shown here, suggest that modest local trends lead to slight increases in the false positive rate of CBS (and Multi), but less so for LLR, and to slight decreases in the power of detection of all methods. The local trends in the last four rows have a large amplitude and small frequency. Without these local trends, the change-points would be easy to detect, and the up-down pairs are ideal for CBS. The local power approximation of Section 3.2 indicates that in the absence of the local trends local power to detect each of the four change-points averages about 0.95. Indeed, each change-point is detected in these experiments, but there is a striking increase in false positives for CBS, presumably because the top-down nature of the search for change-points makes it particularly vulnerable to oscillations in the background.

Table 6 provides the outcomes of 1000 simulations for detecting M change-points randomly located from 0 to 500. The sizes of the changes are normally distributed with mean

TABLE 6

Random change-points, $m = 500$, $b_{\mathrm{BS}} = 3.25$, $b_{\mathrm{LLR}} = 4.83$, $b_{\mathrm{LLR\text{-}F}} = 4.83$, $b_{\mathrm{SLLR}} = 4.33$, $b_{\mathrm{WBS}} = 4.62$, $b_{\mathrm{SaRa}} = 4.42$, $b_{\mathrm{CBS}} = 4.36$, $b_{\mathrm{Multi}} = 1.57$. *The three rows in each entry are the number of times that the number of change-points is correctly detected, under detected and over detected, respectively, in* 1000 *repetitions. The accompanying numbers in parentheses are the number of change-points missed (false negative errors) and the number of over detections (false positive errors). E(asy) denotes the number of repetitions where all methods detected the correct number of change-points; I(mpossible) gives the number of repetitions where no method detected the correct number of change-points*

M	BS	LLR	LLR-F	SLLR	WBS	SaRa	CBS	Multi	E/I
0	955	949	954	953	939	948	953	961	846/3
	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
	45 (47)	51 (68)	46 (59)	47 (83)	61 (79)	52 (54)	47 (90)	39 (74)	
3	739	835	833	848	820	791	805	820	574/65
	90 (114)	133 (146)	136 (150)	110 (122)	128 (141)	175 (197)	105 (114)	92 (104)	
	171 (183)	32 (38)	31 (37)	42 (70)	52 (60)	34 (35)	90 (143)	88 (122)	
5	625	656	646	675	643	572	634	652	352/150
	190 (257)	325 (388)	335 (401)	277 (325)	306 (371)	404 (495)	255 (305)	248 (302)	
	185 (215)	19 (20)	19 (20)	48 (60)	51 (51)	24 (27)	111 (142)	100 (120)	
8	408	324	316	404	344	257	412	400	107/316
	396 (647)	670 (1037)	678 (1052)	567 (835)	633 (961)	735 (1255)	472 (682)	491 (732)	
	196 (235)	6 (7)	6 (6)	29 (37)	23 (24)	8 (8)	116 (144)	109 (126)	

value 2.5ξ, where the values ξ are independently ± 1 with probability 1/2 and variance 0.5. The first method is the classical binary segmentation method with the threshold $b = 3.25$, so the probability of a false positive is about 0.05 (Siegmund (1985), Chapter XI). The second uses the LLR statistic with segmentation based on the smallest value of $k - i$ for which the statistic exceeds the 0.05 level threshold $b_{\mathrm{LLR}} = 4.83$; the third is a faster LLR procedure introduced in the following paragraph; the fourth is SLLR with the threshold 4.33; the fifth is the Wild Binary Segmentation (WBS) procedure of Fryzlewicz (2014) with 5000 random segments and threshold $b_{\mathrm{WBS}} = 4.62$, which by our simulations also provides a false positive rate of about 0.05 when there are no change-points. (This threshold is close to, but slightly larger than the value $1.3[2\log(m)]^{1/2} = 4.58$ recommended by Fryzlewicz, which was presumably also determined by numerical experimentation.) The fifth is SaRa with the (simulated) threshold 4.42. The last two procedures are CBS and Multi with thresholds that, according to (2.9), control the false positive error rates to be approximately 0.05 when there are no change-points.

Since LLR requires order m^3 computations, it can be slow for large values of m. A possible speed-up is based on the observation that for the large backgrounds required to detect relatively small changes, determining the exact background does not seem to be important. Suppose that in considering a fixed value of j, to determine an appropriate k, we choose $k = j + 1$, then choose a new value of k recursively as the old value plus $\max(1, [(k - j)/10])$, where $[x]$ denotes the largest integer less than or equal to x. Thus, for $k - j < 20$, we choose every integer, then every second integer for $k - j < 30$, etc. The computational complexity of this procedure is of order $m(\log(m))^2\ell^2$. In Table 6, this procedure with $\ell = 10$ is denoted LLR-F. A second appealing speed up would be to use a random number of intervals along the lines of WBS, but (unlike WBS) impose a no overlap condition in the process of change-point detection and identification.

All procedures lose power with an increasing number of change-points. The top down procedures, BS, CBS and Multi seem to lose less power, but the price they pay is an increasing number of false positives. This is especially severe for BS. By raising the threshold for BS, we are able to control the false positive rate, but at a substantial loss of power when there are several change-points. For the LLR threshold of $b = 4.83$ suggested by a referee, a simulation produced in the cell for $M = 5$ the numbers 550, 414 (620), 27 (36) and for $M = 8$ the numbers 218, 751 (1579), 31 (33), which are uniformly worse than for LLR. This and other numerical experiments, in addition to the difficulties with up-down pairs of changes discussed above, led us to believe that BS could not compete with the other procedures we have studied unless the change-points are small in number and well separated. The bottom up procedures, LLR and SaRa, have decreasing false positive rates with an increasing number of change-points, but they experience a greater loss of power than the top-down procedures. SLLR and WBS, which are not strictly top-down nor bottom-up, seem slightly more stable than the others in regard to both false positive control and power. As expected, SaRa has problems with detection of near-by change-points.

