4-2008

Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis

Tze Leung Lai

Haipeng Xing

Nancy Zhang
University of Pennsylvania

Recommended Citation

Lai, T., Xing, H., & Zhang, N. (2008). Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis. *Biostatistics, 9* (2), 290-307. http://dx.doi.org/10.1093/biostatistics/kxm031

Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis

Abstract

Array-based comparative genomic hybridization (array-CGH) is a high throughput, high resolution technique for studying the genetics of cancer. Analysis of array-CGH data typically involves estimation of the underlying chromosome copy numbers from the log fluorescence ratios and segmenting the chromosome into regions with the same copy number at each location. We propose for the analysis of array-CGH data, a new stochastic segmentation model and an associated estimation procedure that has attractive statistical and computational properties. An important benefit of this Bayesian segmentation model is that it yields explicit formulas for posterior means, which can be used to estimate the signal directly without performing segmentation. Other quantities relating to the posterior distribution that are useful for providing confidence assessments of any given segmentation can also be estimated by using our method. We propose an approximation method whose computation time is linear in sequence length which makes our method practically applicable to the new higher density arrays. Simulation studies and applications to real array-CGH data illustrate the advantages of the proposed approach.

Keywords

Array-CGH, Bayesian inference, hidden Markov models, jump probabilities

Disciplines

Biostatistics

Stochastic segmentation models for array-based comparative genomic hybridization data analysis

Tze Leung Lai

Department of Statistics and Cancer Center, Stanford University,
Stanford, CA 94305-4065, USA
lait@stat.stanford.edu

Haipeng Xing

Department of Statistics, Columbia University, New York, NY10027, USA
xing@stat.columbia.edu

Nancy Zhang*

Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA
nzhang@stat.stanford.edu

SUMMARY

Array-based comparative genomic hybridization (array-CGH) is a high throughput, high resolution technique for studying the genetics of cancer. Analysis of array-CGH data typically involves estimation of the underlying chromosome copy numbers from the log fluorescence ratios and segmenting the chromosome into regions with the same copy number at each location. We propose for the analysis of array-CGH data a new stochastic segmentation model and an associated estimation procedure that has attractive statistical and computational properties. An important benefit of this Bayesian segmentation model is that it yields explicit formulas for posterior means, which can be used to estimate the signal directly without performing segmentation. Other quantities relating to the posterior distribution that are useful for providing confidence assessments of any given segmentation can also be estimated using our method. Simulation studies and applications to real array-CGH data illustrate the advantages of the proposed approach.

Key words and phrases: Array-CGH; Bayesian inference; Hidden Markov models; Jump probabilities.

**To whom correspondence should be addressed.*

1. INTRODUCTION

Array-based comparative genomic hybridization (array-CGH) has become a useful technology in studying the genetics of cancer. For a given cell sample, array-CGH allows quantitative measurement of the average genomic DNA copy number at thousands of locations linearly ordered along the chromosomes. Typically, a test genomic DNA pool (e.g. genomic DNA from tumor cell sample) and a diploid reference genomic DNA pool are differentially labeled with dyes. These two dye-labeled samples are mixed and hybridized to a microarray chip, which is spotted with genomic targets that map to known locations on a global scale throughout the genome. The hybridized chip is then scanned, and the ratio of the test and reference fluorescence intensities for each genomic target is calculated. The ratio of the intensities of the dyes is a surrogate for the ratio of the abundance of the DNA sample labeled with the dyes. The review by Pinkel and Albertson (2005) summarizes recent developments in this technology and its potential applications.

The first step in the analysis of array-CGH data is the estimation of the real copy number at each probe location from the log intensity measurements. Note that by "copy number" we actually refer to a continuous quantity that is the average copy number at a given location over all of the cells in the sample, which is often a heterogeneous population of cells with different copy numbers at any given genome location. In the last few years, several statistical approaches have been proposed for this problem, including hidden Markov models (HMM, Fridlyand et al. (2004)), recursive change-point detection (CBS, Olshen et al. (2004)), a Gaussian model-based approach (GLAD, Hupe et al. (2004)), hierarchical tree-style clustering (CLAC, Wang et al. (2005)), wavelet approximation (Hsu et al. (2005)), a Bayes regression approach (Wen et al. (2005)), and a pseudo-likelihood approach to Gaussian mixture models (Engler et al. (2006)). Most of these methods approach this problem through a segmentation perspective: they divide the genome into linearly contiguous segments with the same copy number. An important statistical problem in the implementation of such methods is determination of the number of segments, which is sometimes referred to as the smoothness of the segmentation. Information-based model selection (Picard et al. (2005), Zhang and Siegmund (2006)) has been proposed as a guideline to this issue. The reviews by Willenbrock and Fridlyand (2005) and by W.R. Lai et al. (2005) independently survey the effectiveness of existing methods on simulation and real data. Most methods produce a segmentation of the data but offer no way of assessing confidence in the segmentation. For complex aberration profiles, the different methods vary greatly on the location of breakpoints and the estimated signal level, which suggests that a framework for inference is crucial.

In this paper we propose for the analysis of array-CGH data a new stochastic segmentation model and an associated inference framework that has attractive statistical and computational properties. We view array-CGH experiments as producing, for each cell sample, an ordered sequence of (t, y_t) pairs, where t represents the location in the genome and y_t represents the log ratio of the test versus reference spot intensities for the genomic target from that location. The segmentation model in Section 2 assumes that $y_t = \theta_t + \sigma\epsilon_t$, in which ϵ_t are independent standard normal random variables and θ_t is an unknown step function whose prior distribution is given by a jump process with a baseline state and changed states. We assume that the baseline state is 0, since when there are no copy number changes the signal should be $\log 1 = 0$. From the baseline state the process can jump to a changed state that has a Gaussian prior. From a changed state it can jump to another changed state or jump back to the baseline.

Since the copy number of a homogeneous sample of normal cells should be 2 at all genomic locations, giving a signal of 0, the assumption of a zero baseline state is natural. Without making this assumption, most existing methods rely on a merging step after the segmentation to eliminate the small fluctuations around the baseline. The review by Willenbrock and Fridlyand (2005) suggests that ideally, a merging step should be incorporated into the initial segmentation so that not only are the results more interpretable but the additional information may allow higher sensitivity. This is accomplished in our method through the assumption of a baseline state. Whether non-baseline states with close mean levels should be merged is questionable. Inhomogeneity and micro-evolution within a cell sample may cause the copy number changes at different locations in the genome to have different mixture components.

An important benefit of our Bayesian segmentation model is that we can use the posterior distributions of the number and locations of the change-points to provide confidence assessments of a segmentation. Moreover, the posterior probability of copy number change can be readily computed for each genomic target, providing an easily interpretable value that can be used to rank or weight the genomic targets for downstream analysis. This quantity arises naturally from the model and is intuitively appealing and useful for probe-level analysis.

The Bayesian segmentation model contains certain hyperparameters. Their estimation is considered in Section 3 where other implementation issues are also discussed. In Section 4 we apply our method to several real arrayCGH-data sets and illustrate the usefulness of confidence assessments for different scenarios. Section 5 evaluates the performance of our method on simulated data that are generated from our and other models. Some concluding remarks are given in Section 6, in which we also compare our approach with existing methods in the literature.

2. A STOCHASTIC CHANGE-POINT MODEL WITH KNOWN BASELINE

2.1 Model with known baseline and unknown changed states

We assume a change-point model where the baseline state is known to be 0. When the signal leaves the baseline, it moves to a non-zero state; when the next jump occurs, the signal may move back to the baseline or jump to another non-zero state. Suppose the log fluorescence ratios y_t follow the model

$$y_t = \theta_t + \sigma\epsilon_t, \qquad \epsilon_t \sim N(0,1), \tag{1}$$

where θ_t is a piecewise constant function of t. To describe the dynamics of θ_t, we use the transition probability matrix

$$P = \begin{pmatrix} 1-p & \frac{1}{2}p & \frac{1}{2}p \\ c & a & b \\ c & b & a \end{pmatrix}. \tag{2}$$

The matrix P specifies that, at time t, if the state θ_t is in the 0 (baseline) state, then at time $t+1$, θ_{t+1} stays in the 0 state with probability $1-p$, or jumps to a nonzero state which follows $N(\mu, v)$ with probability p. To allow the possibility of jumping from a non-zero state to a different non-zero state, we simply assume that the process can jump from the baseline state with probability $p/2$ to either of two nonzero states that have the same prior distribution $N(\mu, v)$. If $\theta_t \neq 0$, then at time $t+1$, it can stay in the last state with probability a, or jump to another nonzero state with probability b, or jump back to the baseline state with probability c.

The probability vector $\tilde{\pi} = (c/(p+c), \frac{1}{2}p/(p+c), \frac{1}{2}p/(p+c))$ satisfies $\tilde{\pi}P = \tilde{\pi}$, and therefore $\tilde{\pi}$ corresponds to the stationary distribution associated with P. Note also that

$$\tilde{\pi}(x)P(x,y) = \tilde{\pi}(y)P(y,x),$$

so the three-state Markov chain with transition probability matrix P and initialized at $\tilde{\pi}$ is reversible. This implies that the Markov chain $\{\theta_t\}$ has a stationary distribution π that assigns probability $c/(p+c)$ to the baseline value 0 and probability $p/(p+c)$ to a $N(\mu, v)$

random variable. Moreover, under the additional assumption that θ_0 is initialized at the stationary distribution, $\{\theta_t\}$ is a reversible Markov chain; this property provides substantial simplification for the smoothing formulas in Section 2.3.

