9-2010

Detecting Simultaneous Change-Points in Multiple Sequences

Nancy Zhang
University of Pennsylvania

David O. Siegmund
Stanford University

Hanlee P. Ji
Stanford University

Jun Li
University of Michigan

Recommended Citation

Zhang, N., Siegmund, D. O., Ji, H. P., & Li, J. (2010). Detecting Simultaneous Change-Points in Multiple Sequences. *Biometrika, 97* (3), 631-645. http://dx.doi.org/10.1093/biomet/asq025

Detecting Simultaneous Change-Points in Multiple Sequences

Abstract

We discuss the problem of detecting local signals that occur at the same location in multiple one dimensional noisy sequences, with particular attention to relatively weak signals that may occur in only a fraction of the sequences. We propose statistics that combine data across sequences and show that they have better power properties and provide a more easily interpreted summary of the data than do procedures based on a separate analysis for each sequence. In particular, we examine the case where the signal is a temporary shift in the mean of independent Gaussian observations. The formulation of the model is motivated by the problem of detecting recurrent DNA copy number variants in multiple samples, and our results are illustrated by applications to data involving DNA copy number changes.

Keywords

boundary crossing, changepoint detection, DNA copy number, meta-analysis, scan statistic, segementation

Disciplines

Biostatistics | Statistics and Probability

Detecting Simultaneous Change-points in Multiple Sequences

Nancy R. Zhang

Department of Statistics

Stanford University

Stanford, CA 94305-4065, USA

nzhang@stat.stanford.edu

David O. Siegmund

Department of Statistics

Stanford University

Stanford, CA 94305-4065, USA

dos@stat.stanford.edu

Hanlee Ji

Department of Medicine, Division of Oncology

Stanford University School of Medicine

Stanford, CA 94305, USA

genomics_ji@stanford.edu

Jun Li

Department of Human Genetics

University of Michigan

Ann Arbor, MI 48109-0618, USA

junzli@umich.edu

SUMMARY

We discuss the problem of detecting local signals that occur at the same location in multiple one dimensional noisy sequences, with particular attention to relatively weak signals that may occur in only a fraction of the sequences. We propose statistics that combine data across sequences and show that they have better power properties and provide a more easily interpreted summary of the data than do procedures based on a separate analysis for each sequence. In particular, we examine the case where the signal is a temporary shift in the mean of independent Gaussian observations. The formulation of the model is motivated by the problem of detecting recurrent DNA copy number variants in multiple samples, and our results are illustrated by applications to data involving DNA copy number changes.

Key words and phrases: Scan statistics, change-point detection, segmentation, meta-analysis, DNA copy number

1 Introduction

We study in this paper the statistical problem of detecting local signals that occur at the same location in multiple noisy sequences. Of particular importance are cases where the shared signal occurs in only a fraction of the sequences. This inquiry is motivated by current problems in biology, where high-throughput genomic profiles are collected for cohorts of biological samples, and it may be of interest to pool data across samples to boost power for detecting simultaneously occurring signals. We start by describing a few motivating applications.

1. **Detection of DNA copy number variation:** DNA copy number variants (CNVs), i.e. gains and losses of chromosomal segments, are an important class of genetic variation. Various laboratory techniques have been developed for measuring the quantity of DNA present in a population of cells, relative to the expected quantity of two copies (for autosomal chromosomes). These measurements are taken at a set of probes, each mapping to a specific location in the genome. The data thus produced are a set of linear profiles, one for each biological sample in the study.

   While there are many published methods for CNV detection, most deal with samples one at a time (Fridlyand et al., 2004; Olshen et al., 2004; Wang et al., 2005; Picard et al., 2005; Hsu et al., 2005; Guha et al., 2006; Engler et al., 2006; Wen et al., 2006; Broët & Richardson, 2006; Lai et al., 2007; Tibshirani & Wang, 2008). Two independent comparative reviews (Lai et al., 2005; Willenbrock & Fridlyand, 2005) of single sample methods concluded that the published methods often disagree, and both reviews concluded that the circular binary segmentation

(CBS) method of Olshen et al. (2004); Venkatraman & Olshen (2007) performs well.

Since CNVs are often shared across individuals, we would like to scan all profiles simultaneously to detect shared CNVs and to obtain a sparse multi-sample summary that can serve as the overall molecular signature for the cohort of samples. In most of the literature, cross sample analyses are done post-segmentation (e.g. Diskin et al. (2006), Wang et al. (2008), and Newton et al. (1998),Newton & Lee (2000)). An important exception where data is pooled across samples during the segmentation step is Shah et al. (2007), who proposed hierarchical hidden Markov models that make specific assumptions about the occurrences, durations and amplitudes of variant intervals. More background on this problem is given in Section 4.

2. **Transcription profiling using tiling arrays:** High-density genomic tiling microarrays cover a complete genome with densely tiled oligonucleotide probes. These arrays can be used to assay in an unbiased manner multiple types of activity on the genome, including transcription, DNA-protein-binding, and chromatin modification (references needed). As for CNV detection, tiling array data are often collected for multiple samples in one study. It is also frequently of interest to detect common regions of activity, and to pool data across samples to locate weak signals (Piccolboni, 2008; Huber et al., 2006).

3. **Meta-analysis of multiple linkage studies:** Whole genome linkage studies seek to identify genetic regions that may contain susceptibility genes for diseases or genes that contribute to other traits of interest. Often, several linkage studies

with modest sample sizes are reported, with differing results for the same genomic region. This is not surprising, since the power of detection by individual studies is often modest. Wise et al. (1999) and Badner & Gershon (2002) proposed statistical criterion for the simultaneous analysis of multiple genome scans.

The scenarios described above involve situations where a simultaneous scan for a shared signal across multiple linear profiles can potentially improve robustness and power by borrowing strength across profiles. Within individual profiles, the signal of interest, as well as the noise structure, may vary across applications. In this paper, we look at the specific problem of detecting an abrupt shift in mean when the noise within each profile is assumed to be independent and identically distributed Gaussian. The mean shift model can be directly applied to the problem of CNV detection described in Example 1. With modifications for correlated errors and probe-level effects, the methods can potentially also apply to transcription profiling using tiling arrays. The meta-analysis of multiple linkage studies can be viewed in similar light, but would need to acount for the diversity of study designs. All of these applications have their own set of idiosyncracies that must be factored into the models, but we hope to convey some common themes that extend across applications.

Motivated by the comparative evaluation reported in Lai et al. (2005) and Willenbrock & Fridlyand (2005), the relative simplicity of the CBS algorithm, and our past experience in using that algorithm to provide input for our BIC related model selection criterion (Zhang and Siegmund, 2006), we adopt the conceptual foundations of Olshen et al. (2004). In this paper we focus on the issue of borrowing information across multiple scans.

In Section 2, we build on existing change-point methods to formulate some simple test statistics, and provide approximations to their significance level and power. In Section 3, we evaluate the approximations and study the power of these statistics using numerical calculations and simulations. In Section 4, we provide a more extensive review of the scientific issues involved with studies of copy number variation and illustrate our methods on two data sets. After summarizing some general conclusions in Section 5, we sketch in Section 6 the steps in deriving the approximations stated in Section 2.

2 Methods

2.1 Problem Formulation

Let the observed data be a two dimensional array $\{y_{it} : \quad 1 \leq i \leq N, \quad 1 \leq t \leq T\}$, where y_{it} is the data point for the i-th profile at location t, N is the total number of profiles, and T is the total number of locations. We assume that for each i, the random variables $\boldsymbol{y}_i = \{y_{it} : \quad 1 \leq t \leq T\}$ are mutually independent and normally distributed with mean values μ_{it} and variances σ_i^2. We further assume that under the null hypothesis, for each profile i, the random variables y_{it} are identically distributed with "baseline" mean value μ_i. The alternative hypothesis of interest is that there exist values $1 \leq \tau_1 < \tau_2 \leq T$ and a set of profiles $\mathcal{J} \subset \{1, \ldots, N\}$, such that for $i \in \mathcal{J}$,

$$\mu_{it} = \mu_{i0} + \delta_i I_{\{\tau_1 < t \leq \tau_2\}}, \tag{2.1}$$

where the δ_i are non-zero constants and μ_{i0} is the background mean level for sample i. Under the alternative hypothesis we refer to $(\tau_1, \tau_2]$ as a variant interval and $\mathcal{J}$ the set of carriers associated with the interval. If the alternative hypothesis is true, we are interested primarily in detecting this situation and in estimating the variant interval, and secondarily in determining the carriers.

This model is motivated by the analysis of DNA copy number data, for which we provide more background in Section 4. In that application, each profile is usually a different biological sample, with the data points mapping to locations along chromosomes. The change-points τ_1, τ_2 demarcate CNVs. Empirical evidence suggest that the baseline means and sample variances differ substantially across samples. We also found that the shifts in mean differ across the carriers for a given CNV. For example, Figure 1a shows the sample means $\bar{y}_{i,\tau_1:\tau_2} = (y_{i,\tau_1+1} + \cdots + y_{i,\tau_2})/(\tau_2 - \tau_1)$ within two known copy number variant intervals for a set of 62 samples described in Section 4.3. The triangles mark the sample means within the CNV for validated carriers. Observe that the locations of the triangles vary over a wide range. This motivates the allocation of a separate δ_i parameter for each carrier at any given CNV.

In many applications, there are usually multiple variant intervals defined by different τ_1 and τ_2, and $\mathcal{J}$. In DNA copy number data, the magnitude of change differs widely across CNVs for any given sample. Figure 1b presents empirical evidence for this fact: Two samples are shown. For each sample, a histogram of $\{y_{it} : t = 1, \ldots, T\}$ is plotted. The triangles mark the location of validated CNVs in that sample. Observe that the locations of the triangles vary substantially. This motivates the estimation of a separate $\delta_i(\tau_1, \tau_2)$ for each interval τ_1, τ_2. We describe our test statistics first for the simple case where there is at most one variant interval. Then, we build on these test statistics to obtain segmentation algorithms similar to CBS (Olshen et al., 2004) for cases where multiple variant intervals can occur.

2.2 The Sum-of-Chisquares Statistic

First, we recall existing methods for the analysis of only one profile, where temporarily we suppress the dependence of our notation on the profile indicator i. For the data sequence $\{y_1, \ldots, y_T\}$, let $S_t = y_1 + \ldots + y_t$, $\bar{y}_t = S_t/t$, and $\hat{\sigma}^2 = T^{-1} \sum_1^T (y_t - \bar{y}_T)^2$. The test statistic used in Olshen et al. (2004) and Zhang & Siegmund (2007) is

$$\max_{s,t} U^2(s,t), \tag{2.2}$$

where

$$U(s,t) = \hat{\sigma}^{-1}\{S_t - S_s - (t-s)\bar{y}_T\}/[(t-s)\{1-(t-s)/T\}]^{1/2}, \tag{2.3}$$

and the max is taken over $1 \leq s < t \leq T, \;\; t - s \leq T_0$. Here $T_0 < T$ is an assumed upper bound on the length of the variant interval, which in some contexts may be much smaller than T.

If the error standard deviation σ were known and could be used in the definition of $U(s,t)$, (2.2) would be the likelihood ratio statistic. In practice σ must be estimated. Since T is relatively large in typical applications, we shall for theoretical developments treat σ as known. Then, we can without loss of generality set $\sigma = 1$. Numerical studies suggest that this is a reasonable simplification.