If scientific considerations, knowledge of similar data, and/or a plot of the data give a rough guide to the number of change-points, one might compensate for the behavior discussed above by decreasing the threshold of the bottom-up methods or increasing it for the top-down methods if multiple change-points can be anticipated.

Although the simple counts in Table 6 without an indication of accuracy of the detections are not definitive, as we see in Tables 4 and 5, in most cases accuracy of estimation of the change-points is less an issue than the errors of over or under detection.

4.2. *Array CGH data.* In this section, we present examples involving changes in copy number from array CGH data.

We first consider the test cases GBM29 and GBM31 used by Lai et al. (2005) to compare different methods of segmentation.

For GBM29, the total length of the sequence is 193. The estimated standard deviation is 0.76. The theoretical 0.05 thresholds for LLR and SLLR are 4.53 and 4.07, while that for CBS is 4.12 and for Multi is 1.45. Change-points are detected at

$$81, 85, 89, 96, 123, 133$$

by all methods.

For GBM31, the length of the sequence is 797. The estimated standard deviation is 0.38. All methods, except the multiscale statistic detect the same set of change-points, at

$$317, 318, 538, 727, 728.$$

The third change-point is a relatively small change apparently indicating a long region of loss of copy number; the first two and last two change-points are large spikes. Only one of the two is detected by the multiscale statistic, which is designed to favor detection of longer intervals.

We have also tested our methods on the BT474 cell line data from Snijders et al. (2003). See Pollack et al. (1999, 2002) for a different experimental technique involving BT474 and a discussion of the implications for breast cancer. This cell line has also been used by, for example, Zhao et al. (2004), who based their experimental technique on SNPs rather than array CGH.

For a scan of the entire genome, which involves slightly more than 2000 observations, we detect 63 change-points with LLR at a 0.05 genome-wide significance threshold of $b = 5.2$; and we detect 67 using SLLR with a threshold of 4.7. However, the data are organized by chromosomal location, and it turns out that the estimated standard deviation varies considerably from chromosome to chromosome. Although the cited literature typically involves scans of the entire genome, we find a scan of each chromosome using the estimated standard deviation of that chromosome more reasonable.

We continue to use genome wide thresholds, which are 4.68 for CBS and 1.67 for Multi; but we now use standard deviations specific to each chromosome. Particularly interesting are chromosome 17, where an increase in copy number appears to have implications for the severity of breast cancer, and chromosome 20, which appears to contain a second increase in copy number embedded in a modest increase in copy number. For chromosome 17, there are $m = 87$ observations, with an estimated standard deviation of 0.51. According to LLR, SLLR, CBS and Multi, there is an increase in copy number at the 35th observation (17q11.2–12), with a change back to baseline just two observations later. There is a second increase at the 50th observation (17q21.3) and a return to the baseline at the 66th (17q23). Chromosome 20 contains $m = 85$ observations, and the standard deviation is 0.59. LLR, SLLR, CBS and Multi again agree and detect a *decrease* in copy number from the 38th (20q11.2) to the 52nd observation, followed by an increase from the 53rd (20q13) to the 68th (20q13.1). From the 69th observation, there is an even larger increase until the 82nd (20q13.3), then a return to roughly the baseline value for the last three observations.

Also interesting are chromosomes 4, 5 and 11, all of which have several changes, and some of the changes are followed by a second change after only a few observations. On chromosome 4, there are 162 observations and an estimated standard deviation of 0.19. At the 0.05 global significance level, LLR and SLLR detected changes at 7, 8, 59, 61, 141, 143 and 155. CBS and Multi detected the same changes with the exception of 143, which both missed. On chromosome 5, there were 99 observations and an estimated standard deviation of 0.16. Changes were detected by all four methods at 25, 45, 51 and 65. CBS and Multi also detected paired changes at 87 and 91. The first of these was missed by LLR, and both were

missed by SLLR. On chromosome 11, there are 181 observations and an estimated standard deviation of 0.34. Changes were detected by all four methods at 91, 124, 139, 144, 162, 165. In this case, SLLR also detected changes at 6 and 163. Looking at a plot of the cumulative sum of the data and the proximity of the statistic to the detection threshold suggests that the change at 163 is a false positive. The putative change at 6 is also borderline, but looks real in the cumulative sum plot.

To illustrate our confidence region calculations, we consider Chromosome 3, where there are 85 observations and change-points are detected at 19, 39 and 44. The estimated size of the change at 44 is $\hat{\delta} = 2.25$, while the changes at 19 and at 39 are estimated to be substantially larger. For simplicity, we (conservatively) use the single estimated difference, $\hat{\delta} = 2.25$, so from the theory developed above, the critical constant for a 95% joint conditional confidence region for the three change-points is 4.63. Using this threshold, a joint confidence region consists of the exact point estimates 19 and 39, and the union of 43, 44 and 45. For Chromosome 15, change-points are detected at observations 43 and 57, where the smaller change is estimated to be about 2.3 and the other only slightly larger. For the approximate threshold of $b = 3.6$, we found a 95% joint confidence region to consist of the four pairs 42 or 43 and 56 or 57. For Chromosome 20, where we detected change-points at 38, 52, 68 and 82, the smallest value of $\hat{\delta}$ is 2.1 at 68. Using this single estimator, we find that the critical constant for a 95% joint confidence region for four change-points is $b = 5.9$. The union of the values that in various 4-tuples form the joint confidence region are 38, 39, 51, 52, 66, 67, 68 and 82.