2.2 Filtering estimate of signal

Let $K_t = \max\{s \leq t : \theta_s = \cdots = \theta_t, \theta_{s-1} \neq \theta_s\}$ denote the nearest change-point at a location less than or equal to t. Let $\mathcal{Y}_n = (y_1, \ldots, y_n)$ and $\mathcal{Y}_{i,j} = (y_i, \ldots, y_j)$. Define

$$p_t = P(\theta_{K_t} = 0|\mathcal{Y}_t) = P(\theta_t = 0|\mathcal{Y}_t), \qquad q_{i,t} = P(\theta_{K_t} \neq 0, K_t = i|\mathcal{Y}_t) \tag{3}$$

for $1 \leq i \leq t$. Since the conditional distribution of θ_t, given $\mathcal{Y}_t$ and the event that $K_t = i$ and $\theta_{K_t} \neq 0$, is $N(\mu_{i,t}, v_{i,t})$, where

$$v_{i,j} = \left(\frac{1}{v} + \frac{j-i+1}{\sigma^2}\right)^{-1}, \qquad \mu_{i,j} = (\frac{\mu}{v} + \sum_{k=i}^{j} \frac{y_k}{\sigma^2})v_{i,j} \tag{4}$$

for $j \geq i$, it follows that the posterior distribution of θ_t given $\mathcal{Y}_t$ is a mixture of normal distributions and a point mass at 0:

$$\theta_t|\mathcal{Y}_t \sim p_t\delta_0 + \sum_{i=1}^{t} q_{i,t}N(\mu_{i,t}, v_{i,t}), \tag{5}$$

where δ_x denotes the probability distribution that assigns probability 1 to x. Let $\phi_{\mu,v}$ denote the density function of the $N(\mu, v)$ distribution, i.e., $\phi_{\mu,v}(y) = (2\pi v)^{-1/2}\exp\{-\frac{1}{2}(y-\mu)^2/v\}$. Making use of $p_t + \sum_{i=1}^{t} q_{i,t} = 1$ and $y_t = \theta_t + \sigma\epsilon_t$, we show in Appendix A that the conditional probabilities p_t and $q_{i,t}$ can be determined by the recursions

$$\begin{aligned} p_t &\propto p_t^* := (1-p)p_{t-1} + cq_{t-1}, \\ q_{i,t} &\propto q_{i,t}^* := \begin{cases} (pp_{t-1} + bq_{t-1})\psi/\psi_{t,t}, & i = t, \\ aq_{i,t-1}\psi_{i,t-1}/\psi_{i,t}, & i < t, \end{cases} \end{aligned} \tag{6}$$

where $q_t = \sum_{i=1}^{t} q_{i,t} = 1 - p_t$, $\psi = \phi_{\mu,v}(0)$ and $\psi_{i,j} = \phi_{\mu_{i,j},v_{i,j}}(0)$ for $i \leq j$. Specifically, $p_t = p_t^*/\left[p_t^* + \sum_{i=1}^{t} q_{i,t}^*\right]$ and $q_{i,t} = q_{i,t}^*/\left[p_t^* + \sum_{i=1}^{t} q_{i,t}^*\right]$. By (3) and (5),

$$P(\theta_t = 0|\mathcal{Y}_t) = p_t, \qquad E(\theta_t|\mathcal{Y}_t) = \sum_{i=1}^{t} q_{i,t}\mu_{i,t}. \tag{7}$$

2.3 Smoothing estimate of signal

As indicated at the end of Section 2.1, $\{\theta_t\}$ is a reversible Markov chain. Therefore we can reverse time and obtain a backward filter that is analogous to (5):

$$\theta_{t+1}|\mathcal{Y}_{t+1,n} \sim \widetilde{p}_{t+1}\delta_0 + \sum_{j=t+1}^{n} \widetilde{q}_{j,t+1}N(\mu_{t+1,j}, v_{t+1,j}), \tag{8}$$

in which the weights $\widetilde{p}_s, \widetilde{q}_{j,s}$ can be obtained by backward induction using the time-reversed counterpart of (6):

$$\widetilde{p}_s \propto \widetilde{p}_s^* := (1-p)\widetilde{p}_{s+1} + c\widetilde{q}_{s+1},$$
$$\widetilde{q}_{j,s} \propto \widetilde{q}_{j,s}^* := \begin{cases} (p\widetilde{p}_{s+1} + b\widetilde{q}_{s+1})\psi/\psi_{s,s} & j = s, \\ a\widetilde{q}_{j,s+1}\psi_{s+1,j}/\psi_{s,j} & j > s, \end{cases}$$

where $\widetilde{q}_{s+1} = \sum_{j=s+1}^{n} \widetilde{q}_{j,s+1} = 1-\widetilde{p}_{s+1}$. Since $P(\theta_t \in A|\mathcal{Y}_{t+1,n}) = \int P(\theta_t \in A|\theta_{t+1})dP(\theta_{t+1}|\mathcal{Y}_{t+1,n})$, it follows from (8) and the reversibility of $\{\theta_t\}$ that

$$\theta_t|\mathcal{Y}_{t+1,n} \sim [(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}]\delta_0 + (p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1})N(\mu, v) + a\sum_{j=t+1}^{n} \widetilde{q}_{j,t+1}N(\mu_{t+1,j}, v_{t+1,j}). \tag{9}$$

We can use Bayes' theorem to combine the forward filter (5) with its backward variant (9) to derive the posterior distribution of θ_t given $\mathcal{Y}_n$ $(1 \leq t \leq n)$, which is a mixture of normal distributions and a point mass at 0:

$$\theta_t|\mathcal{Y}_n \sim \alpha_t\delta_0 + \sum_{1\leq i\leq t\leq j\leq n} \beta_{ijt}N(\mu_{ij}, v_{ij}). \tag{10}$$

In particular, by Bayes' theorem,

$$\begin{aligned} \alpha_t = P(\theta_t = 0|\mathcal{Y}_n) &\propto P(\theta_t = 0|\mathcal{Y}_t)P(\theta_t = 0|\mathcal{Y}_{t+1,n})/\pi(0) \\ &= p_t[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}]/[c/(p+c)]. \end{aligned} \tag{11}$$

Applying a similar argument to the density function of the absolutely continuous component of the posterior distribution of θ_t given $\mathcal{Y}_n$ yields a formula that is proportional to β_{ijt}. The details are given in Appendix A, which shows that

$$\begin{aligned} &\alpha_t = \alpha_t^*/A_t, \qquad \beta_{ijt} = \beta_{ijt}^*/A_t, \qquad A_t = \alpha_t^* + \sum_{1\leq i\leq t\leq j\leq n} \beta_{ijt}^*, \\ &\alpha_t^* = p_t[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}]/c, \\ &\beta_{ijt}^* = \begin{cases} q_{i,t}(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1})/p, & i \leq t = j, \\ aq_{i,t}\widetilde{q}_{j,t+1}\psi_{i,t}\psi_{t+1,j}/(p\psi\psi_{i,j}), & i \leq t < j. \end{cases} \end{aligned} \tag{12}$$

From (10), it follows that

$$P(\theta_t = 0|\mathcal{Y}_n) = \alpha_t, \qquad E(\theta_t|\mathcal{Y}_n) = \sum_{1\le i\le t\le j\le n} \beta_{ijt}\mu_{ij}. \tag{13}$$

2.4 Inference on segmentation and parameter subsequences

The α_t and β_{ijt} in (12) are posterior probabilities that are useful for inference. As shown in (13), $\alpha_t = P(\theta_t = 0|\mathcal{Y}_n)$. Moreover, the derivation of (12) in Appendix A shows that, for $i \le t \le j$,

$$\beta_{ijt} = P(C_{ij}|\mathcal{Y}_n), \quad \text{where } C_{ij} = \{\theta_i = \cdots = \theta_j \ne 0, \theta_i \ne \theta_{i-1}, \theta_j \ne \theta_{j+1}\}. \tag{14}$$

For the problem of classifying location t as 0 (no copy number change, or normal), G (copy number gain) or L (copy number loss) considered by Engler et al. (2006), although the posterior probability $\alpha_t = P(\theta_t = 0|\mathcal{Y}_t)$ seems to provide an essential ingredient for constructing the Bayes classification rule, in practice gain, loss and no change actually include a margin w beyond which the location is considered aberrant due to copy number gain or loss. Specifically, location t is considered as G if $\theta_t > w$, as L if $\theta_t < -w$, and as 0 if $|\theta_t| \le w$. The choice of w is often based on statistical (e.g., w is some multiple of σ) and biological considerations. With G, L and 0 defined in this way, the Bayes rule R is a "soft" classifier determined by the posterior probabilities in the following:

(R) Classify location t as $\arg\max_s P(s|\mathcal{Y}_n)$, where $s = G$ on $\{\theta_t > w\}$, $s = L$ on $\{\theta_t < -w\}$ and $s = 0$ on $\{|\theta_t| \le w\}$.

Let $[i, j]$ denote the segment whose beginning and ending locations are i and j, respectively. We can use $P(C_{ij}|\mathcal{Y}_n)$ to provide confidence assessments of the abnormality (due to copy number change) of a segment $[i, j]$ obtained by a segmentation procedure of the type described in the second paragraph of Section 1. Typically these segmentation procedures allow some fuzziness in the specified endpoints i, j of the segment, in the sense that the actual endpoints may not be i and j but should be somewhere around them. To make this more precise, suppose the endpoints i and i' (or j and j') are considered "equivalent" if they differ by at most k locations, where $k = \min(k^*, \lfloor (j-i)/2 \rfloor)$ and k^* represents some pre-specified precision. Then we can use $P(\bigcup_{(i',j'):|i-i'|\le k,|j-j'|\le k} C_{i'j'}|\mathcal{Y}_n)$ to provide a posterior "confidence level" of an abnormal segment $[i, j]$ identified by a segmentation procedure, whose endpoints are specified up to the above equivalence. Since $k \le \lfloor (j-i)/2 \rfloor$, these events $C_{i'j'}$ are disjoint and therefore

$$\sum_{(i',j'):|i-i'|\le k,|j-j'|\le k} P(C_{i'j'}|\mathcal{Y}_n) = P(\bigcup_{(i',j'):|i-i'|\le k,|j-j'|\le k} C_{i'j'}|\mathcal{Y}_n). \tag{15}$$