Now consider the model (2.1) for the original problem involving N sequences. To test the null hypothesis that $\delta_i = 0$ for all $1 \leq i \leq N$ versus the alternative that for some values of $\tau_1 < \tau_2$ at least some δ_i are not zero, a direct generalization of (2.2) is

$$\max_{s<t} \sum_{i=1}^{N} U_i^2(s,t), \tag{2.4}$$

where $U_i(s,t)$ is the sequence specific statistic defined as in (2.3) for the ith sequence. As in the single profile case, if the variances were known, (2.4) would be the generalized log likelihood ratio statistic. For each fixed $s < t$, the null distribution of the indicated sum is approximately χ^2 with N degrees of freedom. Since N can potentially be large, it will be convenient to consider the standardized statistics

$$Z(s,t) = \sum_{i=1}^{N} \{U_i(s,t)^2 - 1\}/(2N)^{1/2}. \tag{2.5}$$

We will refer to (2.5) as the sum of χ^2 statistic. Large values of $Z(s,t)$ are evidence against the null hypothesis. If the null hypothesis is rejected, the maximum likelihood estimate of the location of the variant interval is $(s^*, t^*) = \arg\max_{s,t} Z(s,t)$.

Before turning to approximations for the significance level and power of a multi-sample scan using $Z(s,t)$, we consider a weighted version of this statistic suggested by a mixture model.

2.3 The Weighted Sum-of-Chisquares Statistic

Conducting a separate analysis for each individual sequence requires that each sample show strong evidence for the detection of a variant interval. The sum of chi-squares statistic goes to the other extreme of favoring situations where many samples have relatively weak evidence. This is because for assessing whether location $[s,t]$ contains a variant interval in some of the sequences, all individuals contribute a "vote" in terms of adding a chi-square to the sum. However, in most cases, the carriers of a variant interval make up only a (small) fraction of the samples. Further, in the case of DNA copy number data, low-amplitude fluctuations shared by many samples are more likely to be due to measurement artifacts rather than biologically interesting signals. We propose an intermediate statistic that requires individual sequences to show strong enough evidence for a variant interval in order to have substantial vote in the pooled scan. In this section, we examine a mixture model that naturally gives rise to such a weighted statistic.

Let $Q_i(s,t)$ denote the indicator that $i \in \mathcal{J}$, i.e. that sample i is a carrier of the aligned change segment at s,t. If $Q_i(s,t)$ were observable, the log likelihood ratio statistic would be

$$\max_{s,t} \sum_{i=1}^{N} \log[\{1 - Q_i(s,t)\} + Q_i(s,t) e^{U_i^2(s,t)/2}] = \max_{s,t} \sum_{i=1}^{N} Q_i(s,t) U_i^2(s,t)/2. \tag{2.6}$$

In the above statistic, a sample contributes to the sum if and only if it is a carrier. Since $Q_i(s,t)$ is not observed, we propose to estimate it as follows. Let $p \in [0,1)$ be a pre-specified prior probability that $Q_i(s,t) = 1$. We consider the posterior distribution of $Q_i(s,t)$, given the data. After maximizing the posterior mean of $Q_i(s,t)$ with respect to the unknown parameters δ, μ, we get estimates

$$\hat{Q}_i(s,t) = \max_{\delta,\mu} E[Q_i(s,t)|y] = \exp\{U_i^2(s,t)/2\}/[r_p + \exp\{U_i^2(s,t)/2\}], \tag{2.7}$$

where $r_p = (1-p)/p$ denotes the prior odds of $Q_i(s,t) = 0$ versus $Q_i(s,t) = 1$. Substituting $Q_i(s,t)$ by its estimated value in (2.6) leads to our weighted sum of chi-squares statistic:

$$\max_{s,t} \sum_{i=1}^{N} w_p[U_i(s,t)]U_i^2(s,t), \tag{2.8}$$

where

$$w_p(x) = \exp(x^2/2)/\{r_p + \exp(x^2/2)\}. \tag{2.9}$$

Note that small values of p require a more substantial apparent signal from a given sample before that sample is allowed to make an important contribution to the overall statistic. For $p = 1$, we obtain the unweighted sum of chi-squares statistic (2.4).

Assuming that under the null hypothesis $U_i^2(s,t)$ has exactly a chi-square distribution, we can easily compute through numerical integration the expectation $\mu_p = \mathrm{E}[w_p(U)U^2]$ and variance $\sigma_p^2 = \mathrm{var}(w_p(U)U^2]$ of the summands in (2.8), and use them to standardize (2.8) to obtain

$$Z^{(p)}(s,t) = \left[\sum_{i=1}^{N} w_p\{U_i(s,t)\}U_i^2(s,t) - N\mu_p\right] \Big/ \sigma_p N^{1/2}. \tag{2.10}$$

This leads us to the weighted scan statistic $\max_{s<t} Z^{(p)}(s,t)$. In Section 2.5 we will show via numerical studies that the weighted sum of chi-square statistic has higher power than the unweighted statistic when only a small subset of all profiles carry the variant interval.

2.4 Approximations for the Significance Level

In this section, we describe analytic approximations to the significance level for scan statistics of the form (2.8). These approximations will be evaluated numerically in Section 3, and proved in Section 6.

Let $\psi(\theta)$ be the log moment generating function of the standardized weighted chi-square distribution:

$$\psi(\theta) = \log E(\exp[\theta\{f(U) - \mu_p\}/\sigma_p]), \tag{2.11}$$

where $f(U) = w_p(U)U^2$, $U \sim$ N(0,1), and $w_p(\cdot)$ is the weight function defined in (2.9). Let $\theta = \theta_{b,N}$ be chosen to satisfy $\dot{\psi}(\theta) = b/N^{1/2}$. Let $\mathcal{I} = N\{\theta\dot{\psi}(\theta) - \psi(\theta)\}$. Define

$$\beta = (2\sigma^2)^{-1}[E\{f(U)f'(U)U\} - E\{f(U)f''(U)\}].$$

All of these quantities can be computed numerically for any given function f.

Then, an approximation to the significance level of the statistic $Z_{\max} = \max_{\substack{0<s<t<T \\ c_1T<t-s<c_2T}} Z_{s,t}^{(p)}$ is

$$\begin{aligned} \mathrm{pr}\,(Z_{\max} > b) \;\; \approx \;\; & [2\pi\ddot{\psi}(\theta)]^{-1/2}e^{-\mathcal{I}}b^3\beta^2 \\ & \int_{c_1}^{c_2} \frac{1}{u^2(1-u)}\nu^2\left[\frac{b(2\beta/T)^{1/2}}{\{u(1-u)\}^{1/2}}\right] du, \end{aligned} \tag{2.12}$$

where the function $\nu(x)$ is defined in Siegmund (1985, p. 85). A simple approximation for numerical calculations is $\nu(x) \approx [(2/x)\{\Phi(x/2) - 1/2\}]/\{(x/2)\Phi(x/2) + \varphi(x/2)\}$, where φ and Φ are the standard normal density and distribution function, respectively.

In (2.12), $[2\pi\ddot{\psi}(\theta)]^{-1/2}e^{-\mathcal{I}}$ arises from an approximation of the marginal probability

$\mathrm{pr}(Z_{s,t}^{(p)} > b)$ for the specific window (s,t). The rest of (2.12) is a multiple testing correction for taking the maximum over all possible windows. The marginal term is based on large deviation techniques, while the multiple testing correction is derived by approximating the local increments of $\{Z_{s,t}^{(p)}\}$ by a Gaussian process. For a numerical example, consider the unweighted case, where $f(U) = U^2$. Take $T = 1000$, $T_0 = 100$ and $N = 200$. The approximation (2.12) gives a 0.05 significance threshold threshold of $b \approx 5.09$. A 1000 repetition simulation experiment gives as p-value for this threshold the value 0.047. In Section 3 we report more extensive simulations.

While (2.12) seems reasonable as a rough approximation, its heuristic derivation involving a combination of Gaussian and non-Gaussian calculations cannot be made mathematically rigorous. We propose a second approximation for the sum of chi-squares statistic, which is theoretically more satisfactory and much more accurate in situations where N is small. Since this alternative approximation relies on the radial symmetry and other special properties of the sum of chi-squares statistic, it can not be applied to the weighted chi-squares. The approximation given in Siegmund (1988) for $N = 1$ can be directly generalized to arbitrary N, but the generalization is overly conservative, since it in effect approximates spheres locally by their tangent planes. To describe the improved approximation, let $\tilde{Z}(s,t) = \{\sum_1^N U_i^2(s,t)\}^{1/2}$ denote the usual Euclidean norm of the vector $(U_1(s,t), \ldots, U_N(s,t))'$. Then

$$\mathrm{pr}\left(\max_{\substack{0<s<t<T \\ c_1 T < t-s < c_2 T}} \tilde{Z}_{s,t} > b\right) \approx .5b^4(1 - \frac{N-1}{b^2})^3 f_N(b^2) \int_{c_1}^{c_2} \frac{1}{u^2(1-u)} \nu^2 \left[\frac{b\{1-(N-1)/b^2\}}{\{Tu(1-u)\}^{1/2}}\right] du, \tag{2.13}$$

where f_N is the chi-square density with N degrees of freedom. See Section 6.2 for a derivation of this approximation.

2.5 Power

Assuming that there is a variant interval at $(\tau_1, \tau_2]$, we now consider the power of the unweighted and weighted chisquares statistics in detecting this interval. As an approximation to the power, we consider the probability

$$\mathrm{pr}\{Z^{(p)}(\tau_1, \tau_2) > b\},$$

where $b = b_p$ is the threshold chosen to achieve a pre-chosen significance level, say 0.05. This probability is a lower bound on the true power, which also involves the (relatively small) probability that $Z^{(p)}(s,t) < b$ for $(s,t] = (\tau_1, \tau_2]$, but exceeds b for nearby s, t. By the central limit theorem, we approximate this probability by regarding $Z^{(p)}$ as normally distributed with mean and variance that can be numerically computed for a given alternative distribution. For example, consider the sum of chi-squares statistic ($p = 1$), where explicit analytic formulas can be easily obtained: The expectation of $Z(\tau_1, \tau_2)$ is $(\tau_2 - \tau_1)\sum_{i\in\mathcal{J}}(\delta_i/\sigma_i)^2/(2N)^{1/2}$, and the variance is $1 + 2(\tau_2 - \tau_1)\sum_{i\in\mathcal{J}}(\delta_i/\sigma_i)^2$. To simplify the numerical examples to follow, we assume that $\delta_i/\sigma_i = \Delta$ for all $i \in \mathcal{J}$. Also let π denote the cardinality of $\mathcal{J}$ divided by N, i.e., the proportion of sequences having true variant intervals at $(\tau_1, \tau_2]$. The expectation of $Z(\tau_1, \tau_2)$ then becomes

$$N^{1/2}\pi(\tau_2 - \tau_1)\Delta^2/2^{1/2},$$

which depends in a simple way on the the proportion of carriers, the length of the variant interval, and the magnitude of change. For the weighted sum of chi-squares statistic the expectation will again be directly proportional to the product of $N^{1/2}$, π, and a function of ξ, although now this function must be computed numerically.

For a simple numerical illustration, consider the example discussed in Section 2.4, for which 0.05 level significance thresholds are $b = 5.09$ and $b = 7.8$ respectively for the unweighted and weighted ($r_p = 100$) statistics. Suppose $\xi = 3$ and $\pi = 0.07$. The approximate power for the unweighted statistic is 0.79, and increases to 0.94 for the weighted statistic. For a smaller effect size and a larger proportion of carriers, the relation can be reversed. For $\xi = 2$ and $\pi = 0.15$, the approximate power for the unweighted statistic is 0.73, and decreases to 0.63 for the weighted statistic. More extensive numerical comparisons are given in Section 3.

2.6 Search Algorithm for Multiple Variant Intervals

In general the data contains several, possibly nested, variant intervals. We now describe algorithms for simultaneously segmenting multiple sequences. In design of the algorithms, we found it useful to distinguish between two scenarios: In the first scenario, the variant intervals are short and reasonably well separated. For example, in the analysis of DNA copy number data collected from normal tissue samples, the copy number variants usually involve changes of small magnitude over short segments that are well separated along the genome. In this case simultaneous detection of all variant intervals can be achieved by a straightforward scanning procedure, as implemented in Algorithm 2.1 below.