REMARK. In studying copy number variation, it is customary to plot the locus by locus measurements, which should be about equal to zero when the copy number is two, with positive values indicative of amplifications and negative values indicative of deletions. There may be advantages to plotting the consecutive partial sums also and looking for a change in slope to indicate an increase or decrease in copy number. This plot is substantially smoother, and changes in slope that are candidates for change-points in copy number are often easier to see than in a plot of the raw data. The disadvantage is that it is sometimes difficult to infer the regions of normal copy number, which are regions where the slope should be zero although it seems that it is always different from zero.

4.3. *Simulations for confidence intervals.* In order to illustrate the size of the joint confidence regions introduced in Section 3, we consider in Table 7 some parameter settings related to Table 3. The upper part of the table, like Table 2, gives the estimated coverage probability based on 10,000 simulations for examples where the threshold a has been selected so our theoretical approximation gives the probability 0.05. The lower part of the table gives the probability from 1000 simulations that the indicated values of t_1, t_2 are *not* contained in the confidence region. We have chosen values of t_i for which this probability is about 0.5, so one can regard the difference between t_i and τ_i as a rough measure of the size of the confidence region when all other parameters are set to their correct values. Recall that $\delta_i = \mu_i - \mu_{i-1}$ denotes the size of the change at τ_i.

The rows beginning with 0.65 are particularly interesting, since they show that the relatively small change at $\tau_1 = 138$ compared with very large change at $\tau_2 = 225$ leads to substantially more uncertainty in the value of τ_1 compared to the value of τ_2.

4.4. *Comparison with other confidence intervals.* Frick, Munk and Sieling (2014) suggested a different method to construct a confidence region jointly for the change-points and the mean values of the observations in the segments connecting those change-points. For each candidate set of change-points τ and mean values μ, they suggest an application of their multiscale statistic

$$\max\left(\frac{|S_j - S_i - (j-i)\mu|}{(j-i)^{1/2}} - \left[2\log(3m/(j-i))\right]^{1/2}\right), \tag{4.2}$$

TABLE 7

Likelihood ratio based joint confidence intervals. $\hat{p}$ is the simulated probability that the parameters t_1 and t_2 are rejected when the true parameter values are τ_1 and τ_2. Nominal confidence level is 0.05. *Simulations are based on* 10,000 (1000) *repetitions in the first four* (*last* 12) *rows*

δ_1	δ_2	a	τ_1, τ_2	t_1, t_2	$\hat{p}$ (Monte Carlo)
2.13	1.33	6.4	9, 33	9, 33	0.049
2.5	4.0	5.35	87, 104	87, 104	0.051
0.65	2.5	6.65	138, 225	138, 225	0.047
1.73	2.13	6.23	57, 66	57, 66	0.049
2.13	1.33	6.4	9, 33	7, 33	0.59
2.13	1.33	6.4	9, 33	11, 33	0.58
2.13	1.33	6.4	9, 33	9, 29	0.47
2.13	1.33	6.4	9, 33	9, 37	0.44
0.65	2.5	6.65	138, 225	138, 227	0.75
0.65	2.5	6.65	138, 225	138, 223	0.73
0.65	2.5	6.65	138, 225	120, 225	0.49
0.65	2.5	6.65	138, 225	156, 225	0.46
2.5	4.0	5.35	87, 104	87, 102	0.43
2.5	4.0	5.35	87, 104	87, 106	0.44
2.5	4.0	5.35	87, 104	86, 104	0.89
2.5	4.0	5.35	87, 104	88, 104	0.89

where the maximum is taken over all $i < j$ within one of the segments of $(0, \tau_1], \ldots, (\tau_M, m]$, and μ is the hypothesized mean value in the segment. This is in effect a test of the hypothesis that there are no change-points in the hypothesized segments $(0, \tau_1], \ldots, (\tau_M, m]$ and the mean values are as hypothesized. Worsley (1986) discusses a similar idea under the assumption that there is a single change-point, and one is interested only in a confidence region for the change-point, not a joint confidence region for change-points and means. (Note that our approximation (3.8) allows us to condition on the sum of the observations in the interval under investigation, and hence use these ideas to obtain joint confidence regions for the change-points alone.)

It is difficult to make a comparison of the two methods. In Table 8, we compare our confidence region defined by (3.2) with that using (4.2) in a small number of examples. We set $m = 200$, $\tau_1 = 50$, $\tau_2 = 100$ and consider values of the δ_i that are large enough that most of

TABLE 8

Power to detect departure from true parameter values: $\tau = (50, 100)$ *and* μ *as given*; t *and* ξ *are hypothesized values of* τ *and* μ. *The subscript* 1 *indicates the likelihood ratio procedure, while* 2 *indicates the procedure based on* (4.2). *Simulations are based on* 10,000 *repetitions*

μ	ξ	t	$\widehat{\text{Power}}_1$ (Monte Carlo)	$\widehat{\text{Power}}_2$ (Monte Carlo)
0.0, 1.0, 0.0	0.0, 1.0, 0.0	55, 95	0.64	0.08
0.0, 1.0, 0.0	0.1, 0.9, −0.2	55, 95,	0.87	0.40
0.0, 1.0, 0.0	0.1, 0.9, −0.2	40, 100	0.75	0.32
0.0, 1.2, 2.0	0.0, 1.2, 2.0	47, 105	0.47	0.044
0.0, 1.2, 2.0	0.0, 1.5, 1.9	47, 105	0.75	0.32
0.0, 1.5, 0.75	0.1, 1.4, 0.9	40, 97	0.96	0.69
0.0, 1.5, 0.75	0.0, 1.5, 0.75	44, 98	0.81	0.32
0.0, 1.2, −0.1	0.1, 1.1, 0.1	48, 103	0.68	0.22
0.0, 1.1, 0.1	−0.2, 1.0, 0.0	52, 115	0.91	0.44
0.0, 1.0, 2.0	−0.1, 1.1, 2.1	45, 110	0.87	0.24

the time we will detect two change-points. The problem becomes one of locating them and estimating the mean values. For our confidence regions, we choose the thresholds $a = 7.2$ so that the probability in (3.6) equals 0.05. This threshold was confirmed by simulation. Moreover, for the statistic (4.2), we chose the threshold $b_2 = 1.44$ for which a 20,000 repetition simulation experiment gave the probability 0.05. This threshold is slightly larger than the theoretical approximation 1.41.