Whereas C_{ij} relates to the property that all locations in the segment $[i,j]$ have the same copy number $\neq 2$ and that $\theta_i \neq \theta_{i-1}$ and $\theta_j \neq \theta_{j+1}$, one may want to make inferences on other properties of a genomic segment that is not identified by a segmentation procedure. A fundamental entity from which these inferences on genomic regions can be derived is the posterior distribution of the parameter sequence $\{\theta_t : 1 \le t \le n\}$ given $\mathcal{Y}_n$. It is shown in Appendix B that this posterior distribution is that of an inhomogeneous Markov chain whose initial distribution is π and whose transition probabilities are given by

$$\theta_t|\theta_{t-1}, \mathcal{Y}_n \sim a_t\delta_0 + c_t\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\delta_{\theta_{t-1}} + \sum_{j=t}^{n} b_{jt}N(\mu_{t,j}, v_{t,j}), \tag{16}$$

in which $a_t = a_t^*/B_t$, $c_t = c_t^*/B_t$, $b_{jt} = b_{jt}^*/B_t$ and

$$\begin{aligned}
B_t &= a_t^* + c_t^*\mathbf{1}_{\{\theta_{t-1}\neq 0\}} + \sum_{j=t}^{n} b_{jt}^*,\\
a_t^* &= \phi_{0,\sigma^2}(y_t)\Big[(1-p)\mathbf{1}_{\{\theta_{t-1}=0\}} + c\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\Big]\Big[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}\Big]\Big/c,\\
c_t^* &= a\phi_{\theta_{t-1},\sigma^2}(y_t)\Big\{\big(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1}\big) + a\sum_{j=t+1}^{n}\widetilde{q}_{j,t+1}\frac{\phi_{\mu_{t+1,j},v_{t+1,j}}(\theta_{t-1})}{\phi_{\mu,v}(\theta_{t-1})}\Big\}\Big/p,\\
b_{jt}^* &= \Big[p\mathbf{1}_{\{\theta_{t-1}=0\}} + b\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\Big]\phi_{0,\sigma^2}(y_t)\cdot\begin{cases}(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1})\psi/(p\psi_{t,t}), & j=t,\\ a\widetilde{q}_{j,t+1}\psi_{t+1,j}/(p\psi_{t,j}), & j>t,\end{cases}
\end{aligned}$$

using the same notation as that in (12).

Making use of the transition probabilities (16) of the inhomogeneous Markov chain, we can use the following recursive procedure to sample from the joint posterior distribution of the parameters $\theta_{t_1}, \ldots, \theta_{t_2}$ (given $\mathcal{Y}_n$) in a segment $[t_1, t_2]$. Initialize at location $t = t_1$ by sampling θ_t from the distribution (10) for $\theta_t|\mathcal{Y}_n$. At location $t_1 < t \le t_2$, if $\theta_{t-1} = 0$, sample θ_t from $N(\mu_{t,j}, v_{t,j})$ with probability b_{jt} for $t \le j \le n$, and set $\theta_t = 0$ with probability a_t. If $\theta_{t-1} \neq 0$, set $\theta_t = \theta_{t-1}$ with probability c_t, set $\theta_t = 0$ with probability a_t, and sample from $N(\mu_{t,j}, v_{t,j})$ with probability b_{jt} for $t \le j \le n$. The posterior distribution of $(\theta_{t_1}, \ldots, \theta_{t_2})$ given $\mathcal{Y}_n$ evaluated from a large number of simulated trajectoies sampled from it can be used for statistical inference on the segment $[t_1, t_2]$. Some specific applications are given in Section 4.2, in which the special case $t_1 = 1$ and $t_2 = n$ covers an entire geonome.

2.5 BCMIX approximations

Although the Bayes filter (5) uses a recursive updating formula (6) for the weights $q_{i,t}(1 \le i \le t)$, the number of weights increases with t, resulting in unbounded computational complexity and memory requirements in estimating θ_t as t keeps increasing. A simple idea to maintain bounded complexity is to keep only a fixed number k of weights at every stage t (which is tantamount to setting the other weights to be 0). Following Lai et al. (2005) who consider the case without a baseline state, we keep the most recent m weights $q_{i,t}$ (with $t - m < i \le t$) and the largest $k - m$ of the remaining weights, where $1 \le m < k$. Specifically, the updating formula (6) for the weights $q_{i,t}$ is modified as follows to obtain a bounded complexity mixture (BCMIX) approximation. Let $\mathcal{K}_{t-1}$ denote the set of indices i for which $q_{i,t-1}$ is kept at stage $t-1$; thus $\mathcal{K}_{t-1} \supset \{t-1, , \cdots, t-m\}$. At stage t, define $q^*_{i,t}$ by (6) for $i \in \{t\} \cup \mathcal{K}_{t-1}$ and let i_t be the index not belonging to $\{t, t-1, \cdots, t-m+1\}$ such that

$$q^*_{i_t,t} = \min\{q^*_{i,t} : j \in \mathcal{K}_{t-1} \quad \text{and} \quad j \le t-m\}, \tag{17}$$

choosing i_t to be the one farthest from t if the minimizing set in (17) has more than one element. Define $\mathcal{K}_t = \{t\} \cup (\mathcal{K}_{t-1} - \{i_t\})$ and let

$$p_t = p^*_t \Big/ \Big(p^*_t + \sum_{j \in \{t\} \cup \mathcal{K}_{t-1}} q^*_{j,t}\Big),$$
$$q_{i,t} = \Big(q^*_{i,t} \Big/ \sum_{j \in \mathcal{K}_t} q^*_{j,t}\Big)\Big(\sum_{j \in \{t\} \cup \mathcal{K}_{t-1}} q^*_{j,t} \Big/ \big[p^*_t + \sum_{j \in \{t\} \cup \mathcal{K}_{t-1}} q^*_{j,t}\big]\Big), i \in \mathcal{K}_t.$$

For the smoothing estimate $E(\theta_t|\mathcal{Y}_n)$ and its associated posterior distribution, we can construct BCMIX approximations by combining forward and backward BCMIX filters, which have index sets $\mathcal{K}_t$ for the forward filter and $\widetilde{\mathcal{K}}_{t+1}$ for the backward filter at stage t. The BCMIX approximation $\alpha_t \delta_0 + \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}} \beta_{ijt} N(\mu_{ij}, v_{ij})$ to (10) is defined by

$$\alpha_t = \alpha^*_t / A_t, \qquad \beta_{ijt} = \beta^*_{ijt} / A_t, \qquad A_t = \alpha^*_t + \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \widetilde{\mathcal{K}}_{t+1}} \beta^*_{ijt},$$
$$\alpha^*_t = p_t[(1-p)\widetilde{p}_{t+1} + c\widetilde{q}_{t+1}]/c,$$
$$\beta^*_{ijt} = \begin{cases} q_{i,t}(p\widetilde{p}_{t+1} + b\widetilde{q}_{t+1})/p, & i \in \mathcal{K}_t, j = t, \\ aq_{i,t}\widetilde{q}_{j,t+1}\psi_{i,t}\psi_{t+1,j}/(p\psi\psi_{i,j}), & i \in \mathcal{K}_t, j \in \widetilde{\mathcal{K}}_{t+1}. \end{cases}$$

3. ESTIMATION OF HYPERPARAMETERS AND IMPLEMENTATION

It is shown in Appendix A that the conditional density function of y_t given $\mathcal{Y}_{t-1}$ is

$$f(y_t|\mathcal{Y}_{t-1}) = (p_t^* + \sum_{i=1}^{t} q_{i,t}^*)\phi_{0,\sigma^2}(y_t), \tag{18}$$

where p_t^* and q_{it}^* are given by (6) and are functions of the hyperparameter vector $\Phi = (p, b, c, \mu, v, \sigma^2)$. Given Φ and the observed data $\mathcal{Y}_n$, the log likelihood function is

$$l(\Phi) = \sum_{t=1}^{n} \log f(y_t|\mathcal{Y}_{t-1}) = \sum_{t=1}^{n} \log \Big\{ (p_t^* + \sum_{i=1}^{t} q_{i,t}^*)\phi_{0,\sigma^2}(y_t) \Big\}, \tag{19}$$

in which $f(\cdot|\cdot)$ denotes conditional density function. Maximizing (19) over Φ yields the maximum likelihood estimate $\widehat{\Phi}$.

Since Φ is a 6-dimensional vector and the functions $p_t^*(\Phi)$ and $q_{i,t}^*(\Phi)$ have to be computed recursively for $1 \le t \le n$, direct maximization of (19) may be computationally expensive due to the curse of dimensionality. An alternative approach is to use the EM algorithm which exploits the much simpler structure of the log likelihood $l_c(\Phi)$ of the complete data $\{(y_t, \theta_t), 1 \le t \le n\}$:

$$\begin{aligned} l_c(\Phi) = {} & -\frac{1}{2}\sum_{t=1}^{n}\Big\{\frac{(y_t-\theta_t)^2}{\sigma^2} + \log(2\pi\sigma^2)\Big\} - \frac{1}{2}\sum_{t=1}^{n}\Big\{\frac{(\theta_t-\mu)^2}{v} + \log(2\pi v)\Big\}\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}} \\ & + \textstyle\sum_{t=1}^{n}\Big\{[\log(1-p)]\mathbf{1}_{\{\theta_t=\theta_{t-1}=0\}} + (\log p)\mathbf{1}_{\{\theta_t\neq\theta_{t-1}=0\}}\Big\} \\ & + \textstyle\sum_{t=}^{n}\Big\{[\log(1-b-c)]\mathbf{1}_{\{\theta_t=\theta_{t-1}\neq 0\}} + (\log c)\mathbf{1}_{\{\theta_t=0\neq\theta_{t-1}\}} + (\log b)\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\neq 0\}}\Big\}. \end{aligned} \tag{20}$$

Since $l_c(\Phi)$ decomposes into normal and multinomial components, the E-step of the EM algorithm involves $E\big((\theta_t-\mu)^2|\mathcal{Y}_n\big)$, $E\big((\theta_t-y_t)^2|\mathcal{Y}_n\big)$ and the conditional probabilities

$$P(\theta_t = 0 = \theta_{t-1}|\mathcal{Y}_n) = \frac{(1-p)p_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \quad P(\theta_t = 0 \neq \theta_{t-1}|\mathcal{Y}_n) = \frac{cq_{t-1}\alpha_t}{(1-p)p_{t-1} + cq_{t-1}}, \tag{21}$$

$$P(\theta_t \neq \theta_{t-1} = 0|\mathcal{Y}_n) = c\widetilde{q}_t\alpha_{t-1}\Big/\{(1-p)\widetilde{p}_t + c\widetilde{q}_t\}, \tag{22}$$

$$P(0 \neq \theta_t \neq \theta_{t-1} \neq 0|\mathcal{Y}_n) = (\sum_{j=t}^{n}\beta_{tjt})bq_{t-1}\Big/\{bq_{t-1} + pp_{t-1}\}, \tag{23}$$

together with $P(\theta_t = \theta_{t-1} \neq 0|\mathcal{Y}_n)$, which is determined by the property that those five conditional probability have to sum up to 1. The proof of (21) – (23) is given in Appendix A.