In the second scenario, the variant intervals comprise a substantial portion of the sequences being analyzed, and changes may be overlapping or nested. An example is DNA copy number data collected from cancer samples, where somatic aberrations often span entire chromosomes or chromosome arms. In these cases the more complex Algorithm 2.2, which involves a recursion, works better. Algorithm 2.2 resembles the iterative CBS procedure proposed by Olshen *et al.* (2004) for segmentation of a single sequence. For multiple sequences it requires that in the course of the recursion we identify which sequences are carriers of the variant intervals, for which we discuss possible solutions below.

Algorithm 2.1. *Fix a global significance level* α*, a parameter* $p \in (0,1]$ *(or equivalently* r_p*), a maximum window size* $T_0 < T$*, and an overlap fraction* $0 < f < 1$*.*

1. *For each* $\{(s,t): \ 1 \leq s < t \leq T, \ t-s < T_0\}$*, compute* $z_{s,t,\mathrm{obs}}$*, the observed value of (2.10), and let* $p_{s,t} = \mathrm{pr}(Z_{\max} > z_{s,t,\mathrm{obs}})$ *denote the global p-value associated with the interval* $(s,t]$*.*

2. *Let* $\mathcal{S} = \{(s,t) : p_{s,t} < \alpha\}$*. Rank the pairs in* $\mathcal{S}$ *from smallest p-value to largest.*

3. *Starting from the first element in* $\mathcal{S}$*, if it overlaps by more than* f *with any of the segments ranked before it in* $\mathcal{S}$*, eliminate it from* $\mathcal{S}$*.*

The set of variant intervals reported would be the final set $\mathcal{S}$*.*

Algorithm 2.2. *Fix the global significance level* α*, parameter* p*, and a maximum window* $T_0 < T$*. We denote by* $\boldsymbol{y}_{h:k}$ *the matrix* $\{y_{i,t} : \ 1 \leq i \leq N, \ h \leq t \leq k\}$*.*

1. *Initialize* $T_1 = 1$ *and* $T_2 = T$*.*

2. *Compute*

$$Z_{\max} = \max_{\substack{T_1 \le s < t \le T_2 \\ 1 \le t-s \le T_0}} \{Z^{(p)}(s,t)\}.$$

   Let (s^*, t^*) *be the maximizing interval.*

3. *If the p-value of* $Z_{\max}$*, as computed using the approximations in Section 2.4, is less than* α*, then for each* $(u,v) \in \{(T_1, s^*-1),\ (s^*,t^*),\ (t^*+1, T_2)\}$*, do:*

   (a) *Determine which samples carry the variation, as described below. If a sample carries the variation, let* $\hat{\boldsymbol{y}}_{i,u:v} = \bar{\boldsymbol{y}}_{i,u:v}$*, and for the other samples let* $\hat{\boldsymbol{y}}_{i,u:v} = \bar{\boldsymbol{y}}_{i,T_1:T_2}$*. Let* $\boldsymbol{y}'_{u:v} = \boldsymbol{y}_{u:v} - \hat{\boldsymbol{y}}_{u:v}$*.*

   (b) *Repeat steps 2-3 for* $T_1 = u$*,* $T_2 = v$ *and the newly normalized* $\boldsymbol{y}'_{u:v}$*.*

This second algorithm is understandably slower than Algorithm 2.1, because recomputation of $Z_{\max}$ for each of the three sub-segments every time a changed segment is found is an $O(NTT_0)$ operation. Thus, if T is large, and if there are many variant intervals, the algorithm is much slower than Algorithm 2.1. However, it is as fast as separately applying CBS (Olshen et al., 2004) to each of the individual sequences.

When a variant interval is identified across samples, it is often of interest to determine its carriers. This is in fact a necessary part of Algorithm 2.2. One approach is to classify as carrier those samples whose statistic $U^2_{i,s,t}$ falls above a suitable threshold. A second approach for classifying sequences is to use the absolute difference in median between points inside $(s,t]$ and points outside that interval. Certain experimentally verified CNVs contain only one or two SNPs, but are shared across a substantial proportion of the samples. These short intervals are a significant part of the motivation for multi-sample analysis, although they are often ignored in favor of long intervals of small change if $U^2_{i,s,t}$ is used by itself to select individual samples.

For application to DNA copy number data in Section 4, we use a combination of both types of thresholding. If a multi-sample scan identifies a variant interval at $(s, t]$, we decide that the ith sample is a carrier if both of the following two conditions hold:

1. the difference in median between inside and outside of the region is greater than $\delta^{\mu}\hat{\sigma}_i$,
2. the p-value of the sequence and interval specific chi-square statistic, $U^2_i(s, t)$, is less than δ^{χ^2}.

We chose the thresholds δ^{μ} and δ^{χ^2} to achieve the best performance on a set of validation data described in Section 4.3. These rules for classifying the samples rely on two quantities: The effect size (shift in mean divided by standard deviation) and length of the interval. Figure 2 shows the region in the (effect size) × (interval length) plane where a sample would be classified as a carrier, using values of δ^{μ} and δ^{χ^2} that work well on the validation data set. Figure 2 also shows the detection curve for a single sample scan of the entire genome containing 500,000 Illumina probes at a maximum window size of 200 and global p-value of 0.01. The area between the two detection boundaries are those (effect size) × (interval length) combinations that are missed in a single sample scan, but possibly detectable in a multi-sample scan through the pooling of information across samples.

These classification rules are designed specifically for analysis of DNA copy number data. For other types of data, different rules for identifying sequences carrying the variant intervals, perhaps incorporating problem specific knowledge and objectives, may be appropriate.

3 Numerical Experiments

We used Monte Carlo simulations to test the accuracy of the significance level approximation (2.12) for $N = 100$, $T = 500$, $T_0 = 50$, and prior odds $r_p \in \{0, 10, 100\}$. The results are plotted in Figure 3. We see that the analytic approximations agree reasonably well with Monte Carlo results for small p-values (< 0.05). The quality of the approximation degrades with an increase in r_p. However, for the range of r_p that we examined in this numerical experiment, the analytical approximations are close enough to provide useful practical guidelines.

Approximation (2.12), which involves the assumption of normality in estimating the multiple-testing correction, performs well when N is large. For small values of N, approximation (2.13) should be used for the sum-of-chisquares test statistic. Figure 4 compares Monte Carlo simulation results with approximations (2.12) and (2.13) at values of $N = 3, 20$ and 40. As expected, the approximation (2.12) is too conservative for small values of N, but (2.13) is fairly accurate.

We also evaluated the power of the test under the same settings ($N = 100$, $T = 500$, $T_0 = 50$) at different levels of signal strength ξ and population frequency π, as defined in Section 2.5, and at different values for the prior odds r_p. Power is evaluated at significance level 0.05, with the rejection threshold computed theoretically via (2.12). It is our expectation that, due to the effect of r_p in downweighting small chi-square values, one would gain power using larger r_p if the proportion of carriers π in the set of samples is small. Figure 5, which shows the curve of power versus ξ at different levels of π and different r_p, confirms this expectation. We expect that for different sample sizes, sequence lengths, or significance levels, the relationship between the curves would be different. However, the power can be approximated via a fast and easy formula, and thus in applications can be directly computed for the scenario of interest.

4 Analysis of DNA Copy Number Data

4.1 Scientific Background and Pre-processing of CNV Data

DNA copy number variants (CNVs) are an important class of genetic variation (recently reviewed in Scherer et al. (2007)), and may underlie a broad spectrum of human traits and diseases (Fanciulli et al., 2007; Perry et al., 2007; Hollox et al., 2007). Some CNVs are inherited and, as for other forms of genetic variation, can attain high allele frequencies in the population and become common or "recurrent" CNVs (Khaja et al., 2007; Redon et al., 2006; Conrad et al., 2006; McCarroll et al., 2006). Some other CNVs are de novo, i.e, generated by germline mutations and are observed in a child but not in his/her parents (Turner et al., 2007). Finally, there is also the category of somatic CNVs, most noticeably those occurring in cancer cells (reviewed in Pinkel & Albertson (2005)). These CNVs may confer growth advantages and are observed in a high proportion of cells in a given tumor sample, or in a large number of samples of a given kind of tumor.

In this paper, we focus on the de novo detection of inherited CNVs. Since these CNVs are population level polymorphisms due to a single mutation event in the history of the cohort, the break points should be exactly shared between samples. These CNVs are usually relatively short and often involve only 1 copy changes. Since the signal within each sample is weak, a joint analysis across samples can boost power.

Although not the focus of this paper, we note that Algorithm 2.2 is very useful for obtaining a sparse cross-sample summary for a set of samples. Thus, it is especially useful for copy number analysis in cancer genomes, where a sparse representation of the genome profiles of a set of tumors is often needed for downstream analysis, such as for building predictive models of clinical outcome. In current literature, this is usually done post segmentation. However, chromosomal breakage does not occur at random in the genome. Instead, they appear in "hot spots" distributed unevenly across the chromosomes, and often re-use the same CNV-prone breakpoint junctions. Such patterns of recurrence result from the sequence-specific nature of some of the key molecular mechanisms responsible for producing new CNVs, such as nonallelic homologous recombination and retroposition events (Korbel et al., 2007). As long as some CNV boundaries are shared across samples, a joint analysis which incorporates information across all samples analyzed is likely to be statistically more robust.

Existing approaches for cross-sample analysis of DNA copy number fall into the following categories:

1. Frequency plots: In this approach, each sample is separately segmented into regions of amplification, deletion, or normal copy number. Then, the smoothed profiles are aligned, and the frequency of amplification or deletion across the sample cohort is plotted versus the location of each segment. Regions where the frequency is above a certain threshold are considered regions of interest, and kept for downstream analyses.

2. STAC Diskin et al. (2006): Each sample is separately segmented into regions of amplification, deletion, or normal copy number. Then, the samples are aligned,

high-frequency aberrations are defined, and a permutation method is used to assess the significance of high-frequency aberrant regions in the sample set.

3. HMM based methods: In Shah et al. (2007), a multi-layer hierarchical hidden Markov model is used to segment all samples simultaneously. This method involves much more restrictive assumptions on the way that copy number changes are shared across samples. For example, it assumes that all carriers of a given CNV must have a change in the same direction, which is often violated in copy number data from normal samples, as seen in the example in Section 4.4. It also assumes that all deletions (or gains) for a given sample must have the same underlying mean, which we also show to be violated in our data set in Figure 1(c,d). A hidden Markov model based approach is also proposed in Wang et al. (2008), where the change-points are not assumed to be shared across samples. While Wang et al. (2008) focused on the analysis of cancer data, they mentioned that a shared change-point model would be desirable for the detection of inherited CNVs, as well as noted the enormous computational task that is inherent to a hidden Markov model solution for this problem. The output of the methods from both Shah et al. (2007) and Wang et al. (2008) is a plot by location of the probability of aberration in any of the samples.

In summary, with the exception of Shah et al. (2007); Wang et al. (2008), most current studies take the following approach: First, each sample is processed by using existing copy number estimation methods. Then, the smoothed profiles for the samples are aligned, and recurrent regions are identified as where the frequency of aberration across samples is high. We argue, and will show preliminary evidence below, that it is beneficial to pool data across samples during the initial segmentation step. We propose an alternative to hidden Markov models Shah et al. (2007) that can computationally handle thousands of samples simultaneously, rely on less restrictive model assumptions, and involve more transparent tuning parameters.

4.2 Data Preprocessing

CNV data contains well documented artifacts, which needs to be removed by preprocessing. One artifact is local trends, which were first noted in the statistics literature by Olshen et al. (2004). These local trends correlate with GC content (Bengtsson et al., 2008), and manifest as local low magnitude shifts in mean that is reproducible across samples. We observe that on many platforms (including Illumina and Affymetrix), the local trends in normal samples can be well estimated by the first principal component of the matrix of y values. This is because in normal samples, CNVs are very short as compared to the total sequence length, and thus the variation in the data is dominated by the local trends. Curiously, in most of the data sets we encountered, the local trends seem to fluctuate continuously in very few dimensions that is captured by the first few principal components. Thus, we normalize the data by reducing it to the residuals of its projection on the first 2 principal components.