Since a direct comparison of these regions in terms of size is conceptually complicated and technically demanding, we use the relation of confidence regions to hypothesis testing to compare them in terms of power. Under specific hypothetical, but incorrect, values of the change-points and mean values the power of the test of the true values represents the probability that the hypothetical values do not lie in the confidence region. Hence the procedure with larger power is preferred. From Table 8, it seems clear that for the parameter settings analyzed, the likelihood ratio procedure is preferable.

5. Exponential families. A natural generalization of the methods of this paper involve data from exponential families, where there usually is the option to pursue analogous methods or to use a normal approximation. We first develop the analogous theory and discuss the second possibility below.

Assume $X_1, \dots, X_m$ are independent and from a one-parameter exponential family of distributions $\{F_\theta : \theta \in \Theta\}$ where

$$\frac{dF_\theta}{du}(x) = \exp\bigl(\theta x - \psi(\theta)\bigr), \quad x \in \mathbb{R}, \theta \in \Theta,$$

u is a σ-finite measure on the real line and Θ is an open interval. For $0 \le i < j < k \le m$, the likelihood ratio statistic to test whether j is a change-point in the local background (i, k) is

$$\begin{aligned}\ell_{i,j,k} = {} & (j-i) \sup_{\theta_1 \in \Theta} \left(\theta_1 \frac{S_j - S_i}{j-i} - \psi(\theta_1) \right) + (k-j) \sup_{\theta_2 \in \Theta} \left(\theta_2 \frac{S_k - S_j}{k-j} - \psi(\theta_2) \right) \\ & - (k-i) \sup_{\theta \in \Theta} \left(\theta \frac{S_k - S_i}{k-i} - \psi(\theta) \right).\end{aligned}$$

In the following, we use $\mathbb{P}_\theta$ ($\mathbb{E}_\theta$, resp.) to denote the probability (expectation, resp.) calculated when $X_i \sim F_\theta, \forall i$. Following the proof of (2.1), we suggest the following approximation to the p-value of $\max_{i,j,k} \ell_{i,j,k}$:

$$\begin{aligned}&\mathbb{P}_\theta \Biggl(\max_{\substack{i<j<k \\ m_0 \le j-i, k-j \le m_1}} \ell_{i,j,k} \ge \frac{b^2}{2} \Biggr) \\ &\quad \sim \varphi(b) \sum_{\substack{m_0 \le n_1, n_2 \le m_1: \\ n_1+n_2 \le m}} (m - n_1 - n_2) \\ &\qquad \times \sum_{\theta_1, \theta_2} \frac{a(\theta_1, \theta) a(\theta_1, \theta_2) a(\theta, \theta_2)}{[n_1(\theta_1 - \theta)^2 \psi''(\theta_1) + n_2(\theta_2 - \theta)^2 \psi''(\theta_2)]^{1/2}},\end{aligned} \tag{5.1}$$

where the second summation is over two pairs of $\theta_1 < \theta_2$, which are assumed to exist (see the remark below), solving

$$\begin{cases} \psi'(\theta_1) n_1 + \psi'(\theta_2) n_2 = \psi'(\theta)(n_1 + n_2), \\ n_1 \bigl[\theta_1 \psi'(\theta_1) - \psi(\theta_1)\bigr] + n_2 \bigl[\theta_2 \psi'(\theta_2) - \psi(\theta_2)\bigr] \\ \quad - (n_1 + n_2) \bigl[\theta \psi'(\theta) - \psi(\theta)\bigr] = b^2/2, \end{cases} \tag{5.2}$$

TABLE 9
Exponential distribution with rate λ

λ	m	m_1	b	p_{Approx}	$\hat{p}$ (Monte Carlo)
1	500	50	4.72	0.049	0.061
1	500	100	4.78	0.048	0.053
1	1000	100	4.95	0.047	0.048

and for $\theta_1 < \theta_2$,

$$a(\theta_1, \theta_2) = \exp\left(-\sum_1^\infty n^{-1} \mathbb{E}_{\theta_2} e^{-[(\theta_2-\theta_1)S_n - n(\psi(\theta_2)-\psi(\theta_1))]^+} \right).$$

We use Theorem 8.51 of Siegmund (1985) and Theorem A of Tu and Siegmund (1999) to compute $a(\theta_1, \theta_2)$ numerically for nonarithmetic and arithmetic random variables, respectively.

REMARK. For those n_1 and n_2 such that the solutions to (5.2) do not exist, we first find the smallest $\theta' > \theta$ such that the solutions to (5.2) with θ replaced by θ' exist. We denote the solutions by θ_1' and θ_2'. Then the proposed approximation is the RHS(5.1) with θ, θ_1, θ_2 replaced by θ', θ_1', θ_2' respectively, and multiplied by $\mathbb{P}_\theta(S_{n_1+n_2}/(n_1+n_2) \geq \psi'(\theta'))$.

5.1. *Simulations*. We first consider the exponential distribution with rate λ. Observing that in (5.1), both the probability and its approximation do not depend on λ, we choose $\lambda = 1$ without loss of generality. We fix $m_0 = 1$. In Table 9, with different values of m, m_1 and b, p denotes the RHS(5.1) and $\hat{p}$ denotes the simulated p-value with 2000 repetitions. We see from Table 9 that our approximation to the p-values are reasonably accurate, especially when m and m_1 are large. From these results, it appears that use of a normal model would also be reasonable, especially for larger m_1 and m. For the three examples in Table 9, the approximation of Theorem 2.1 gives the probabilities 0.056, 0.053, 0.053, respectively.