In view of (20), the M-step of the EM algorithm involves the closed-form updating formulas

$$\begin{aligned}
1-\widehat{p}_{\text{new}} &= \left[\Sigma_1^n P(\theta_t=\theta_{t-1}=0|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right]\Big/\left[\Sigma_1^n P(\theta_{t-1}=0|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right],\\
\widehat{a}_{\text{new}} &= \left[\Sigma_1^n P(\theta_t=\theta_{t-1}\neq 0|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right]\Big/\left[\Sigma_1^n P(\theta_{t-1}\neq 0|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right],\\
\widehat{c}_{\text{new}} &= \left[\Sigma_1^n P(\theta_t=0\neq\theta_{t-1}|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right]\Big/\left[\Sigma_1^n P(\theta_{t-1}\neq 0|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right], \widehat{b}_{\text{new}}=1-\widehat{a}_{\text{new}}-\widehat{c}_{\text{new}},\\
\widehat{\mu}_{\text{new}} &= \left[\Sigma_1^n E(\theta_t\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}}|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}}\right]\Big/\left[\Sigma_1^n P(0\neq\theta_t\neq\theta_{t-1}|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right],\\
\widehat{v}_{\text{new}} &= \left[\Sigma_1^n E\{(\theta_t-\widehat{\mu}_{\text{old}})^2\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}}|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}}\}\right]\Big/\left[\Sigma_1^n P(0\neq\theta_t\neq\theta_{t-1}|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right],\\
\widehat{\sigma}^2_{\text{new}} &= \Sigma_{t=1}^n\left[E((y_t-\theta_t)^2|\mathcal{Y}_n,\widehat{\Phi}_{\text{old}})\right]\Big/n.
\end{aligned}\tag{24}$$

It is shown in Appendix A that

$$\begin{aligned}
E(\theta_t\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}}|\mathcal{Y}_n) &= \sum_{t\leq j\leq n}\beta_{tjt}\mu_{t,j},\\
E((\theta_t-\mu)^2\mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}}|\mathcal{Y}_n) &= \sum_{t\leq j\leq n}\beta_{tjt}(\mu_{t,j}^2+v_{t,j}-2\mu\mu_{t,j}+\mu^2),
\end{aligned}\tag{25}$$

which can be applied to compute $\widehat{\mu}_{\text{new}}$ and $\widehat{v}_{\text{new}}$ in (24). The iterative scheme (24) is carried out until convergence or until some prescribed upper bound on the number of iterations is reached.

To speed up the computations involved in the preceding EM algorithm, one can use the BCMIX approximations in Section 2.5 instead of the full recursions to determine $q_{i,t}$, $\widetilde{q}_{j,t}$, etc. Moreover, one can accelerate the EM algorithm by using a hybrid approach that combines EM with some classical optimization technique, e.g., quasi-Newton methods as in Lange (1995). Applications to array-CGH data have shown that the EM estimates of μ, v, σ^2 and b typically converge quite fast. This suggests switching, after these parameter estimates stabilize, from the EM algorithm to global search for the optimizing p and c, which are particularly important as they represent relative frequencies of departures from, and returns to, the baseline state. The global search in this hybrid procedure uses (19) as a function only of p and c, with the other parameter estimates fixed at the time of switch from EM.

4. APPLICATIONS TO REAL DATA SETS

We now illustrate our method and examine its performance on several real array CGH data sets. In Section 4.1, we consider the BAC array hybridizations of the Coriel cell lines from Snijders et al. (2001), which are taken from 15 primary breast tumors. Out of these 15 cell lines, we use the 9 which have known karyotype data. These cell lines have been used extensively for validation purposes in numerous methodological studies, since the true karyotypes are known. However, because the chromosomal aberration profile in this data set is relatively simple, most methods give similar segmentations and good estimates of the true signals.

In Section 4.2 we use the BAC array hybridization of the BT474 cell line, taken from Snijders et al. (2003), to illustrate some of the inferential procedures that are possible with our method. The cell line is taken from tumors with more complicated aberration profiles than those of the Coriel cell lines, as is evident from the array-CGH plots in Figures 2 and 3. These more challenging data sets do not reveal obvious segmentations, and thus a framework for inference becomes crucial.

4.1 Coriel breast cancer data

The 9 cell lines that we used in our study are: GM13330, GM13031, GM07081, GM05296, GM03563, GM03134, GM01750, GM01535, and GM01524. From the karyotype information, we can estimate the true signal level θ_i as follows: If a probe i lies in a region where the karyotype is 2, we set $\theta_i = 0$. Otherwise, the probe lies in a changed region for which the boundaries are known, and we set θ_i to be the mean of all probes in that region. Note that we need to estimate the true signal from the data even when the true copy number is known, because of the nonlinear relationship between measured fluorescence ratio and copy number that may differ slightly across data sets (Pinkel and Albertson (2005)).

To estimate the hyperparameters of our model, which will be denoted by SCP (stochastic change-point model) in the sequel, we note that since the Coriel cell lines have a relatively small number of aberrant segments, the probability p of jumping from the zero state to a nonzero state should be small and the probability b of jumping from a nonzero state to another nonzero state should be even substantially smaller. Therefore, we set $b = 0$, ruling out jumps from an infrequent nonzero state to another nonzero state, and use the hybrid procedure described in the last paragraph of Section 3, limiting the global search to $10^{-4} < p < 0.005$ and $10^{-4} < c < 0.05$.

Figure 1 plots the posterior means $E(\theta_t|\mathcal{Y}_n), 1 \leq t \leq n$, for the nine cell lines. Although we have also computed the 2.5% and 97.5% quantiles of the posterior distribution of θ_t given $\mathcal{Y}_n$, they are too close to $E(\theta_t|\mathcal{Y}_n)$ to be plotted distinctly in the figure. The qth quantile of the posterior distribution of θ_t given $\mathcal{Y}_n$, which is a mixture of normal distributions and a point mass at 0 given by (10), is obtained by solving the equation

$$\alpha_t \mathbf{1}_{\{x \geq 0\}} + \sum_{1 \leq i \leq t \leq j \leq n} \beta_{ijt} \int_{-\infty}^{x} \phi_{\mu_{i,j}, v_{i,j}}(z) dz = q.$$

We also calculate the L_1 distance $\sum_{t=1}^{n} |\widehat{\theta}_t - \theta_t|$ for the estimated signal produced by a given method. The methods that we choose for comparison are the HMM-based algorithm (HMM) of Fridlyand et al. (2004) and the CBS algorithm of Olshen et al. (2004). For HMM, we start with 5 states in the hidden Markov model and apply the state merging step with a merging threshold of 0.25 as recommended in Fridlyand et al. (2004). For the CBS algorithm, we used the default parameters for the dnacopy software in Bioconductor. Table 1, which lists the L_1 distances for each cell line and method, shows that by assuming a known baseline, our model provides a better fit to these data than previous methods.

INSERT TABLE 1 AND FIGURE 1 ABOUT HERE

4.2 Breast cancer cell line BT474

For the cell line BT474, we use the EM algorithm in Section 3 to estimate the hyperparameters. With initial values of (p, a, b) at $(0.05, 0.995, 0.0025)$ and (μ, v, σ^2) at $(0.065, 0.087, 0.020)$, the EM algorithm stops after 7 iterations according to a convergence criterion, yielding $\widehat{p} = 0.7196, \widehat{a} = 0.9147, \widehat{c} = 0.0662, \widehat{\mu} = 0.3063, \widehat{v} = 0.5668, \widehat{\sigma}^2 = 0.0152$.

The top plots of Figures 2 and 3 show the array CGH profile for chromosomes 17 and 20, with the estimated mean levels and the 2.5% and 97.5% quantiles of the posterior distribution of true signals computed by our model. Because of the complexity of this cell line, previous methods disagree widely on the correct segmentation, as can be seen by comparing the segmentations given by CBS, HMM, and our method on these two chromosomes. Because of the complexity of the BT474 profile, it is important for a statistical method to be able to assess the confidence in a particular segmentation.

A striking difference between our method and CBS and HMM is that our method can capture sawtooth patterns such as those found in the q arm of chromosomes 17 and 20 (see Figure 2, 50-70 Mb region, and Figure 3, 40-60 Mb region). The sawtooth patterns are smoothed out by CBS and HMM, primarily because these methods aim to segment the data, while our method estimates the true signal without imposing a segmentation. These sawtooth patterns are very frequently seen in highly rearranged breast tumors, and are generally recognized as a real phenomena and not system noise. They have generated much biological interest because they may provide clues to the specific path that cells took to acquire them. For example, Hicks et al. (2005) discuss the possible biological mechanisms that generated such patterns, which they also found by ROMA CGH.

Through our model, we provide a framework for multiple levels of inference. At the genome level, it is often of interest to rank the detected chromosomal aberrations by the confidence that it is a true aberration. This allows a prioritization of downstream studies and experiments so that they can be targeted to genomic regions of higher statistical significance. We make use of the formulas in Section 2.4 to calculate $P(C_{ij}|\mathcal{Y}_n)$, which is the posterior probability of an aberration with the left boundary i and the right boundary j, for all $i < j$ on the same chromosome arm. In Table 2, the aberrations detected in BT474 are ranked by $P(C_{ij}|\mathcal{Y}_n)$. Comparing Table 2 and Figures 2 and 3, we see that the aberrations that are visually evident in chromosomes 17 and 20 are also ranked high in the table. For example, the focal aberration on chromosome 17, which contain the well studied ERBB2 amplicon, is at the top of the ranking with a probability of 1. Other segments that top the list, mostly amplicons on chromosomes 11, 17, and 20, are well known, as they have been identified by previous studies and in other breast cancer cell lines (e.g. Pollack et al. (1999), Pinkel et al. (1998)). In comparison, segmental duplications and deletions, such as those on chromosomes 9 and X, are ranked lower than the focal aberrations in the list. This is desirable if biologists wish to zoom in on a narrow region that has undergone strong selective pressure.