Still another artifact is badly behaving individual SNPs, which give observations that are quite different from background in both cross-sample mean and variance. We standardize each SNP to have median 0 and inter-quartile range 1, to ameliorate the effect of badly performing SNPs.

4.3 Detection Accuracy of Inherited CNVs

We assess the accuracy of our CNV detection method on a set of 62 Illumina 550K Beadchips. The experiments were performed on DNA samples extracted from lymphoblastoid cell lines derived from healthy individuals, and were used as part of the Quality Assessment panel in a genomewide association study recently carried out at the Stanford Human Genome Center. The 62 samples represent

- 10 sets of (child, parent, parent) trios,
- 16 pairs of technical replicates for 16 independent DNA samples.

To assess detection accuracy, we compare CNVs identified for the two technical replicates of the same individual and those identified for the child with those identified for the parents. It is not possible to estimate type 1 and type 2 error rates from the data, but it is possible to define other measures of accuracy. Specifically, we define "inconsistency" of detections of CNVs in individual samples as follows:

- If a detected CNV in one of the replicate pairs is not detected in the second sample of the pair, the CNV is considered inconsistent. In this case, either the detection is a false positive or there is a false negative in the other sample.
- If a detected CNV in the child is not detected in at least one of the parents, it is considered inconsistent. In this case, neglecting the rare event that the detection represents a de novo mutation, either the detection made in the child is a false positive or there is a false negative in one or both of the parents.

In this way, detections made in the child samples and in the replicate sample pairs can be classified as consistent or inconsistent. The detections made in the parent samples are used only to validate the detections made in the child samples, and are not counted towards the total number of detections. Detection accuracy is thus assessed by plotting the number of consistent versus inconsistent detections, and different methods can be compared in such a plot. As described in the previous Section, after an interval of CNV is found at a location $(s, t]$, one still needs to decide, for each sample, whether it carries the CNV, and the method for doing this affects the level of consistency. For example, if all of the samples are classified as "changed" at all CNV locations, then there would be no inconsistencies. The preceding section suggests some practical thresholding solutions for classification of samples given that a location is detected as variant.

Figure 6 shows the results for different settings of the parameters r_p and the sample detection thresholds. The horizontal axis is the number of consistent detections and the vertical axis is the number of inconsistent detections. For example, if a variant interval is found, and 5 samples are determined to have that variant interval, it contributes 5 detections to the total. If 3 of those detections are validated, then that adds 3 to the number on the horizontal axis. Note that in the parent child trios, a parent can validate a child but not vice versa. Each line in the graph represents a different setting for r_p, and dots on the line refer to performance at varying δ^{μ}_{MIN}, where δ^{μ}_{MIN} is the parameter described in the previous Section. As δ^{μ}_{MIN} decreases, the total number of detections, as well as the number of inconsistencies, increases.

Figure 6 also plots the results obtained by segmenting each sample individually using CBS. The curve for CBS is obtained by varying the p-value parameter in the CBS algorithm. We can see by comparing the multi-sample segmentation Algorithm

2.1 and CBS that pooling information across samples does indeed improve accuracy. For example, with 200 inconsistent detections, CBS finds fewer than 200 consistent variations, while Algorithm 2.1 with $r_p = 0$ finds more than 400 consistent variations. For these data, the best value for r_p in terms of achieving the highest proportion of consistent detections is $r_p = 0$.

We expect that CNVs found in these samples are inherited changes. Consistent with this expectation, we found that long variant intervals are rare and that there is increased power to detect short intervals from pooling data across samples. For a summary of the length (number of SNPs spanned by the variation) and proportion (number of samples that carry the variation) for all consistent detections made in the child and replicate pair samples, see Figure 7 .

Figure 8 shows heat maps of example regions from these data. Heatmaps of the entire data set and the estimated change-points, which can be found in the online supplement, show plots of the data in blocks of 1000. Figure 8 shows a few smaller regions in finer detail. For each heatmap, the rows correspond to samples, and the columns correspond to SNPs. The top panel is the raw data, the bottom panel is the estimated copy numbers. The copy number estimated were obtained with $r_p = 0$, $\delta^{\chi^2}_{MIN} = 10^{-4}$, and $\delta^{\mu}_{MIN} = 0.4$. Each panel has two sample sets separated by a horizontal blue line. The samples above the blue line are the (child, parent, parent) trios; each trio is plotted together in that order. The samples below the blue line are the replicate pairs; the pairs are laid next to each other. Therefore the CNVs detected above the blue line should occur in 2-out-of-3's, whereas the CNVs detected below the blue line should occur in pairs. As one can verify from inspection of the heatmaps, most of the consistent CNVs are very short and occur in only a small fraction of this cohort.

A substantial fraction of the detections are inconsistent. From visual inspection, we believe that many of the inconsistencies are caused by two types of experimental artifacts: (1) Low quality SNPs, which have higher variance than the rest of the data and produce a larger sum-of-(weighted)-chisquares statistic. Fortunately, on many platforms these SNPs can be flagged in the data normalization step and thrown out. (2) Local trends, which have been observed in Illumina Beadarray data as well as in other platforms. These trends occur for various reasons that are not well understood (Olshen et al. (2004)), and are often shared by samples that are processed in the same batch. Checking whether samples that carry a low frequency variation belong to the same batch is a good way to spot this artifact.

4.4 Example Analysis of a Complex Region

As is well documented in the Database of Genomic Variants (Iafrate et al., 2004), chromosome 22 contains a complex region of nested deletion at cytoband 22q11 that has several variant forms in the population. Many samples among the 62 sample data set we described in Section 4.3 carry this variant region, as is clearly noticeable in the heatmap of Figure 9. Since this variant interval contains nested changes, Algorithm 2.2 is preferred to Algorithm 2.1 for its analysis. We use this example to illustrate the application of Algorithm 2.2.

We consider only the first 2000 SNPs mapping to chromosome 22, shown completely in the top panel of Figure 9. We applied Algorithm 2.2 to this region with parameters

$\alpha = 0.001$, $r_p = 0$, $\delta^{\mu} = 0.2$, and $\delta^{\chi^2} = 0.001$. The segmentation is shown in the lower panel of Figure 9. From the segmentation, we see that there are 3 visually noticeable variant regions. The first region is at SNPs $(416, 442)$, which corresponds to positions 17,017-17,368 kilobases. Compared to the rest of the cohort, both gains and losses in this region are observed. The second region spans SNPs 996 to 1329 (positions 20706 to 21549 kilobases), and contains several layers of nested deletions with change-points at SNPs $(1167, 1217, 1309, 1321)$. corresponding to chromosomes positions $(20996, 21110, 21379, 21436)$ Kb. Comparing the top and bottom panels of Figure 9, we see that the recursive Algorithm 2.2 reconstructs this complex region quite well. The third visible copy number variant is SNPs 1830-1880 (at positions 23986-24234 kilobases), which also has at least 3 copy number levels in this sample set. All of the copy number estimates in the child and replicate samples for these three variant regions are validated.

5 Conclusions

We have discussed a general statistical problem: simultaneous detection of shared change-points that define variant intervals in a subset of a collection of sequences. We have shown the potential advantages of such an approach to the analysis of chromosomal copy number variation of DNA sequences.

The formulation we have chosen was motivated by the success of Olshen et al. (2004) in their analysis of CNVs in single sequences. It is doubtful that any one approach can be optimal in problems of this complexity, and it would be useful to extend other single sequence methods to deal with multiple sequences. A useful version of hidden

Markov models would be particularly welcome. There is one multi-sequence HMM of which we are aware (Shah et al., 2007) and for which there is readily available and easily used software. However, in our experience it would not run in any reasonable length of time on even moderate numbers of sequences. We are also developing a multi-sequence version of the Bayes Information Criterion for model selection that we used for single sequence analysis (Zhang and Siegmund, 2006). There are a number of ad hoc modifications of single sequence methods that have been suggested for dealing with multiple sequences. It would be interesting to make a comparison of these methods along the lines of Lai et al. (2007) for single sequences.

We have concentrated on inherited CNVs, a majority of which occur in relatively short and non-overlapping intervals across individuals. We also studied cancer-related CNVs, which are often substantially longer and more complex. The straightforward Algorithm 2.1 seems sufficient to detect most inherited CNVs. The substantially more complex Algorithm 2.2, developed to deal with cancer related CNVs, contains two free parameters. While these parameters are to some extent arbitrary, they are easily interpreted and compared. Additional empirical experimentation may be required to determine the stability of the parameter values we have used.

The advantages of simultaneous analysis of multiple sequences is most apparent for inherited CNVs, which are hypothesized to align because of a common mutational origin and which often give a signal too weak to be detected in single sequence analysis. Cancer related CNVs are typically longer and can often be detected in single sequence analysis. In this case a potential advantage of simultaneous analysis is a relatively clean cross-sample summary of the data for downstream calculations trying to discover the relationship between CNVs and cancer phenotypes.

6 Appendix

6.1 Proof of (2.12)

Here we sketch the theoretical arguments leading to the approximation (2.12). The analysis given is a slight modification of the methods used in Siegmund (1988). Instead of the process $Z^{(p)}(s,t)$, we consider a more general process

$$Z^{f,N}_{s,t} = \{N^{1/2}\sigma_f\}^{-1}\sum_{i=1}^{N}[f\{U_i(s,t)\} - \mu_f],$$

where $U_i(s,t)$ is the χ-distributed random variable defined as in (2.3) for sample i, f is an arbitrary "well-behaved" function, and $\mu_f = \mathrm{E}f\{U_i(s,t)\}$, $\sigma_f^2 = \mathrm{var}f[U_i(s,t)]$. For simplicity of notation we sometimes omit f and N in our notation and simply write $Z(s,t)$, μ, and σ. Since $Z_{s,t}$ is a mean and variance standardized sum of N independent and identically distributed processes, for large N, $Z_{s,t}$ is approximately a Gaussian process on the two dimensional indexing set $D = \{(s,t) : 1 < s < t < T, \quad t - s < T_0\}$ with zero mean and covariance function

$$\rho(s,t,u,v) = \mathrm{cov}[Z_{s,t}, Z_{s,t}] = \frac{\mathrm{cov}[f\{U_1(s,t)\}, f\{U_1(u,v)\}]}{\sigma_f^2}. \tag{6.1}$$

Note that because of the standardization, for any $(s,t) \in D$, $\rho(s,t,s,t) = \mathrm{var}[Z_{s,t}] = 1$. Let

$$J = J(s,t) = \{(u,v) \in D : v < t \text{ or } v = t \text{ and } u < s\}.$$

Then,

$$\begin{aligned}
\mathrm{pr}(\max_{s,t\in D} Z_{s,t} > b) &= \sum_{s,t\in D}\int_0^\infty \mathrm{pr}(Z_{s,t}\in b+dx)\mathrm{pr}(\max_{u,v\in J_{s,t}} Z_{u,v} < b|Z_{s,t}=b+x) \\
&= \sum_{s,t}\frac{1}{b}\int_0^\infty \mathrm{pr}(Z_{s,t}\in b+dx/b)\mathrm{pr}(\max_{u,v\in J_{s,t}} Z_{u,v} < b|Z_{s,t}=b+x/b) \\
&\approx \frac{\varphi(b)}{b}\sum_{s,t}\int_0^\infty e^{-x}\mathrm{pr}(\max_{u,v\in J_{s,t}} b(Z_{u,v}-Z_{s,t}) < -x|Z_{s,t}=b)dx \quad (6.2) \\
& \qquad (6.3)
\end{aligned}$$

In (6.3), we applied the Gaussian approximation to the marginal distribution of $Z_{s,t}$. Later, we will give an improved approximation that corrects for non-normality. This correction is important because in most cases the distribution of $f(U)$ is highly skewed. We now treat the term inside the integral in (6.3). Again we regard Z as a Gaussian random field. Under this assumption the conditional mean and variance of $b(Z_{u,v}-Z_{s,t})$ are

$$\begin{aligned}
\mathrm{E}\{b(Z_{u,v}-Z_{s,t})|Z_{s,t}=b\} &= b^2\{\rho(s,t,u,v)-1\}, &(6.4)\\
\mathrm{var}\{b(Z_{u,v}-Z_{s,t})|Z_{s,t}=b\} &= b^2\{1-\rho^2(s,t,u,v)\}. &(6.5)
\end{aligned}$$