Next, we consider the inverse Gaussian distribution with fixed shape parameter $\lambda = 10$. We fix $m_0 = 1$. With different values of the mean μ, m, m_1 and b, p denotes the RHS(5.1) and $\hat{p}$ denotes the simulated p-value with 2000 repetitions. We can see from Table 10 that both the theoretical and simulated p-values are reasonably robust against variations in the mean μ.

Since the computation of appropriate thresholds for nonnormal exponential families is somewhat complicated, one may also consider the use of normal approximations, which in some cases appears to work quite well. Following are two examples where a Gaussian approximation to the signed square root of the likelihood ratio statistic seems to perform admirably.

TABLE 10
Inverse Gaussian distribution with shape parameter $\lambda = 10$

μ	m	m_1	b	p_{Approx}	$\hat{p}$ (Monte Carlo)
1	500	50	4.72	0.051	0.043
5	500	50	4.72	0.036	0.027
1	500	100	4.78	0.049	0.037
5	500	100	4.78	0.035	0.031
1	1000	100	4.95	0.049	0.050
5	1000	100	4.95	0.036	0.034

For the detection of CG rich regions in genomic studies, as mentioned in the Introduction, the sequences are very long and the exact boundary between regions has little biological significance. Hence one often forms groups of consecutive Bernoulli variables. Following Elhaik, Graur and Josić (2010), we have used groups of 33 consecutive Bernoulli variables. Since the values of the Bernoulli parameters p are usually neither extremely small nor extremely large, possibilities that might indicate a Poisson approximation, we have tentatively assumed that we can use the theory developed above for the normal distribution. Since the Bernoulli variances depend on p, and hence must be estimated each time that parameter changes, it turns out that the skewness of the binomial distribution when p is not in the immediate neighborhood of 1/2 can make an approximation of the distribution of the scaled value of $[S_j - S_i - (j-i)(S_k - S_i)/(k-i)]$ by a normal distribution unsatisfactory, unless the size of the groups is relatively large. To avoid this problem, we have used the signed square roots of the log likelihood ratio statistics, which behave very much like a Gaussian process with a relatively stable variance. Simulations of this process indicate that the approximation is quite satisfactory and offer no new insights, so we omit the details.

The copy number data discussed in this paper was all obtained by comparative genomic hybridization. To achieve greater resolution, many present day studies use sequence data (e.g., Zhang et al. (2016)), which often utilize models built from Poisson processes. The simplest of these is concerned with detection of a change from a background rate for a Poisson process. Since the background rate varies with genomic position due to variation in sequencing depth, local detection procedures along the lines of LLR may be useful. Like the binomial distribution, to detect changes in the rate of a Poisson process, simulations support an approximation based on a normal approximation to the signed square root of the (generalized) log likelihood ratio statistic. For 500 observations, $b = 4.83$, and the mean of the Poisson distribution equal to 10, 400 simulations gave the significance value 0.0475, when our normal approximation gives the value 0.05. Calculation of Kullback–Leibler information suggests that for detecting changes from 10 to 20 and back to 10 in well separated intervals, interval lengths of 6 and 7 are borderline detectable. Several simulations of this case involving two pairs of change-points lead to successful detections of all four change-points, while the differences between the estimates of the change-points and the true values totaled 1–3 observations.

5.2. *Changes in a normal mean and variance.* An interesting, but considerably more complex example, is to allow for simultaneous changes to both the mean and variance (or mean vector and covariance matrix) of a sequence of independent, normally distributed observations. Although the formulation we have adopted, which assumes a constant value of the variance is much more common, and the copy number data considered above shows little evidence of heteroscedasticity within chromosomes, the recent paper of Du, Kao and Kou (2016), where the possibility of simultaneous changes in the mean and variance is considered, motivates the following brief discussion.

For $0 \le i < j \le m$, let $\ell_{i,j} = -0.5(j-i)\log(\sigma^2) - 0.5\sum_{i+1}^{j}(X_k - \mu)^2/\sigma^2$ denote the log likelihood of $X_{i+1}, \ldots, X_j$, and let $\hat{\ell}_{i,j} = -0.5(j-i)\log(\hat{\sigma}_{i,j}^2) - 0.5\sum_{i+1}^{j}(X_k - \bar{X}_{i,j})^2/\hat{\sigma}_{i,j}^2$ denote the log likelihood with parameters replaced by estimators. When the estimators are the maximum likelihood estimators, the generalized likelihood ratio statistic (which reduces to one-half the square of (1.2) in the case of known $\sigma^2 = 1$) is $\hat{\ell}_{i,j} + \hat{\ell}_{j,k} - \hat{\ell}_{i,k}$, maximized over $i < j < k$. Necessarily, we must take the minimum values of $j - i$ and $k - j$ at least equal to $m_0 = 2$. If one is interested in detecting changes occurring as close together as those studied above, this maximum likelihood ratio statistic is very unstable when there are no changes and $j - i$ or $k - j$ is small, since the maximum likelihood estimator of σ^2 can with substantial probability assume very small values. The consequence is that a suitable threshold to control

the rate of false positives must be so large that the statistic has very poor power to detect changes, and this problem persists even when m_0 is substantially larger than 2.