INSERT TABLE 2 AND FIGURES 2, 3 ABOUT HERE

Finer scale confidence assessments targeted at a specific genome region also arises naturally from our framework. We illustrate this with the data from Chromosome 20 in BT474 (Figure 3). This region of the genome has been under scrutiny in many cancer studies, partly due to the fact that it contains several candidate oncogenes (e.g., AIB1, TFAP2C, and STK15). Figure 3 shows several distinct aberrations in this region for BT474. However, it may be of interest to assess the relative likelihood of a sawtooth pattern consisting of at least one spike within the 40-50 Mb region, as compared to a flat segmentation given by HMM and CBS that assigns a uniform mean to this region. It is biologically meaningful and critical to make these distinctions, because differences in such minute details of the segmentation can point to differences in the history of progression of the tumor, as well as different arrangements of the segments in the genome. The posterior confidence level (15) , with $k^* = 2$, for a single segment in 40-50 MB region proposed by the HMM procedure is

0.000, whereas

$$P\{[i,j] \text{ contains a sub-segment } [i',j'] \text{ such that } \theta_{i'} = \ldots = \theta_{j'} \neq 0,$$
$$\theta_{i'} \neq \theta_{i'-1}, \theta_{j'} \neq \theta_{j'+1}, i'-1 \geq i, j'+1 \leq j\} \geq \max_{i<i'\leq j'<j} P(C_{i'j'}|\mathcal{Y}_n),$$

which exceeds 0.997, 0.999, and 1 respectively for the segments $[i',j']$ = A, B, C within the 40-50 Mb region in Figure 3. Thus, the probability of a spike within the 40-50 Mb region far outweighs the probability of a uniform mean level in that region.

The total number of changes in chromosome copy number is a useful indicator of genome instability, and has been shown to be correlated with many factors such as disease stage, degrees of aneuploidy, and tumor heterogeneity (Fabarius et al. (2003), Pinkel and Albertson (2005)). Existing segmentation algorithms are able to provide an estimate of the number of change-points through a hard segmentation. However, for a complex aberration profile such as BT474, a confidence bound for the number of change-points can be much more informative. For the BT474 cell line, the modified BIC (Zhang and Siegmund, 2006) peaks at 30 change-points. With HMMs, 103 change-points are found if the state-merging step proposed by Fridlyand et al. (2004) is not taken, after the merging of states, the complete procedure from Fridlyand et al. (2004) reports 69 change-points for this data series.

We use the Monte Carlo procedure described in Section 2.4 to construct confidence bounds for the total number of change-points in the genome. Define

$$\kappa = \sum_{i=1}^{n-1} \mathbf{1}_{\{\theta_{t+1} \neq 0, |\theta_t - \theta_{t+1}| > \delta\}}, \tag{26}$$

in which the threshold δ is used to exclude negligibly small jumps that can occur in our Gaussian jump model. The posterior distribution of κ given $\mathcal{Y}_n$ is computed by Monte Carlo, using simulated sequences generated from the fitted model by using (16). Figure 4 shows the histogram of κ, in which we set $\delta = \sqrt{v}$ in (26), calculated for 5000 simulated sequences; recall that our model assumes Gaussian jumps with variance v. The mean and 95% confidence intervals of κ based on these 5000 simulations are 60.24 and $[60.14, 60.35]$.

INSERT FIGURE 4 ABOUT HERE

5. SIMULATIONS

We also tested our method using simulation studies, in which the true signal is known and thus various measures of accuracy can be computed. The simulation data are generated from $y_t = \theta_t + \epsilon_t, \quad 1 \leq t \leq n$, where ϵ_t are i.i.d. $N(0, \sigma^2)$ and one of the following three models is used to generate θ_t: (a) the HMM model of Fridlyand et al. (2004), in which $\{\theta_t\}$ is a finite-state Markov chain; (b) the stochastic change-point model (SCP) described in Section 2; (c) the frequentist model considered by Olshen et al. (2004), in which θ_t is a fixed piecewise constant function.

The parameters of the above models are determined by fitting the model to the BT474 breast cancer cell line. For the HMM model, the parameters consist of the state means $\{\theta_i\}_{i=1}^{K}$, the $K \times K$ state transition matrix, and the noise variance. The number of states $K = 7$ was chosen by AIC. For the stochastic change-point model, the hyperparameters are p, a, b, c, v, and σ defined in Section 2, with values given in Section 4.2. For the frequentist model, the θ_t are the estimated mean levels for BT474 using the CBS algorithm. We used $n = 2056$ for all three models, which is the same length as the complete BT474 data set without missing values. Figure 5 shows an example of a simulation data series generated from each of the three models. We simulated 100 data series from each model for this study, and for our method, we first run the EM algorithm with 20 iterations to estimate the hyperparameters and then compute the posterior means for each simulated sequence.

INSERT TABLES 3,4 AND FIGURE 5 ABOUT HERE

Table 3 gives the L_1 distances of the estimated and true means of HMM, CBS, and our method on each of the 3 simulation models. From these results, we see that our method outperforms CBS and HMM for both the stochastic change-point and frequentist models. For the HMM model, our method loses slightly to CBS, while being better than HMM. Since these simulation data are generated to resemble the level of difficulty in the BT474 cell line, they show that our method can still perform well for the more complex profiles.

We next consider the performance of our classification procedure described in the second paragraph of Section 2.4. We choose the threshold $w = 2\widehat{\sigma}$ for each sequence, in which $\widehat{\sigma}$ is the estimate of σ determined by our method. For the data simulated by HMM and frequentist models, the threshold ranges from 0.35 to 0.40; for the data simulated by our model, the threshold ranges from 0.20 to 0.25. These thresholds are quite small compared to the signal sequence, and are therefore fair parameters to use. Table 4 lists the false positive and false negative rates (in classifying no change versus a gain or a loss) for the three different methods in each of the three simulation models. For CBS and HMM, only a hard segmentation of the data is produced, and thus we assign a probe to a changed state if the absolute value of its estimated mean is above the threshold $w = 2\widehat{\sigma}$.

6. DISCUSSION

We have developed a stochastic change-point model for inference on array-CGH data sets. The model allows exact computation, through recursive formulas given in Section 2.2, of the parameters of the posterior distribution of the signal $\{\theta_t : 1 \leq t \leq n\}$. From the posterior distribution of the signal given the observations, a segmentation of the data and a classification of the probes can be obtained. A Monte Carlo method for sampling from the joint posterior distribution of $\{\theta_t\}$ is given in Section 2.4, which allows inference on almost any quantity of interest to the biologist. An approximation to the exact explicit formulas, using the BCMIX method, allows our method to be executed almost instantaneously for BAC arrays.

In Section 4, we have used the Coriel and BT474 breast cancer data sets to illustrate the application of our method. In particular, we have focused on illustrating the types of inference that are possible with our method. For example, in Section 4.1 we give a method for calculating pointwise marginal confidence intervals for the estimated signal. In Section 4.2, we give a ranking of the most "interesting" aberrations in the complex data set from BT474. The aberrations that top this list are those found by most previous methods, while those that are further down the list have lent to disagreements. Instead of producing a hard segmentation, our method

the biologist the option to investigate it further.

A departure of our model from most previous models is the assumption of a baseline state, which yields a natural classification of genomic regions into "amplified", "deleted", and "normal" states. The model proposed in Engler et al. (2006) also gives such a classification rule. However, it does not provide explicit formulas for the posterior probability of the states at each time point conditioned on the entire data series for Bayesian inference as it relies on "smoothing" via three-probe windows. To circumvent the computational complexity for their model, Engler et al. (2006) have used pseudo-likelihood in lieu of the actual likelihood. In contrast, our model yields explicit formulas for Bayesian analysis and maximum likelihood estimation of the hyperparameters.

We chose to conduct our data analysis at the genome level, rather than at the chromosome level, because the inter-chromosome difference in baseline signal level for BT474 and the Coriel cell lines is negligible for our model. Also, pooling data across chromosomes allows a more accurate estimate of the hyperparameters. The fact that multi-chromosome analysis improves sensitivity has also been shown in Engler et al. (2006). Finally, genome scale analysis allows the detection of copy number changes involving entire chromosome arms, which would be missed in chromosome-level analyses for which no actual changes-points exist.

The data sets used in Section 4 are from BAC arrays, which use bacterial artificial chromosomes as genomic targets. Other platforms for array-CGH have been designed, such as cDNA arrays (Pollack et al., 2002), which measure copy numbers only at transcribed regions of the genome. Wen et al. (2006) have pointed out the need for incorporating possible changes in σ with θ_t in the analysis of cDNA array-CGH data, and have developed for such analysis a Bayesian regression model that relies on Markov chain Monte Carlo for posterior analysis. By making use of the ideas of Lai et al. (2005) to model changes in both the error variance and the regression parameters, it should be possible to extend the methods and results of the present paper to accommodate changes in σ with θ_i and to incorporate possible correlations among the observations. Extending Lai et al.'s (2005) approach to multivariate regression should also enable us to combine both cDNA and BAC array-CGH data in estimating the underlying signal and other inferential tasks. This is a topic for future research.

ACKNOWLEDGMENTS

T.L. Lai's research was supported by the National Science Foundation grant DMS-0305749 and the National Institutes of Health grant CA088890.