One can verify that ρ is not differentiable in u, v at $(u,v)=(s,t)$, but that the left and right derivatives have the same absolute magnitude. Let

$$\begin{aligned}
\rho_1'(s,t) &= \lim_{\delta\uparrow 0}\left|\frac{\rho(s,t,s+\delta,t)-\rho(s,t,s,t)}{\delta}\right|,\\
\rho_1'(s,t) &= \lim_{\delta\uparrow 0}\left|\frac{\rho(s,t,s,t+\delta)-\rho(s,t,s,t)}{\delta}\right| &(6.6)
\end{aligned}$$

then for small values of (ϵ_1, ϵ_2),

$$\rho(s,t,s+\epsilon_1,t+\epsilon_2) \approx 1 + \rho_1'(s,t)\epsilon_1 + \rho_2'(s,t)\epsilon_2,$$

$$\rho^2(s,t,s+\epsilon_1,t+\epsilon_2) \approx 1 + 2\rho_1'(s,t)d\epsilon_1 + 2\rho_2'(s,t)\epsilon_2,$$

and thus,

$$\begin{aligned} \mathrm{E}\{b(Z_{s+\epsilon_1,t+\epsilon_2} - Z_{s,t})|Z_{s,t} = b\} &\approx -b^2\left\{\rho_1'(s,t)\epsilon_1 + \rho_2'(s,t)\epsilon_2\right\}, & (6.7) \\ \mathrm{var}\{b(Z_{s+\epsilon_1,t+\epsilon_2} - Z_{s,t})|Z_{s,t} = b\} &\approx 2b^2\left\{\rho_1'(s,t)\epsilon_1 + \rho_2'(s,t)\epsilon_2\right\}. & (6.8) \end{aligned}$$

In a small neighborhood of $(s,t]$ the conditional process $b(Z_{s+\epsilon_1,t+\epsilon_2} - Z_{s,t})$ behaves like the sum of two independent random walks with negative drifts and with variances equal to twice the absolute drift. Because the drift is negative and of order $O(b^2) = O(T)$, we assume that the probability that $b(Z_{u,v} - Z_{s,t})$ crosses the $-x$ threshold when (u,v) are outside of a $O(1)$ neighborhood of (s,t) is negligible. Therefore, in the $O(1)$ neighborhood of (s,t), Lemma 4 of Siegmund (1988) applies to give the approximation:

$$\begin{aligned} & \mathrm{pr}\{\max_{u,v\in J_{s,t}} b(Z_{u,v} - Z_{s,t}) < -x|Z_{s,t} = b) \\ = \; & \mathrm{pr}\{\max_{n\geq 1} W_n \leq -x\}\mathrm{pr}\{\min_{n\geq 0} W_n + \min_{n\geq 1} W_n' \geq x\}, \end{aligned}$$

where W_n is a random walk with drift $-b^2\rho_1'$ and variance $2b^2\rho_1'$, while W_n' is a second random walk, independent of the first, with drift $-b^2\rho_2'$ and variance $2b^2\rho_2'$. Plugging this into (6.3), we have

$$\frac{\varphi(b)}{b}\sum_{s,t}\int_0^\infty e^{-x}\mathrm{pr}\{\max_{n\geq 1} W_n \leq -x\}\mathrm{pr}\{\min_{n\geq 0} W_n + \min_{n\geq 1} W_n' \geq x\}$$

Then, Lemma (21) from Siegmund (1992) can be used to evaluate the integral above to get

$$\mathrm{pr}\{\max_{s,t\in D} Z(s,t) > b\} \approx \varphi(b)b^3 \sum_{s,t\in D} \rho_1'(s,t)\rho_2'(s,t)\nu\left[b_0\{2\rho_1'(s,t)\}^{1/2}\right]\nu\left[b_0\{2\rho_2'(s,t)\}^{1/2}\right], \tag{6.9}$$

where $b_0 = b/N^{1/2}$.

6.1.1 Approximations of ρ_1', ρ_2'

First we look at the sum of chi-square statistic, where $f(x) = x^2$. In this special case, a simple approximate analytic form for $\rho(s,t,u,v)$ exists. For two values x and y, we use the notation $x \vee y = \min(x,y)$ and $x \wedge y = \max(x,y)$. Then for large T,

$$\rho(s,t,u,v) = \mathrm{cov}(U_{1,s,t}^2, U_{1,u,v}^2) = \frac{\{t\vee v - s\wedge u - (t-s)(v-u)/T\}^2}{(t-s)\{1-(t+s)/T\}(v-u)\{1-(v+u)/T\}}.$$

Computing the one-sided derivatives (6.6) for this correlation function, we have:

$$\rho_1'(s,t) = \rho_2'(s,t) = [(t-s)\{1-(t-s)/T\}]^{-1} \tag{6.10}$$

Noting that $\rho_1'(s,t)$ and $\rho_2'(s,t)$ are both functions only of $k = t - s$, and approximating the summation over (s,t) in (6.9) by an integral, (6.9) becomes

$$\varphi(b)b^3 \int_{\delta_0}^{\delta_1} \frac{1}{u^2(1-u)} \nu^2 \left[\frac{2^{1/2} b_0}{\{u(1-u)\}^{1/2}} \right] du, \tag{6.11}$$

which, after correction for non-Gaussianity of $Z_{s,t}$ as described in Section 6.1.2, would be equivalent to (2.12) in the case of $\beta = 1$.

Next, we consider general functions f where we may not know the explicit analytical form of $\rho[s,t,u,v]$. For small a,

$$\begin{aligned} & \mathrm{E}\{f(U_{s,t})f(U_{s-a,t})\} \\ \approx \;& \mathrm{E}\left[f(U_{s,t}) \left\{ f(U_{s,t}) + \sum_{k=1}^{\infty} f^{(k)}(U_{s,t})(U_{s-a,t} - U_{s,t})^k / k! \right\} \right]. \end{aligned} \tag{6.12}$$

An easy calculation of covariances shows that the numerator of $U_{s,t}$, namely $S_t - S_s - (t-s)S_T/T$ has the same joint distributions as the conditional joint distributions of $S_t - S_s$ given that $S_T = 0$. In what follows, it will be convenient to consider these conditional distributions, for which we will add a subscript of 0 to the usual notation for expectations, variances and covariances. Thus, for example, $\mathrm{E}(\cdot | S_T = 0) = \mathrm{E}_0(\cdot)$.

Let $r = t - s$ and $W_r = S_t - S_s$. The (conditional) distribution of $U_{s-a,t} - U_{s,t}$ is the same as that of

$$\frac{W_{r+a}}{[(r+a)\{1-(r+a)/T\}]^{1/2}} - \frac{W_r}{\{r(1-r/T)]^{1/2}}$$

$$= \frac{W_{r+a} - W_r}{[(r+a)\{1-(r+a)/T\}]^{1/2}} + W_r \left[\frac{1}{[(r+a)\{1-(r+a)/T\}]^{1/2}} - \frac{1}{\{r(1-r/T)\}^{1/2}} \right]$$

$$\approx \frac{W_{r+a}-W_r}{[(r+a)\{1-(r+a)/T\}]^{1/2}} + W_r\left[\frac{-1+2r/T}{\{r(1-r/T)\}^{3/2}}\right]\left(\frac{a}{2}\right). \tag{6.13}$$

Computing the first and second moment of $W_{r+a}-W_r$ conditioned on W_r, we have:

$$\begin{aligned}
\mathrm{E}_0\{W_{r+a}-W_r|W_r\} &= -aW_r/\{T(1-r/T)\},\\
\mathrm{var}_0\{W_{r+a}-W_r|W_r\} &= a(1-a/T)+O(a^2).
\end{aligned}$$

One can also verify that, since W_r is Gaussian, all higher (conditional) moments of $W_{r+a}-W_r$ are $o(a)$ and thus negligible. Therefore, we only need to keep the first two terms in the Taylor series expansion in (6.12). Letting

$$\kappa(r)=\frac{1}{r(1-r/T)},$$

we can use the conditional moments computed above to get

$$\begin{aligned}
\mathrm{E}\{U_{s-a,t}-U_{s,t}|U_{s,t}\} &= \left[\frac{r(1-r/T)}{(r+a)\{1-(r+a)/T\}}\right]^{1/2}\left\{\frac{-a}{T(1-r/T)}\right\}U_{s,t}\\
&\quad+\left\{\frac{-1+2r/T}{r(1-r/T)}\right\}\left(\frac{a}{2}\right)U_{s,t}\\
&\approx a\kappa(r)U_{s,t}/2,\\
\mathrm{E}\{(U_{s-a,t}-U_{s,t})^2|U_{s,t}\} &= \mathrm{E}_0\left[\frac{W_{r+a}-W_r}{[(r+a)\{1-(r+a)/T\}]^{1/2}}|W_r\right]+O(a^2)\\
&= \frac{a(1-a/T)}{(r+a)\{1-(r+a)/T\}}\\
&\approx a\kappa(r)/2
\end{aligned}$$

Plugging the above into (6.12),

$$\begin{aligned}\mathrm{E}\{f(U_{s,t})f(U_{s-a,t})\} &\approx \mathrm{E}\{f(U_{s,t})^2\} - \mathrm{E}\{f(U_{s,t})f'(U_{s,t})U_{s,t}\}\frac{a\kappa(t-s)}{2} \\ &\quad +\mathrm{E}\{f(U_{s,t})f''(U_{s,t})\}\frac{a\kappa(t-s)}{2}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{cov}\{f(U_{s-a,t}), f(U_{s,t})\} &= \mathrm{E}\{f(U_{s-a,t})f(U_{s,t})\} - \mu^2 \\ &= \sigma^2 + [\mathrm{E}\{f(U_{s,t})f''(U_{s,t})\} - \mathrm{E}\{f(U_{s,t})f'(U_{s,t})U_{s,t}\}]\frac{a\kappa(t-s)}{2},\end{aligned}$$

and

$$\begin{aligned}\rho_1'(s,t) &= \lim_{a\to 0} a^{-1}\left|\sigma^{-2}\mathrm{cov}\{f(U_{s-a,t}), f(U_{s,t})] - 1\right| \\ &= \frac{\mathrm{E}\{f(U_{s,t})f'(U_{s,t})U_{s,t}\} - \mathrm{E}\{f(U_{s,t})f''(U_{s,t})\}}{2\sigma^2}\kappa(t-s). \qquad (6.14)\end{aligned}$$

The computation of $\rho_2'(s,t)$ can be carried out exactly as above, since none of the steps except for (6.13) depends on whether we are differentiating s or t, and (6.13) relies only on $r = t-s$. Thus $\partial r/\partial s = -\partial r/\partial t$, although the a in (6.13) becomes $-a$ for $\rho_2'(s,t)$, so $\rho_2'(s,t) = \rho_1'(s,t)$. Given that $U_{s,t}$ is χ distributed and f, f' are known, $\rho_1'(s,t)$ can be computed numerically using this formula.

For example, using (6.14) on the simple one sample change-point case $f(x) = x$, we have

$$\frac{\mathrm{E}\{f(U_{s,t})f'(U_{s,t})U_{s,t}\} - \mathrm{E}\{f(U_{s,t})f''(U_{s,t})\}}{2\sigma^2} = \frac{\mathrm{E}\{U_{s,t}^2\}}{2} = 1/2,$$

giving us $\rho_1'(s,t) = \kappa(t-s)/2$, which, when plugged into (6.9), gives us the significance level approximation in Siegmund (1992). For the sum-of-chisquare statistic (2.5), $f(x) = x^2$ and $\sigma_f^2 = 2$, and therefore,

$$\frac{\mathrm{E}\{f(U_{s,t})f'(U_{s,t})U_{s,t}\} - \mathrm{E}\{f(U_{s,t})f''(U_{s,t})\}}{2\sigma^2} = \frac{2\mathrm{E}\{U_{s,t}^4\} - 2\mathrm{E}\{U_{s,t}^2\}}{4} = 1,$$

giving us $\rho_1'(s,t) = \kappa(t-s)$, which is the same as what we get by differentiating the exact form of the covariance function in (6.10).