A device to ameliorate this problem that maintains the invariance of the likelihood ratio statistic under scale and location changes is to subtract a small constant $c/2$ from the sample size in the denominators of the estimators $\hat{\sigma}^2_{i,j}$ and $\hat{\sigma}^2_{j,k}$, and subtract c from the denominator of $\hat{\sigma}^2_{i,k}$. Then with these new estimators (denoted by a tilde) use the statistic $-(j-i-c/2)\log(\tilde{\sigma}^2_{i,j}) - (k-j-c/2)\log(\tilde{\sigma}^2_{j,k}) + (k-i-c)\log(\tilde{\sigma}^2_{i,k})$. In simulations, we have found that with $m_0 = 2$ and $c \approx 2.7$, this statistic has a false positive rate approximately the same as a two-dimensional version of (1.2), for which the significance level and power approximations of this paper are easily adapted. A similar result holds for the corresponding CBS statistic. If the variance changes by a factor of $1+\Delta$, the difference in mean values, scaled to unit standard deviation, is δ, and π denotes the fraction of observations at unit variance before a change-point, rough law of large numbers arguments indicate a noncentrality parameter in large samples proportional to

$$\pi(1-\pi)\log\bigl(1+\pi(1-\pi)\delta^2+(1-\pi)\Delta\bigr) - (1-\pi)\log(1+\Delta)$$

for the two-dimensional statistic.

If in fact there is no change in the variance the marginal power of the two-dimensional statistic to detect a change-point is approximately 0.2–0.3 less than the power of (1.2). When the variance does change, theoretical calculations and simulations suggest that there is a complex tradeoff that depends on the size of the changes in variance and the relative locations of the various change-points. Finally, there is also the issue that the likelihood ratio statistic that tests for a change in both mean and variance will not be as robust against excess kurtosis as a statistic that tests only for a change in mean value.

Following are the results of a few simulations that indicate the complexity of the problem. The statistics considered are the two-dimensional statistic suggested in this section, the statistic (1.2), and a modified version of (1.2), designed to compensate for the possibility that (1.2) has an excess of false positives. Since (1.2) estimates an average variance, if there is a subinterval where the variance is much larger than that average variance, the statistic (1.2) will use an inappropriately small variance estimate, which may lead to false positives. The modification of (1.2) is as follows: for any $i < k$, when searching for a putative change-point in $[i,k]$, standardize the process by the estimated (maximum likelihood) variance of the observations $X_i,\dots,X_k$. If there is a change-point in the interval, the maximum likelihood estimate may be positively biased, but other possibilities appear to be too unstable when the interval is short. Simulations indicate that the thresholds suggested by Theorem 2.1 are conservative.

In Table 11, the false positive in the sixth row is presumably a reflection of the fact that in the interval between 110 and 135 the variance of the observations is substantially larger than the "average variance" used by (1.2). Although we did not observe this in a number of other simulations, not reported here, this possibility of an inflated false positive error rate appears to be one of the principal disadvantages of using the unmodified (1.2), which otherwise seems to perform very well. The last five rows were based on the test case suggested by Du, Kao and Kou (2016) following an earlier suggestion of Lai et al. (2005), but we have reduced the signal to noise ratio to make more difficult what otherwise would be easy detections. In those last five rows, we see the effect on the two-dimensional statistic of the constant $c \approx 2.7$, which was introduced to reduce false positive errors in short test intervals, but here has an adverse effect on the power. For the modified version of (1.1), the behavior of which is similar to the two-dimensional statistic, the loss of power is presumably due to estimating the variance locally, which leads to large positive biases in (short) intervals containing change-points.

Since multiscale methods are designed to favor detection of change-points in longer over shorter intervals, it is natural to ask if imposition of a multiscale penalty on the square root

TABLE 11

Changes in mean and variance: $m = 200$, *Threshold for* (1.1) *and for the modification suggested above is* $b_1 = 4.54$; *threshold for the two-dimensional statistic is* $b_2 = 4.97$. *Detected change-points are as noted for* (1.2), *for the modification indicated in the text* (*denoted by an asterisk*), *and for the two-dimensional statistic suggested in this section*, *respectively. False positive errors are denoted by an asterisk*

τ	μ	σ	(1.1)	(1.1)*	2-D
38, 88, 108, 132	1.1, 2.7, 1.0, 2.5	1.1, 1.8, 1.1, 1.7	0, 88, 104, 132	0, 86, 0, 132	0, 88, 0, 133
38, 88, 108, 132	1.1, 2.7, 1.0, 2.5	1.1, 1.8, 1.1, 1.7	39, 86, 106, 135	39, 86, 106, 0	39, 88, 107, 0
38, 88, 108, 132	1.1, 2.7, 1.0, 2.5	1.1, 1.8, 1.1, 1.7	0, 0, 0, 132	37, 0, 0, 132	69, 82, 0, 0
38, 88, 108, 132	1.1, 2.7, 1.0, 2.5	1.1, 1.8, 1.1, 1.7	0, 90, 108, 134	36, 90, 0, 134	36, 84, 0, 134
30, 80, 110, 135	1.5, 0.5, 2.5, 1.0	1.5, 1.0, 2.0, 1.2	0, 0, 110, 134	0, 0, 110, 0	0, 0, 110, 136
30, 80, 110, 135	1.5, 0.5, 2.5, 1.0	1.5, 1.0, 2.0, 1.2	30, 74, 110, 115*, 132	30, 74, 110, 132	30, 77, 110, 133
30, 80, 110, 135	1.5, 0.5, 2.5, 1.0	1.5, 1.0, 2.0, 1.2	30, 0, 110, 134	30, 0, 110, 134	30, 0, 110, 134
48, 50, 150, 154	4.0, 0.0, 4.0, 0.0	2.0, 1.0, 2.0, 1.0	48, 50, 150, 154	0, 0, 150, 153	0, 0, 150, 155
48, 50, 150, 154	4.0, 0.0, 4.0, 0.0	2.0, 1.0, 2.0, 1.0	48, 50, 150, 154	48, 50, 150, 154	48, 50, 150, 155
48, 50, 150, 154	4.0, 0.0, 4.0, 0.0	2.0, 1.0, 2.0, 1.0	0, 0, 150, 155	0, 0, 150, 155	0, 0, 150, 154
48, 50, 150, 154	4.0, 0.0, 4.0, 0.0	2.0, 1.0, 2.0, 1.0	48, 50, 149, 154	48, 50, 0, 0	48, 50, 0, 0
48, 50, 150, 154	5.0, 0.0, 5.0, 0.0	2.0, 1.0, 2.0, 1.0	48, 50, 150, 154	48, 50, 150, 154	0, 0, 150, 155

of the likelihood ratio statistic would work here. Some numerical experimentation suggests that the penalty $[4\log(3m/\min(j-i,k-j))]^{1/2}$ allows one to control the false positive rate, and the multiscale statistic performs about as well in these examples as the two-dimensional statistic defined above.