APPENDIX A

Proof of (6), (12), (14), (18), (21), (22), (23) *and* (25)

To prove (6), we make use of

$$\phi_{\mu_1,v_1}(\theta)\phi_{\mu_2,v_2}(\theta) = \phi_{\bar{\mu},\bar{v}}(\theta)\sqrt{\frac{\bar{v}}{v_1 v_2}}\ \exp\left\{\frac{1}{2}\left[\frac{\bar{\mu}^2}{\bar{v}} - \frac{\mu_1^2}{v_1} - \frac{\mu_2^2}{v_2}\right]\right\} = \frac{\phi_{\mu_1}(0)\phi_{\mu_2}(0)}{\phi_{\bar{\mu},\bar{v}}(0)}\ \phi_{\bar{\mu},\bar{v}}(\theta), \quad \text{(A.1)}$$

where $\bar{v} = (v_1^{-1} + v_2^{-1})^{-1}$ and $\bar{\mu} = \bar{v}(\mu_1/v_1 + \mu_2/v_2)$, as can be shown by completing the squares. From (5), it follows that

$$P(\theta_t = 0|\mathcal{Y}_t) \propto \{(1-p)p_{t-1} + cq_{t-1}\}\phi_{0,\sigma^2}(y_t), \quad \text{(A.2)}$$

and the density function of the absolutely continuous component of θ_t is proportional to

$$(pp_{t-1} + bq_{t-1})\phi_{\mu,v}(\theta)\phi_{\theta,\sigma^2}(y_t) + a\sum_{i=1}^{t-1} q_{i,t-1}\phi_{\mu_{i,t-1},v_{i,t-1}}(\theta)\phi_{\theta,\sigma^2}(y_t), \quad \text{(A.3)}$$

with the constant of proportionality equal to the reciprocal of the conditional density function of y_t given $\mathcal{Y}_{t-1}$. From (A.1), it follows that

$$\begin{aligned}\phi_{\mu,v}(\theta)\phi_{\theta,\sigma^2}(y_t) = \phi_{\mu,v}(\theta)\phi_{y_t,\sigma^2}(\theta) &= \phi_{\mu_{t,t},v_{t,t}}(\theta)\{\phi_{\mu,v}(0)\phi_{y_t,\sigma^2}(0)/\phi_{\mu_{t,t},v_{t,t}}(0)\} \\ &= \phi_{\mu_{t,t},v_{t,t}}(\theta)(\psi/\psi_{t,t})\phi_{0,\sigma^2}(y_t),\end{aligned} \quad \text{(A.4)}$$

$$\phi_{\mu_{i,t-1},v_{i,t-1}}(\theta)\phi_{\theta,\sigma^2}(y_t) = \phi_{\mu_{i,t},v_{i,t}}(\theta)(\psi_{i,t-1}/\psi_{i,t})\phi_{0,\sigma^2}(y_t), \tag{A.5}$$

Putting (A.4) and (A.5) into (A.2) and (A.3) then yields (6).

The formula for α_t in (12) has already been proved in (11). Let $f_t(\cdot|\mathcal{Y}_n)$, $f_t(\cdot|\mathcal{Y}_t)$ and $f_t(\cdot|\mathcal{Y}_{t+1,n})$ denote the density functions of the absolutely continuous components of θ_t given $\mathcal{Y}_n$, $\mathcal{Y}_t$, $\mathcal{Y}_{t+1,n}$, respectively, and let $\dot{\pi}$ denote the density function of the absolutely continuous component of π. Then applying Bayes' theorem as in (11),

$$f_t(\theta|\mathcal{Y}_n) \propto f_t(\theta|\mathcal{Y}_t)f_t(\theta|\mathcal{Y}_{t+1,n})/\dot{\pi}(\theta). \tag{A.6}$$

The constant of proportionality in (11) and (A.6) is $g(\mathcal{Y}_t)g_*(\mathcal{Y}_{t+1,n})/g^*(\mathcal{Y}_n)$, where g, g_* and g^* denote the respective joint density functions. As shown in Section 2.1,

$$\dot{\pi}(\theta) = \phi_{\mu,v}(\theta)p/(p+c). \tag{A.7}$$

Simple algebra that involves completing squares as in (A.1) can be used to show that if $v^{-1} < v_1^{-1} + v_2^{-1}$, $\tilde{v} = (v_1^{-1} + v_2^{-1} - v^{-1})^{-1}$ and $\tilde{\mu} = \tilde{v}(\mu_1/v_1 + \mu_2/v_2 - \mu/v)$, then

$$\frac{\phi_{\mu_1,v_1}(\theta)\phi_{\mu_2,v_2}(\theta)}{\phi_{\mu,v}(\theta)} = \phi_{\tilde{\mu},\tilde{v}}(\theta)\sqrt{\frac{v\tilde{v}}{v_1v_2}}\exp\left\{\frac{1}{2}\left[\frac{\tilde{\mu}^2}{\tilde{v}} + \frac{\mu^2}{v} - \frac{\mu_1^2}{v_1} - \frac{\mu_2^2}{v_2}\right]\right\} = \frac{\phi_{\mu_1,v_1}(0)\phi_{\mu_2,v_2}(0)}{\phi_{\tilde{\mu},\tilde{v}}(0)\phi_{\mu,v}(0)}\phi_{\tilde{\mu},\tilde{v}}(\theta). \tag{A.8}$$

Combining (A.6), (A.7) with (5) and (9), and making use of (A.8) in the case $t < j$, we obtain β^*_{ijt} in (12).

Let $\widetilde{K}_t = \min\{s \geq t : \theta_t = \cdots = \theta_s \neq \theta_{s+1}\}$ be the counterpart of K_t (defined at the beginning of Section 2.2) for the time-reversed chain. In view of the preceding argument and (10),

$$\beta_{ijt} = P\{\theta_{K_t} \neq 0, K_t = i, \widetilde{K}_t = j|\mathcal{Y}_n\}. \tag{A.9}$$

From the definitions of K_t and $\widetilde{K}_t$, it follows that the event in (A.9) is the same as C_{ij} defined in (14). Hence (14) holds.

To prove (18), note that

$$P(\theta_t = 0, y_t \in dt|\mathcal{Y}_{t-1}) = P(\theta_t = 0|\mathcal{Y}_{t-1})\phi_{0,\sigma^2}(y_t)dt = p_t^*\phi_{0,\sigma^2}(y_t)dt, \tag{A.10}$$

$$\begin{aligned}
P(\theta_t \neq 0, y_t \in dt|\mathcal{Y}_{t-1}) &= \int f(\theta_t = \theta \neq 0|\mathcal{Y}_{t-1})\phi_{\theta,\sigma^2}(y_t)d\theta dt \\
&= \int \left\{(pp_{t-1} + bq_{t-1})\phi_{\mu,v}(\theta) + a\sum_{i=1}^{t-1} q_{i,t-1}\phi_{\mu_{i,t-1},v_{i,t-1}}(\theta)\right\}\phi_{\theta,\sigma^2}(y_t)d\theta dt \qquad \text{(A.11)} \\
&= \sum_{i=1}^{t} q^*_{i,t}\phi_{0,\sigma^2}(y_t)dt,
\end{aligned}$$

by (A.4), (A.5) and (6). From (A.10) and (A.11), (18) follows.

To prove (21), we modify (A.2) as

$$P(\theta_t = 0, \theta_{t-1} = 0|\mathcal{Y}_t) \propto (1-p)p_{t-1}\phi_{0,\sigma^2}(y_t), \qquad P(\theta_t = 0, \theta_{t-1} \neq 0|\mathcal{Y}_t) \propto cq_{t-1}\phi_{0,\sigma^2}(y_t).$$

Combining this with $P(\theta_t = 0|\mathcal{Y}_{t+1,n})/\pi(0)$ as in (11) yields (21). A similar argument applied to the time reverse chain yield (22). To prove (23), we use a similar argument to obtain

$$P(0 = \theta_{t-1} \neq \theta_t \in d\theta|\mathcal{Y}_t) \propto pp_{t-1}\phi_{\mu,v}(\theta)\phi_{\theta,\sigma^2}(y_t)d\theta,$$
$$P(0 \neq \theta_{t-1} \neq \theta_t \in d\theta|\mathcal{Y}_t) \propto bq_{t-1}\phi_{\mu,v}(\theta)\phi_{\theta,\sigma^2}(y_t)d\theta.$$

We obtain (23) by combining this with $f_t(\theta|\mathcal{Y}_{t+1,n})/\dot{\pi}(\theta)$ and

$$P(\theta_{t-1} \neq \theta_t \neq 0|\mathcal{Y}_t) = \sum_{t\le j\le n} P(\theta_{t-1} \neq \theta_t = \cdots = \theta_j \neq 0, \theta_j \neq \theta_{j+1}|\mathcal{Y}_t) = \sum_{t\le j\le n} \beta_{tjt}.$$

The first equation in (25) follows from

$$E(\theta_t \mathbf{1}_{\{0\neq\theta_t\neq\theta_{t-1}\}}|\mathcal{Y}_n) = \sum_{t\le j\le n} E(\theta_t|K_t, \widetilde{K}_t = j, \theta_t \neq 0, \mathcal{Y}_n)P(K_t = t, \widetilde{K}_t = j, \theta_t \neq 0|\mathcal{Y}_n),$$

noting that $E(\theta_t|K_t, \widetilde{K}_t = j, \theta_t \neq 0, \mathcal{Y}_n) = \mu_{t,j}$ in view of (10). The second equation also follows similarly since $(\theta_t - \mu)^2 = \theta_t^2 - 2\mu\theta_t + \mu^2$.