6.1.2 Correction for non-normality

In (6.3), we used the Gaussian approximation

$$\mathrm{pr}(Z_{s,t} \in b + dx/b) \approx (2\pi)^{-1/2} \exp\{-(b + x/b)^2/2\} dx/b \approx \varphi(b) e^{-x} dx/b.$$

Since in most cases $f(U)$ is highly skewed (e.g. $f(U) = U^2$ in the sum of chi-square statistic), we replace the above with an improved approximation obtained as follows by a standard argument: Let $g(U) = \{f(U) - \mu_f\}/\sigma_f$. Then $Z_{s,t} = \sum_{i=1}^N g(U_{s,t,i})/N^{1/2}$. Let $\psi(\theta)$ be the log moment generating function of $g(U)$, and $\theta = \theta_{b,N}$ be the positive value that satisfies $N^{1/2}\dot{\psi}(\theta_{b,N}) = b$. This root is easily found numerically, since ψ is increasing on $(0, \infty)$ and convex. Let pr_θ be the tilted measure

$$\mathrm{pr}_\theta(g(U) \in dx) = e^{\theta x - \psi(\theta)} \mathrm{pr}\{g(U) \in dx\}$$

and let E_θ denote expectation under this measure. Then by a local central limit theorem

$$\begin{aligned} \mathrm{pr}(Z \in b + dx/b) &= \mathrm{E}_\theta\{e^{-\theta N^{1/2} Z + N\psi(\theta)}; Z \in b + dx/b\} \\ &\approx \{2\pi\ddot{\psi}(\theta)\}^{-1/2} e^{-\theta N^{1/2}(b+x/b)+N\psi(\theta)} dx/b. \end{aligned}$$

A simple linear approximation $\dot{\psi}$ for θ near 0 suggests the approximation $\theta \approx b/N^{1/2}$ (for Gaussian variables this is exact), and hence

$$\mathrm{pr}(Z \in b + dx/b) \approx \{2\pi\ddot{\psi}(\theta)\}^{-1/2} \exp(-\mathcal{I}) e^{-x} dx/b,$$

where $\mathcal{I} = N\{\theta\dot{\psi}(\theta) - \psi(\theta)\}$. This approximation, when used in (6.3) in place of $\varphi(b)\exp(-x)dx/b$ and combined with the appropriate value of β, leads to (2.12).

6.2 Modifications to prove (2.12)

We indicate briefly here modifications of the proof of (2.11) required to prove (2.12). Observe that c in (2.12) is given in terms of b in (2.11) by $c = \{b(2N)^{1/2} + N\}^{1/2}$. This means that c^2 and N are of the same order of magnitude when N is large. Also, the marginal distribution of $\tilde{Z}_{s,t}$ is χ with N degrees of freedom. From a straightforward approximation for large c of $\mathrm{pr}\{\tilde{Z}_{s,t} \in c + dx/c\}$, in which we do not neglect N/c^2 even though c is assumed large, we find that the simple exponential e^{-x} that arises under the integral sign in the last line of (6.2) now becomes $\exp[-x\{1 - (N-1)/c^2\}]$, while the normal density in front of the integral is replaced by the χ density evaluated at c.

Conditioning on $\tilde{Z}_{s,t}$, we now consider a two term Taylor series expansion of the increments $c(\tilde{Z}_{s+\epsilon_1,t+\epsilon_2} - \tilde{Z}_{s,t})$. We can by spherical symmetry assume without loss of generality that all the coordinates of the vector $(U_1(s,t),\ldots,U_N(s,t))'$ are zero except for the first one. The expansion of $c(\tilde{Z}_{s+\epsilon_1,t+\epsilon_2} - \tilde{Z}_{s,t})$ contains linear terms in the first coordinate direction in the form of the sum of two random walks indexed by ϵ_i, $i = 1, 2$ with (negative) means and variances proportional to $c^2/[2(t-s)\{1-(t-s)/T\}]$, cf. (6.6) and (6.7), and independent quadratic terms in the $N-1$ orthogonal directions with means proportional to $(N-1)/[2(t-s)\{1-(t-s)/T\}]$ and variances proportional to $(N-1)/[(t-s)\{1-(t-s)/T\}]^2$. Asymptotically important values of $t-s$ are of order c^2, so stochastic fluctuations of the quadratic terms are negligible. The consequence of adding $(N-1)/[2(t-s)\{1-(t-s)/T\}]$ to the means of the random walks is that both the exponential under the integral and the drift of the local random walks are modified by the same correction factor: $1-(N-1)/c^2$, while the variances of the local random walks remain unchanged. Hence Lemma 4 of Siegmund (1992) applies again to yield (2.12).

Remark. A similar problem was considered by Siegmund and Yakir (2000), but the dimension N was regarded as small and fixed, which made it reasonable to approximate a sphere of large radius in N dimensional Euclidean space locally by tangent hyperplanes. This leads to a similar approximation, but without the factor $1-(N-1)/c^2$, which arises in our analysis because N is sufficiently large that the curvature of the sphere should not be neglected. Numerical examples show that the simpler approximation deteriorates sharply with increasing N, while the accuracy of the new approximation is essentially independent of N, as Figure 2 illustrates.

References

Badner, J. & Gershon, E. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Molecular Psychiatry* **7**, 405–411.

Bengtsson, H., Irizarry, R., Carvalho, B. & Speed, T. (2008). Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**, 759–767.

Broët, P. & Richardson, S. (2006). Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911–918.

Conrad, D., Andrews, T., Carter, N., Hurles, M., & Pritchard, J. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nature Genetics* **38**, 75–81.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert Jr., C. J., Weber, B. L., Maris, J. M. & Grant, G. R. (2006). Stac: A method for testing the significance of dna copy number aberrations across multiple array-cgh experiments. *Genome Research* **16**, 1149–1158.

Engler, D., Mohapatra, G., Louis, D. & Betensky, R. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridications. *Biostatistics* **7**, 399–421.

Fanciulli, M., Norsworthy, P., Petretto, E., Dong, R., Harper, L., Kamesh, L., Heward, J., Gough, S., de Smith, A., Blakemore, A., Froguel, P., Owen, C., Pearce, S., Teixeira, L., Guillevin, L., Cunninghame Graham, D., Pusey, C., Cook, H., Vyse, T. & Aitman, T. (2007). Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature Genetics* **39**, 721–723.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. G. & Jain, A. (2004). Application of hidden markov models to the analysis of the array-cgh data. *Journal of Multivariate Analysis* **90**, 132–153.

Guha, S., Li, Y. & Neuberg, D. (2006). Bayesian hidden markov modeling of array cgh data. *Harvard University Biostatistics Working Paper Series* .

Hollox, E. J. J., Huffmeier, U., Zeeuwen, P. L. J. M. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C. M. C., Traupe, H., de Jongh, G., Martin, Reis, A., Armour, J. A. L. A. & Schalkwijk, J. (2007). Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet* **40**, 23–25.

Hsu, L., Self, S., Grove, D., Randolph, T., Wang, K., Delrow, J., Loo, L. & Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6**, 211–226.

Huber, W., Toedling, J. & Steinmetz, L. M. (2006). Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963–1970.

Iafrate, J. A., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics* **36**, 949–951.

Khaja, R., Zhang, J., MacDonald, J., He, Y., Joseph-George, A., Wei, J., Rafiq, M.A. and, Q. C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R., Lee, C., Scherer, S. & Feuk, L. (2007). Genome assembly comparison to identify structural variants in the human genome. *Nature Genetics* **38**, 1413–1418.

Lai, T. L., Xing, H. & Zhang, N. R. (2007). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290–307.

Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics* **21**, 3763–3770.

McCarroll, S., Hadnott, T., Perry, G., Sabeti, P., Zody, M., Barrett, J., Dallaire, S., Gabriel, S., Lee, C., Daly, M., Altshuler, D. & The International HapMap Consortium (2006). Common deletion polymorphisms in the human genome. *Nature Genetics* **38**, 86–92.

Newton, M., Gould, M., Reznikoff, C. & Haag, J. (1998). On the statistical analysis of allelic-loss data. *Statistics in Medicine* **17**, 1425–1445.

Newton, M. & Lee, Y. (2000). Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**, 1088–1097.

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

Perry, G. H. H., Dominy, N. J. J., Claw, K. G. G., Lee, A. S. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A. A., Mountain, J. L. L., Misra, R., Carter, N. P. P., Lee, C. & Stone, A. C. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nat Genet* .

Picard, F., Robin, S., Lavielle, M., Vaisse, C. & Daudin, J. (2005). A statistical approach for array cgh data analysis. *BMC Bioinformatics* **6**, 27.

Piccolboni, A. (2008). Multivariate segmentation in the analysis of transcription tiling array data. *Journal of Computational Biology* **15**, 845–856.

Pinkel, D. & Albertson, D. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37**, S11–S17.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, D. T., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Gonzalez, J. R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., Macdonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W. & Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature* **444**, 444–454.

Scherer, S., Lee, C., Birney, E., Altshuler, D., Eichler, E., Carter, N. &

Hurles, M. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**, S7–S15.

Shah, S. P., Lam, W. L., Ng, R. T. & Murphy, K. P. (2007). Modeling recurrent dna copy number alterations in array cgh data. *Bioinformatics* **23**, 450–458.

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

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Beck, S. & Hurles, M. E. (2007). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature Genetics* **40**, 90–95.

Venkatraman, E. & Olshen, A. (2007). A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics* **23**, 657–663.

Wang, H., Veldink, J. H., Ophoff, R. A. & Sabatti, C. (2008). Markov models for inferring copy number variations from genotype data on illumina platforms. *Technical Report, Dept. of Statistics, University of California at Los Angeles* .

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. (2005). A method for calling gains and losses in array-cgh data. *Biostatistics* **6**, 45–58.

Wen, C., Wu, Y., Huang, Y., Chen, W., Liu, S., Jiang, S., Juang, J., Lin, C., Fang, W., Hsiung, C. & Chang, I. (2006). A bayes regression approach to array-cgh data. *Statistical Applications in Molecular Biology* **5**.

Willenbrock, H. & Fridlyand, J. (2005). A comparison study: applying segmentation to arraycgh data for downstream analyses. *Bioinformatics* **21**, 4084–4091.

Wise, L., Lanchbury, J. & Lewis, C. (1999). Meta-analysis of genome scans. *Annals of Human Genetics* **63**, 263–272.

Zhang, N. & Siegmund, D. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* .