6. Discussion. We have studied local thresholding procedures for segmenting sequences of independent random variables subject to change-points in the mean. The local likelihood ratio statistic, LLR, begins with subsets of intervals $(i,k]$, computes the maximum log likelihood ratio statistic, $\max_{i<j<k}|Z_{i,j,k}|$, which is compared to a threshold designed to control the probability of a false positive error, and values of j are selected by enforcing a no overlap condition described above. The method is designed to provide a set of detected change-points j and background intervals $(i,k]$, each of which contains only a single putative change-point. The pseudo-sequential procedure SLLR leaves i fixed at 0 or at the most recently discovered candidate change-point, then sequentially with respect to k examines $\max_{i<j<k} Z_{i,j,k}$ until it exceeds a suitable threshold, which suggests a change-point at an appropriate value of j. The statistic LLR has better false positive control than SLLR, although it requires a relatively large threshold, and hence loses some power compared to SLLR, especially when the number of change-points is large.

Our suggested procedures are compared to several other threshold based procedures that attempt to control, with varying degrees of success, the false positive error rate: (i) the Wild Binary Segmentation (WBS) procedure of Fryzlewicz (2014), (ii) the SaRa procedure of Niu and Zhang (2012), (iii) the CBS procedure of Olshen et al. (2004) and (iv) Multi, a related iterative threshold based implementation of the statistic of Frick, Munk and Sieling (2014). Each of these methods has strengths and weaknesses, some obvious, others not so obvious. When the thresholds are chosen so that the probability of exceeding the threshold somewhere when there are no change-points is fixed, say 0.05, the procedures WBS, SLLR, CBS and Multi have the best power of detection. CBS and Multi have less adequate control over the false positive rate, especially when the number of change-points is large, so a substantial number of iterations of the basic search is required. The statistics LLR and SaRa provide strict asymptotic control of the false positive error rate, but LLR has less power than the others, while SaRa suffers a severe loss of power when change-points are close together. CBS and Multi show expected power advantages/disadvantages, with CBS performing better in detecting near-by change-points of large amplitude and Multi performing better in detecting distant change-points of small amplitude. If the mean moves in the same direction at both of relatively nearby change-points, CBS and Multi may pick an intermediate value and fail to detect the second change-point. Our thresholding implementation of Multi is based on the approximation (2.9) and omits the dynamic programming step from the algorithm suggested in Frick, Munk and Sieling (2014). We (and others) found that algorithm to perform poorly when used with default parameters; but our analysis shows that when it is calibrated to have a global false positive rate comparable to the others, it performs competitively.

When the number of change-points is large, the statistics CBS and Multi experience an increase in false positives, while LLR experiences a decrease. This reflects the difference between top-down and bottom-up approaches. The method of proof of Theorem 2.1 shows that a bottom-up versions of CBS (also Multi) can be formulated and would have guaranteed (asymptotic) false positive control, although it would require still higher thresholds than LLR, run more slowly, and suffer a similar loss of power with an increasing number of change-points. After some preliminary experimentation that suggests the disadvantages outweigh, or at least neutralize the potential advantages, we have not pursued this possibility.

If one's goals are somewhat exploratory and do not depend on providing convincing evidence of the existence of particular change-points, a visual and/or rough preliminary analysis

that provides some idea of the number and configuration of change-points may be helpful in choosing a detection threshold that brings the false positive and false negative rates into balance.

Inversion of the log likelihood ratio statistic is used to obtain approximate confidence regions for the locations of the change-points or jointly for the locations and amplitudes of the changes. For the latter case, Frick, Munk and Sieling (2014) suggested a quite different method, which amounts to testing whether there is a change in the hypothesized mean values between any two hypothesized change-points. Numerical examples suggest that for change-points of large amplitude our methods provide more accurate estimates, although our asymptotic control of the confidence level deteriorates if the sizes of the changes or the distances between consecutive change-points are not sufficiently large.

We have conducted a small simulation experiment to study the problem noted in the literature on detection of copy number variations, that there may be local drifts that can give the appearance of change-points where there is none. For these problems, all methods appear to suffer some loss of power, but the methods LLR, SLLR, WBS and SaRa which use a local background seem less likely to experience an increase in false positives than the top down methods CBS and Multi.

We have provided a brief discussion of detection of simultaneous changes in a normal mean and variance, but our analysis to date suggests that the problem is quite complicated and requires additional study.

We have assumed the observations are independent, which appears to be the case for the problems motivating our study, both from the nature of the experiments and from the data themselves. This provides a considerable advantage in estimating the variance, as discussed above. For weakly dependent data, there are roughly two different alternatives, which we are now studying. For short range dependence, if the distance between change-points is proportional to the number of observations, a number of authors (e.g., Robbins, Gallagher and Lund (2016)) have observed that weak convergence arguments can be used to obtain mathematical results similar to those given above, expressed in terms of Brownian motion. Details involve correcting the (estimate of the) variance for autocorrelation of the individual observations. A second approach is to use a low-order autoregressive model, which allows one to pursue an asymptotic likelihood analysis, similar to what we have followed in this paper. Our preliminary studies suggest that both approaches are successful under certain conditions in controlling the false positive rate, but lead to a substantial loss of power, because of biased estimates of the variance and autocorrelation when there are change-points. We expect to provide a thorough discussion of these problems in the future.