APPENDIX B

Proof of (16)

Applying Bayes' theorem as in (11) yields

$$\begin{aligned} a_t =& P(\theta_t = 0|\theta_{t-1}, \mathcal{Y}_n) \propto P(\theta_t = 0|\theta_{t-1}, y_t)P(\theta_t = 0|\mathcal{Y}_{t+1,n})\Big/\pi(0) \\ &\propto \Big[(1-p)\mathbf{1}_{\{\theta_{t-1}=0\}} + c\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\Big]\phi_{0,\sigma^2}(y_t)P(\theta_t = 0|\mathcal{Y}_{t+1,n})\Big/\pi(0), \end{aligned} \tag{B.1}$$

as in (A.2). In the case $\theta_{t-1} \neq 0$ (and therefore, unlike 0, θ_{t-1} is not an atom of the stationary distribution π), a similar argument yields

$$c_t = P(\theta_t = \theta_{t-1}|\theta_{t-1}, \mathcal{Y}_n) \propto a\phi_{\theta_{t-1},\sigma^2}(y_t)f_t(\theta_{t-1}|\mathcal{Y}_{t+1,n})/\dot{\pi}(\theta_{t-1}), \tag{B.2}$$

where $f_t(\cdot|\mathcal{Y}_{t+1,n})$ and $\dot{\pi}$ are the same as in (A.1) and (A.2). Moreover, the absolutely continuous component of the conditional distribution of θ_t given $(\theta_{t-1}, \mathcal{Y}_n)$ has density function proportional to

$$\begin{aligned} &\Big[p\mathbf{1}_{\{\theta_{t-1}=0\}} + b\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\Big]\phi_{\mu,v}(\theta)\phi_{\theta,\sigma^2}(y_t)f_t(\theta|\mathcal{Y}_{t+1,n})\Big/\dot{\pi}(\theta) \\ =&\Big[p\mathbf{1}_{\{\theta_{t-1}=0\}} + b\mathbf{1}_{\{\theta_{t-1}\neq 0\}}\Big]\phi_{0,\sigma^2}(y_t)(\psi/\psi_{t,t})\phi_{\mu_{t,t},v_{t,t}}(\theta)f_t(\theta|\mathcal{Y}_{t+1,n})\Big/\dot{\pi}(\theta), \end{aligned} \tag{B.3}$$

by (A.4). We can then apply (9) and (A.8) to derive (16) from (B.1) - (B.3).

REFERENCES

Engler, D.A., Mohapatra, G., Louis, D.N., and Betensky, R.A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridications (aCGH1374). *Biostatistics*, **7**, 399-421.

Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. G., and Jain, A.N. (2004). Application of Hidden Markov Models to the analysis of the array-CGH data. Special Genomic Issue of *Journal of Multivariate Analysis*, **90**, 132-153.

Hicks, J., Muthuswamy, L., Krasnitz, A., Navin, N., Riggs, M., Grubor, V., Esposito, D., Alexander, J., Troge, J., Wigler, M., Maner, S., Lundin, P., and Zetterberg, A. (2005). High-Resolution ROMA CGH and FISH Analysis of Aneuploid and Diploid Breast Tumors. *Cold Spring Harbor Symposia on Quantitative Biology*, **70**, 51-63.

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

Lai, T.L., Liu, H., and Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica*, **15**, 279-301.

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.

Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, **5**, 1-18.

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. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel, D., and Albertson, D.G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, **37**, Suppl 11-17.

Pinkel, D., Seagraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S., Lijung, B.-M., Gray, J.W., and Albertson, D. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207-211.

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. *Nature Genetics*, **23**, 41-46.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D., and Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263-264.

Snijders, A.M., Fridlyand, J., Mans, D.A., Segraves, R., Jain, A.N., Pinkel, D., and Albertson, D.G. (2003). Shaping of tumorand drug-resistant genomes by instability and selection. *Oncogene*, **22**, 4370-4379.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., and 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.A., and Chang, I. (2006). A Bayes regression approach to array-CGH data. *Statistical Applications in Generics and Molecular Biology*, **5**, No. 1, Article 3.

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

Zhang, N. and Siegmund, D. (2006). A modified Bayes information criterion with applications to comparative genomic hybridization data. *Biometrics*, in press.

Table 1. L_1 distance between true and estimated signals for Coriel cell lines using SCP, CBS, and HMM.

Cell line	SCP	CBS	HMM
GM03563	0.00102	0.01316	0.00449
GM05296	0.01991	0.02892	0.02152
GM01750	0.00166	0.01640	0.00109
GM03134	0.00173	0.00949	0.00605
GM13330	0.00130	0.01532	0.00446
GM01535	0.01858	0.02668	0.02128
GM07081	0.00271	0.00445	0.00511
GM13031	0.00119	0.02052	0.00271
GM01524	0.00087	0.01423	0.00205

Table 2. Ranking of the aberrations in BT474 cell line by the posterior probability $P(C_{ij}|\mathcal{Y}_n)$ (truncated list). The corresponding posterior means $(j - i + 1)^{-1}\sum_{t=i}^{j} E(\theta_t|\mathcal{Y}_n)$ are also shown.

Chrom.	AugKB	Post.	Post.	Chrom.	AugKB	Post.	Post.
Number	Region	Prob.	Mean	Number	Region	Prob.	Mean
6	171756-171756	1	1.9920	4	82270-83314	0.9798	0.8046
11	133531-133531	1	2.2708	9	21709-35926	0.9739	-1.4063
11	134582-134582	1	2.0817	6	175263-200000	0.9695	-0.6204
12	108526-108526	1	1.0247	17	65396-65897	0.9653	1.8895
20	33000-33000	1	2.5181	7	18139-18139	0.9628	0.5312
20	47981-47981	1	-0.8273	23	160000-160000	0.9535	-0.7291
1	280672-280672	1	1.3569	9	17027-19086	0.9531	-0.8835
17	41969-41969	1	3.3500	12	92200-103456	0.9496	0.4371
20	47863-47863	0.9999	2.1967	5	117977-143584	0.9431	-0.5957
11	90509-90509	0.9999	0.6110	17	53252-54381	0.9406	2.1483
20	47986-48254	0.9993	2.4234	17	72037-72403	0.9336	0.8156
20	51687-52266	0.9985	2.0479	11	114497-117620	0.9304	0.9544
20	45154-45351	0.9978	1.8745	20	48941-49016	0.9230	1.7054
7	76562-76562	0.9973	1.1229	4	202817-210000	0.9125	-0.2597
20	56647-56647	0.9971	1.9751	11	84122-85101	0.9068	1.4208
20	57607-57843	0.9971	2.8618	20	49365-50902	0.9035	1.2958
8	41881-41881	0.9953	0.6502	11	82908-83238	0.9006	0.9837
20	47321-47321	0.9939	1.2477	20	32006-32330	0.8824	0.6716
9	80639-80639	0.9932	0.4848	20	65000-65000	0.8719	0.5970
9	38119-38622	0.9917	1.3142	11	94150-111973	0.8629	-0.2965
20	46643-46643	0.9894	0.4967	11	49641-49641	0.8562	0.4699
4	194428-194428	0.9882	0.9438	3	29769-29769	0.8414	-0.5360
20	52686-56017	0.9819	3.3421	17	60359-60633	0.8409	2.0651

Table 3. L_1 distances between the estimated means and true means for simulation data generated using the HMM model of Fridlyand et al. (2004), the stochastic change-point model (SCP), and the frequentist model (CBS) of Olshen et al. (2004). The three methods being compared are CBS, HMM, and SCP.

Simulation Model	CBS	HMM	SCP
HMM	0.0667	0.0967	0.0688
SCP	0.0557	0.1162	0.0334
CBS	0.0532	0.1092	0.0431

Table 4. Misclassification rates for CBS, HMM, and SCP compared on the simulation data generated by the HMM model of Fridlyand et al. (2004), the stochastic change-point model (SCP), and the frequentist model (CBS) of Olshen et al. (2004).

Simulation	CBS		HMM		SCP	
Model	FP	FN	FP	FN	FP	FN
HMM	0.0189	0.0717	0.0077	0.0518	0.0094	0.0343
	(0.0013)	(0.0034)	(0.0008)	(0.0045)	(0.0006)	(0.0012)
SCP	0.0087	0.0820	0.0297	0.3296	0.0647	0.0197
	(0.0026)	(0.0249)	(0.0113)	(0.0459)	(0.0041)	(0.0011)
CBS	0.0235	0.0417	0.0129	0.1175	0.0180	0.0321
	(0.0014)	(0.0017)	(0.0010)	(0.0037)	(0.0010)	(0.0014)

Figure 1. Genome-wide DNA copy-number variation for 9 Coriel breast cancer cell lines.

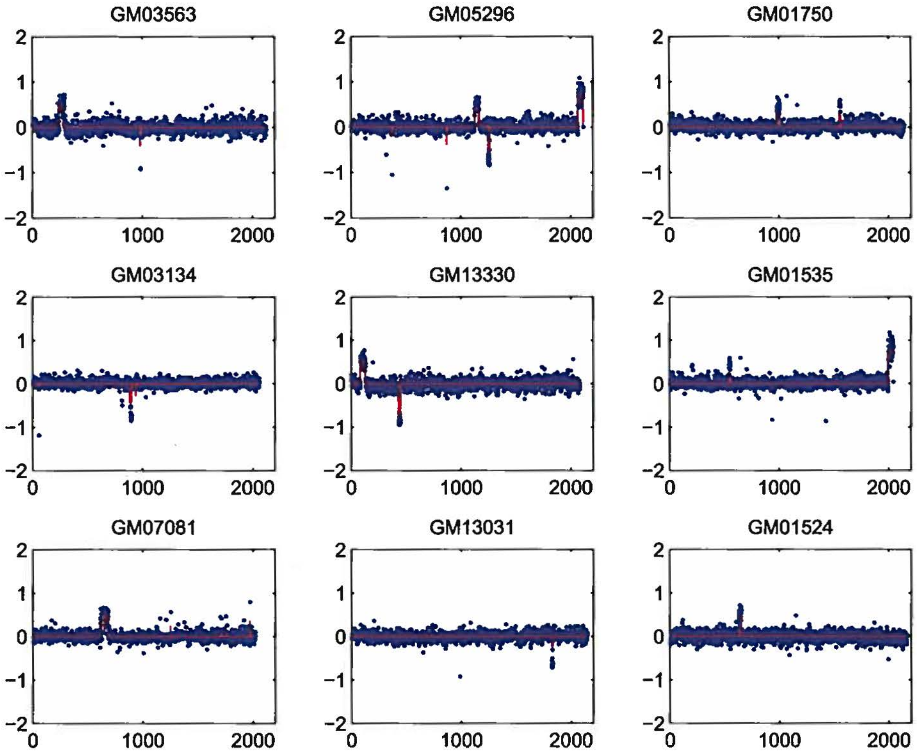


Figure 2. BAC array CGH profile for chromosome 17 in cell line BT474. The lines are the signal levels estimated using SCP (top plot), HMM (middle plot), and CBS (bottom plot). Note that SCP does not smooth out the sawtooth pattern in this region. Also shown in the top plot are the 2.5% and 97.5% quantiles (green lines) of the posterior distribution of θ_t estimated by SCP.