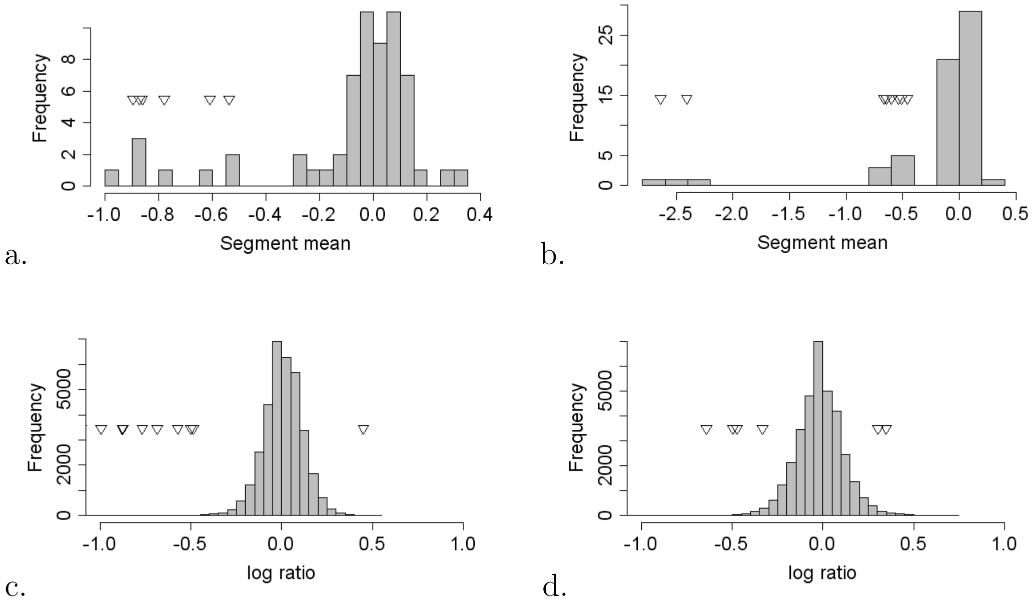


Figure 1. Histograms (a,b) show the sample means within two example copy number variant regions. There are 62 samples, so each histogram represents the counts for 62 numbers. Both CNV are deletion polymorphisms. The triangles show the means for the validated carriers among the samples. Observe that the triangles have a wide spread in values, suggesting that the model needs a separate mean shift for each sample within the same CNV. Figures (c,d) are histograms for $\{y_{it} : \quad t = 1, \dots, T\}$ for two different samples. The triangles show the values of $\delta_i(\tau_1, \tau_2)$ for validated variant intervals τ_1, τ_2 on chromosome 5 for that sample. Observe again that the triangles have a wide spread in values, suggesting that the shift in mean is different across variant intervals within the same sample.

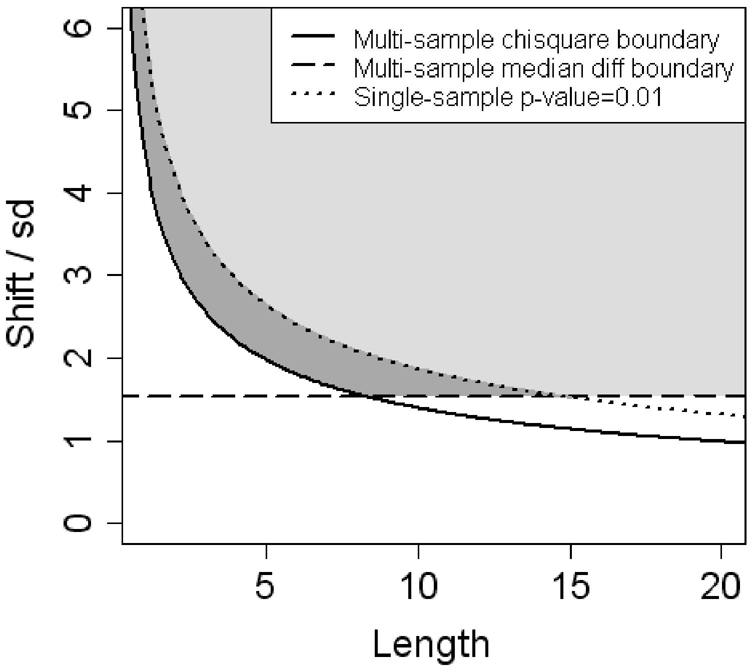
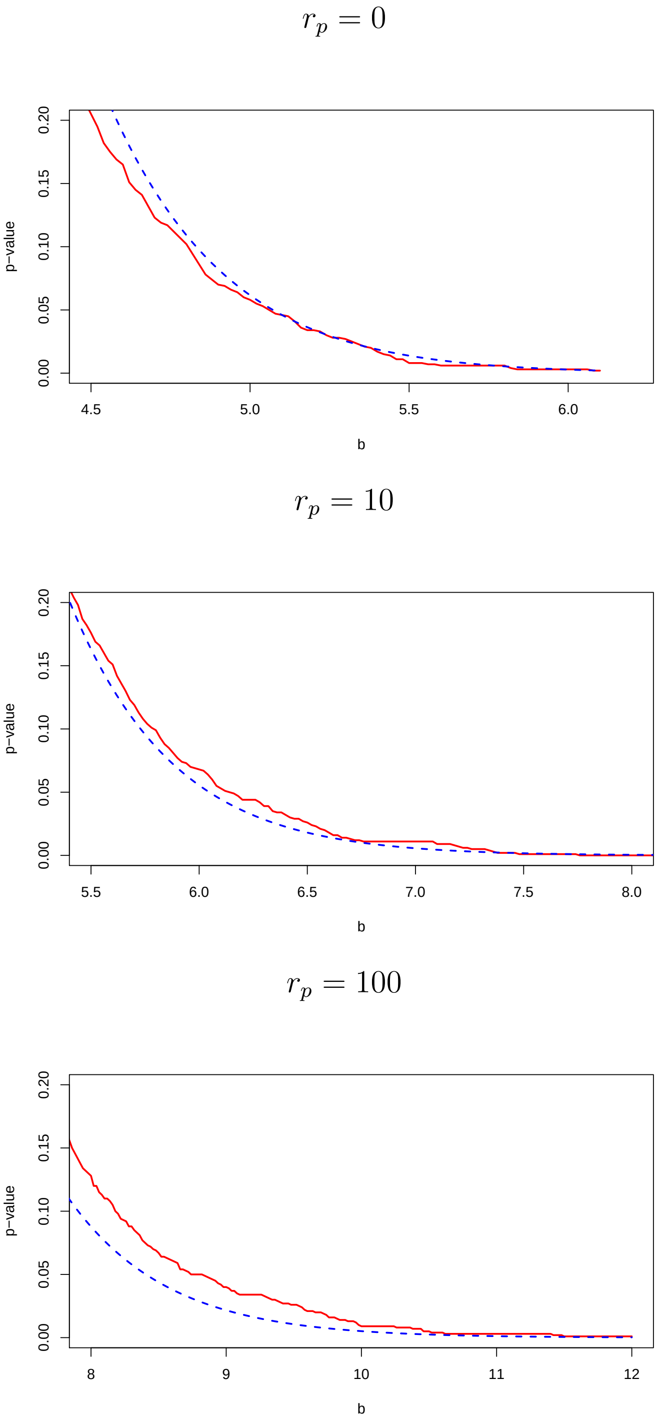


Figure 2. The light gray region shows the values of segment length $(\tau_2 - \tau_1)$ and effect size (δ_i/σ_i) that are classified as carrier for a detected variant interval. This region is determined by setting $\delta^{\chi^2} = 10^{-5}$ and $\delta^{\mu} = 1.5$. The dotted line is the rejection boundary for a single sample scan with $T = 500,000$ data points, $T_0 = 200$, and global p-value of 0.01. The dark gray region between the two boundaries contain those values that are missed in a single sample scan, but possibly detectable in a multi-sample scan.

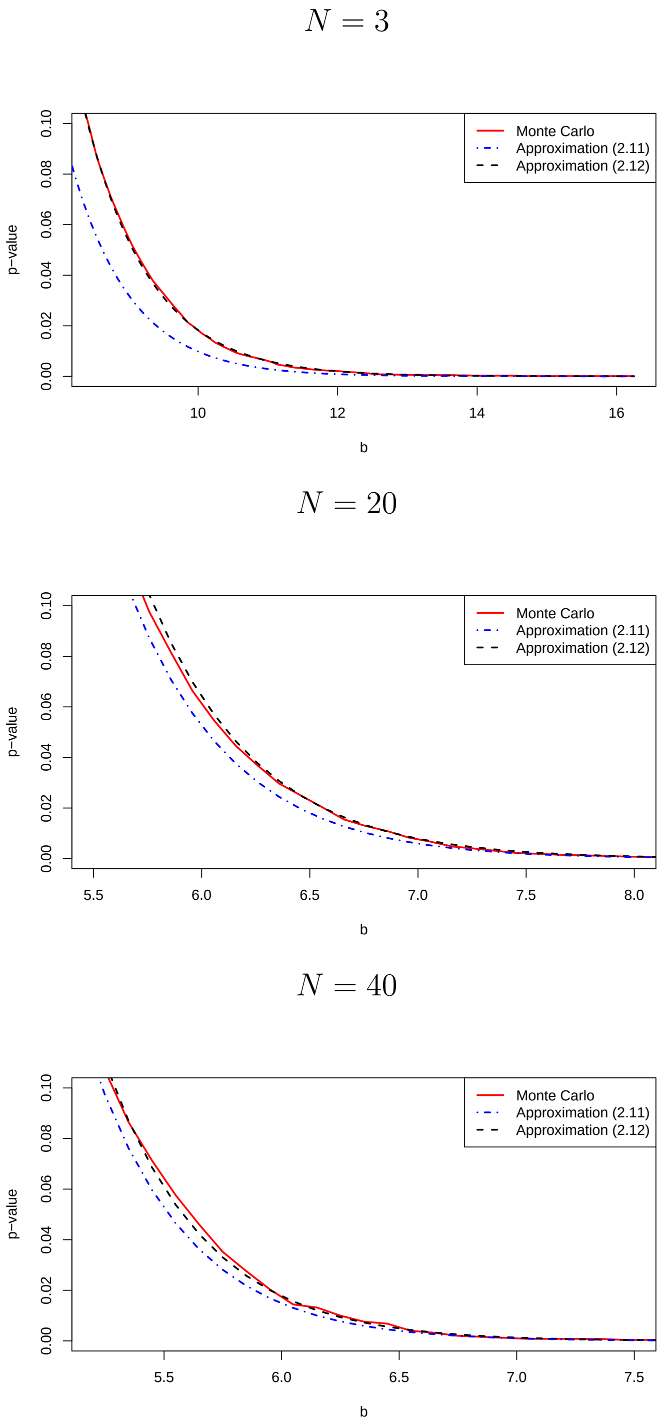


$r_p = 0$

$r_p = 10$

$r_p = 100$

Figure 3. Significance curves approximated by analytic formula and by Monte Carlo at setting $N = 100$, $T = 500$, $T_0 = 50$, and r_p values of 0, 10, and 100.



$N = 3$

$N = 20$

$N = 40$

Figure 4. Significances curve approximated by analytic formula and by Monte Carlo at setting $r_p = 0$, $T = 500$, $T_0 = 50$, and number of samples $N = 3, 20$, and 40.

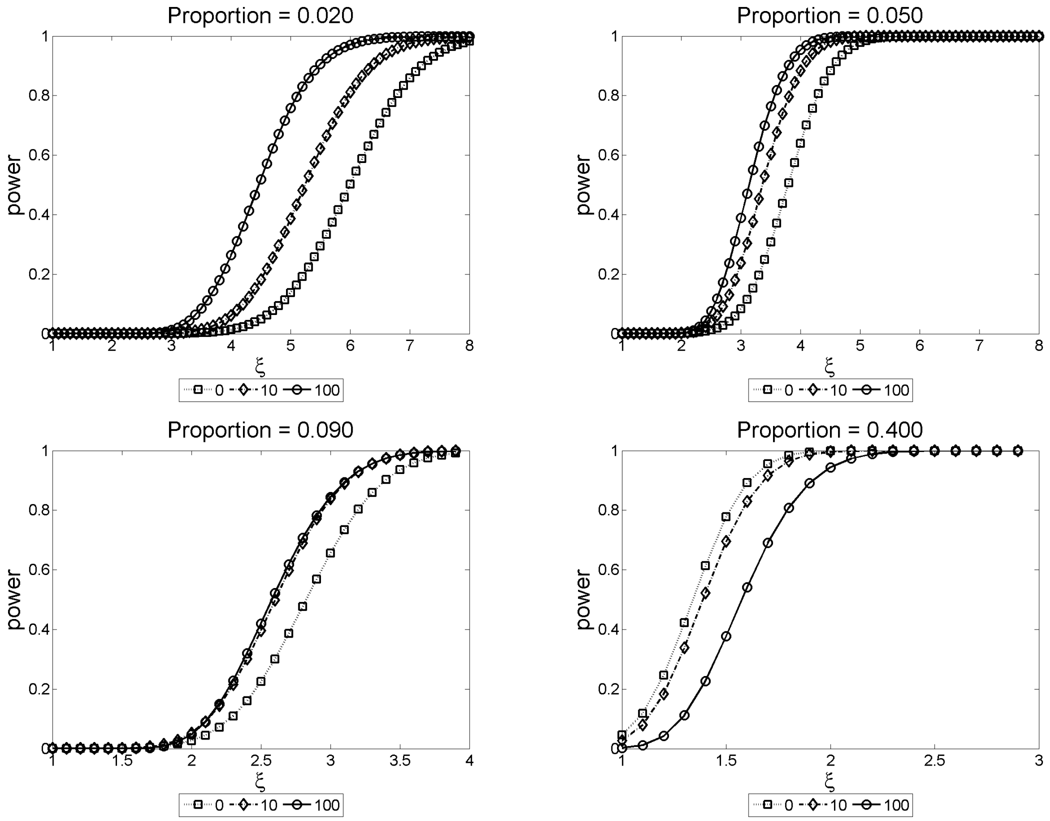


Figure 5. Power curves for varying levels of r_p and varying proportion. In this example, $N = 100$, $T = 500$, and $T_0 = 50$. Significance level is fixed at 0.05.