Another challenging problem is to detect and estimate local signals in spatial data (with or without a temporal variable). For independent observations on a rectangular grid, our methods generalize easily to signals having a rectangular shape with sides parallel to the sides of the grid. A natural question is the extent to which this is adequate for detection of the signals having the many irregular shapes possible in higher dimensions. Another approach would be to smooth the signals, which opens up the possibility of dealing with a much larger set of shapes.

Acknowledgment. The authors thank Nancy Zhang for several helpful discussions and suggestions. We also thank the anonymous referees for their detailed comments which led to many improvements.

The first author was supported in part by NUS grant R-155-000-158-112, CUHK direct grant 4053234 and a CUHK start-up grant.

The third author was supported in part by the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Segmentation and estimation of change-point models: False positive control and confidence regions." (DOI: 10.1214/19-AOS1861SUPP; .pdf). This supplement contains proofs of Theorems 2.1 and 3.1.

REFERENCES

ASTON, J. A. D. and KIRCH, C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.* **6** 1906–1948. MR3058688 https://doi.org/10.1214/12-AOAS565

BARANOWSKI, R., CHEN, Y. and FRYZLEWICZ, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 649–672. MR3961502

CHAN, H. P. and CHEN, H. (2017). Multi-sequence segmentation via score and higher-criticism tests. Available at arXiv:1706.07586v1.

CHURCHILL, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51** 79–94. MR0978904 https://doi.org/10.1016/S0092-8240(89)80049-7

DU, C., KAO, C.-L. M. and KOU, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.* **111** 314–330. MR3494662 https://doi.org/10.1080/01621459.2015.1006365

DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* **29** 124–152. MR1833961 https://doi.org/10.1214/aos/996986504

ELHAIK, E., GRAUR, D. and JOSIĆ, K. (2010). Comparative testing of DNA segmentation algorithms using benchmark simulations. *Mol. Biol. Evol.* **27** 1015–1024.

FANG, X., LI, J. and SIEGMUND, D. (2020). Supplement to "Segmentation and estimation of change-point models: False positive control and confidence regions." https://doi.org/10.1214/19-AOS1861SUPP.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

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

LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R. and PARK, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21** 3763–3770.

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 https://doi.org/10.1214/12-AOAS539

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. *Biostatistics* **5** 557–572.

PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. and DAUDIN, J. J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinform.* **6** 27.

POLLACK, J. R., PEROU, C. M., ALIZADEH, A. A., EISEN, M. B., PERGAMENSCHIKOV, A., WILLIAMS, C. F., JEFFREY, S. S., BOTSTEIN, D. and BROWN, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23** 41–46.

POLLACK, J. R., SØRLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BØRRESEN-DALE, A. L. et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99** 12963–12968.

ROBBINS, M. W., GALLAGHER, C. M. and LUND, R. B. (2016). A general regression changepoint test for time series data. *J. Amer. Statist. Assoc.* **111** 670–683. MR3538696 https://doi.org/10.1080/01621459.2015.1029130

SCHWARTZMAN, A., GAVRILOV, Y. and ADLER, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *Ann. Statist.* **39** 3290–3319. MR3012409 https://doi.org/10.1214/11-AOS943

SIEGMUND, D. (1985). *Sequential Analysis*: *Tests and Confidence Intervals*. *Springer Series in Statistics*. Springer, New York. MR0799155 https://doi.org/10.1007/978-1-4757-1862-1

SIEGMUND, D. (1988a). Approximate tail probabilities for the maxima of some random fields. *Ann. Probab.* **16** 487–501. MR0929059

SIEGMUND, D. (1988b). Confidence sets in change-point problems. *Int. Stat. Rev.* **56** 31–48. MR0963139 https://doi.org/10.2307/1403360

SIEGMUND, D. and YAKIR, B. (2000). Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6** 191–213. MR1748719 https://doi.org/10.2307/3318574

SIEGMUND, D. and YAKIR, B. (2007). *The Statistics of Gene Mapping*. *Statistics for Biology and Health*. Springer, New York. MR2301277

SIEGMUND, D. O., ZHANG, N. R. and YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337 https://doi.org/10.1093/biomet/asr057

SNIJDERS, A. M., FRIDLYAND, J., MANS, D. A., SEGRAVES, R., JAIN, A. N., PINKEL, D. and ALBERTSONN, D. G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22** 4370–4379.

TU, I. and SIEGMUND, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis. *Adv. in Appl. Probab.* **31** 510–531. MR1724565 https://doi.org/10.1239/aap/1029955145

VOSTRIKOVA, L. (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 https://doi.org/10.1093/biomet/73.1.91

YAKIR, B. (2013). *Extremes in Random Fields*: *A Theory and Its Applications*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3241226 https://doi.org/10.1002/9781118720608

ZHANG, Y. and LIU, J. S. (2011). Fast and accurate approximation to significance tests in genome-wide association studies. *J. Amer. Statist. Assoc.* **106** 846–857. MR2894742 https://doi.org/10.1198/jasa.2011.ap10657

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 https://doi.org/10.1111/j.1541-0420.2006.00662.x

ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. With supplementary data available online. *Biometrika* **97** 631–645. MR2672488 https://doi.org/10.1093/biomet/asq025

ZHANG, N. R., YAKIR, B., XIA, L. C. and SIEGMUND, D. (2016). Scan statistics on Poisson random fields with applications in genomics. *Ann. Appl. Stat.* **10** 726–755. MR3528358 https://doi.org/10.1214/15-AOAS892

ZHAO, X., LI, C., PAEZ, J. G., CHIN, K., JÄNNE, P. A., CHEN, T.-H., GIRARD, L., MINNA, J., CHRISTIANI, D. et al. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64** 3060–3071.