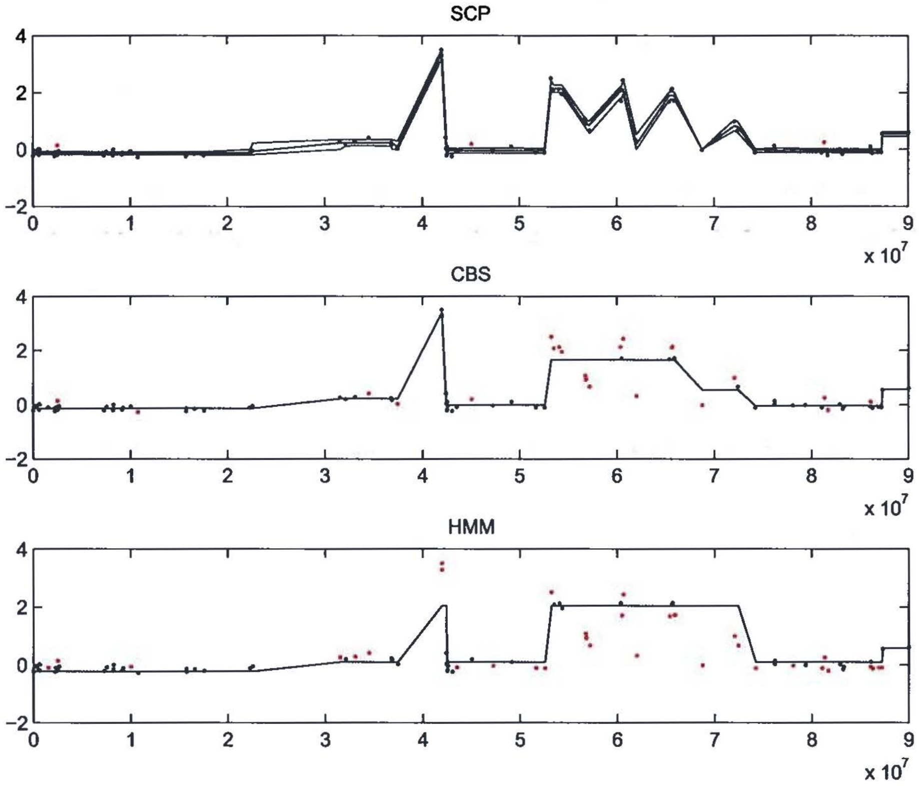


Figure 3. BAC array CGH profile for chromosome 20 in cell line BT474. The lines are the signal levels estimated using SCP (top plot), HMM (middle plot), and CBS (bottom plot). Also shown in the top plot are the 2.5% and 97.5% quantiles (green lines) of the posterior distribution of θ_t estimated by SCP, and the locations A, B, and C analyzed in Section 4.2.

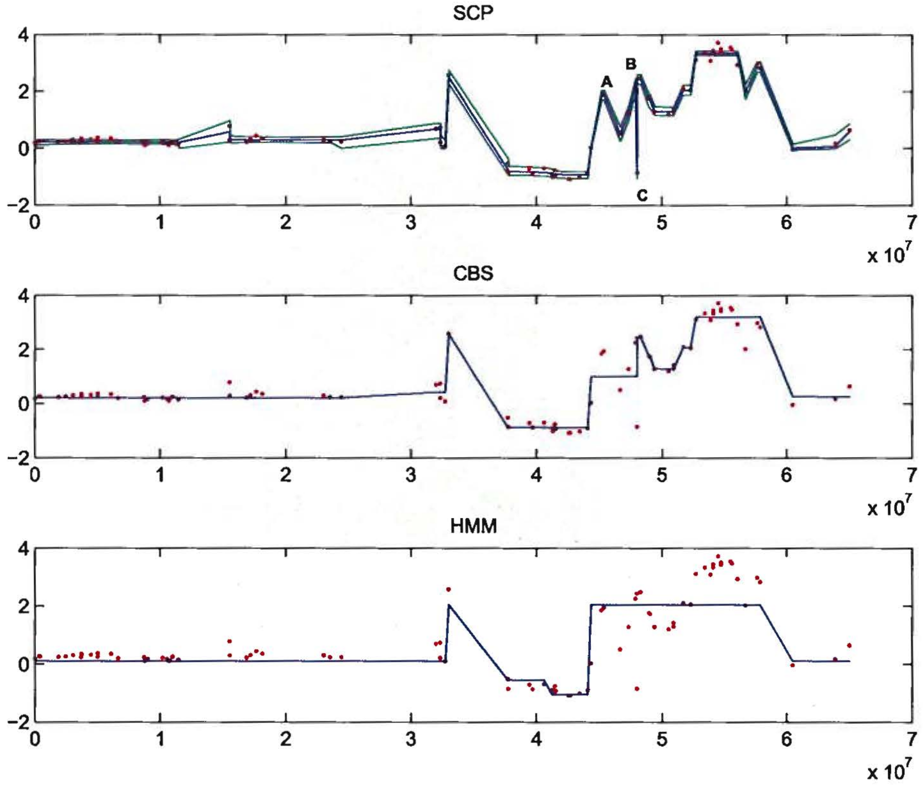


Figure 4. Histogram of number of segments in 5000 signal sequences simulated from the posterior distribution for cell line BT474.

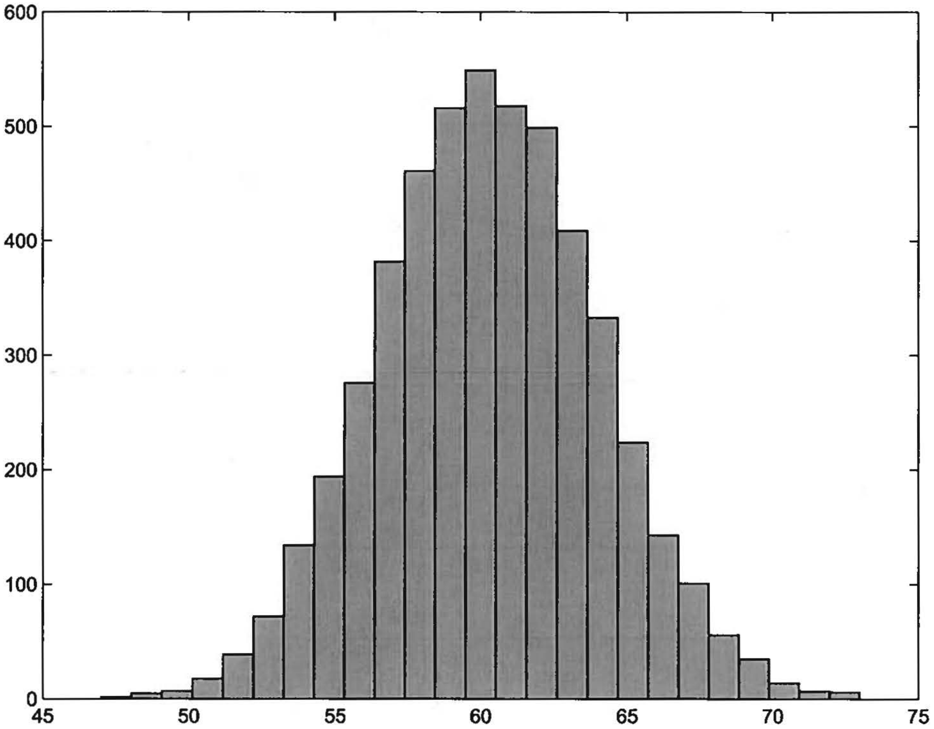
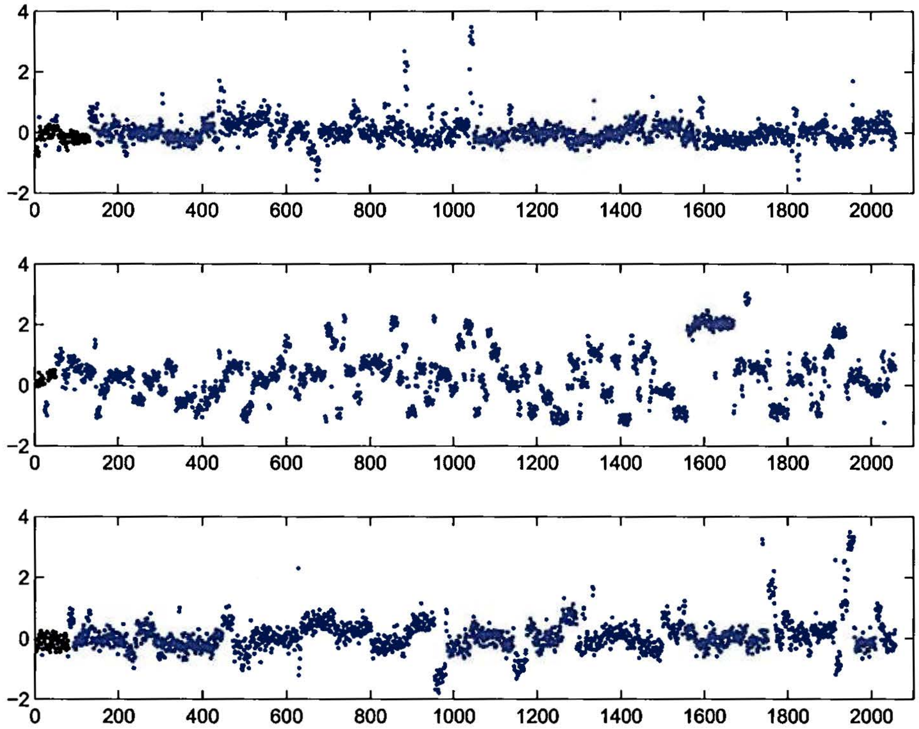


Figure 5. A simulation sequence generated from the HMM model (top plot), the stochastic change-point (SCP) model (middle plot), and the frequentist (CBS) model (bottom plot).


4
2
0
-2
0
200
400
600
800
1000
1200
1400
1600
1800
2000
4
2
0
-2
0
200
400
600
800
1000
1200
1400
1600
1800
2000
4
2
0
-2
0
200
400
600
800
1000
1200
1400
1600
1800
2000