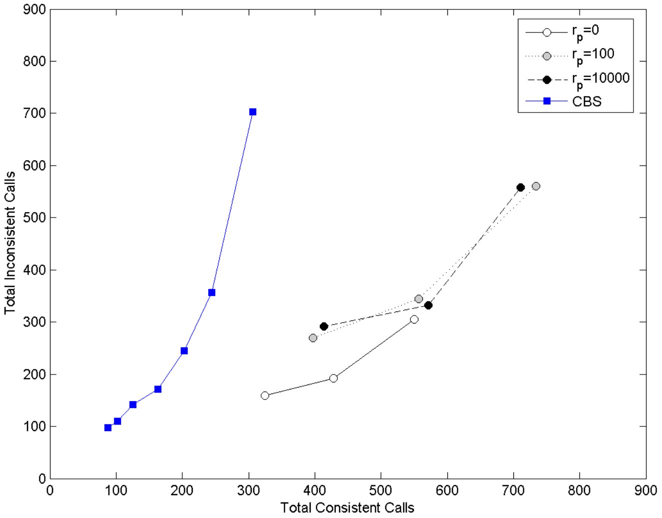


Figure 6. Comparison of a single sample method (CBS, Olshen et al. 2003) with cross sample segmentation using different values for r_p. The global significance value $\alpha = 10^{-3}$. The sample calling thresholds are $\delta^{\mu}_{MIN} \in \{0.2, 0.3, 0.4\}$ and $\delta^{\chi^2}_{MIN} = 10^{-4}$.

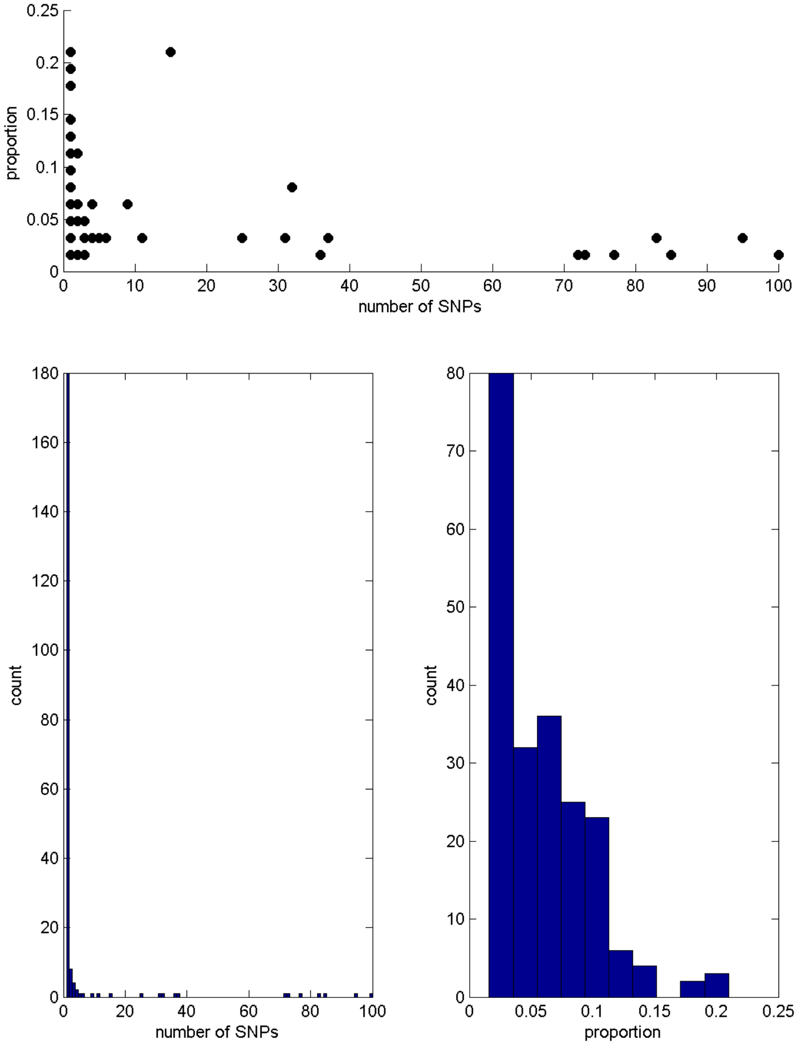


Figure 7. Summary of the consistent CNVs found for validation data set ($r_p = 0$, $\delta_{MIN} = 0.4$). Top: scatter plot of number of snps versus proportion of the 62 sample set that have the anomaly for the total of 211 regions identified. Bottom left: histogram of number of SNPs in each anomalous region. Bottom right: histogram of the proportion (%) of the samples that have the anomaly.

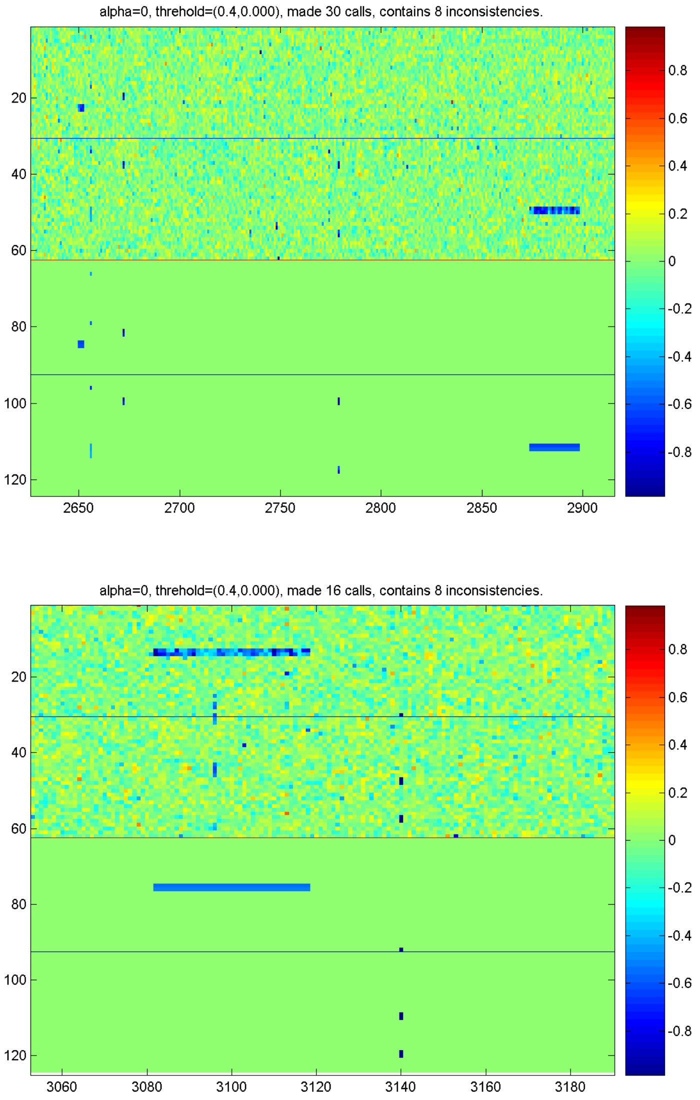


Figure 8. Example of two regions containing both multi-SNP and single SNP copy number variations in the 62 sample validation data. The parameters used were $r_p = 0$, $\delta^{\mu}_{MIN} = 0.4$, and $\delta^{\chi^2}_{MIN} = 10^{-4}$.

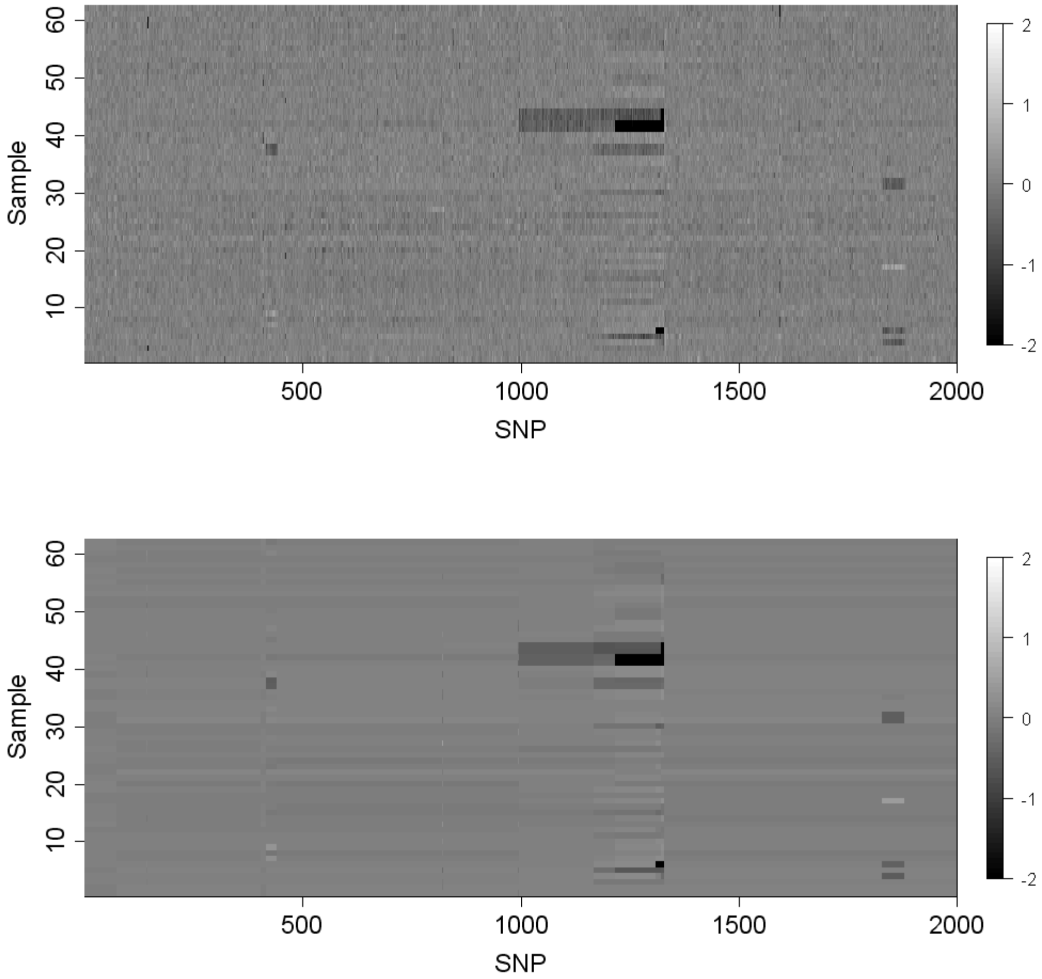


Figure 9. Example 2000 SNP region in cytoband 22q11 containing a complex CNV with nested deletions. Bottom panel shows segmentation given by Algorithm 2.2.