

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*

HAIPENG XING

Department of Statistics, Columbia University, New York, NY 10027, USA

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 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.

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 2 dye-labeled samples are mixed and hybridized

*To whom correspondence should be addressed.

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 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 *and others*, 2004), recursive change-point detection (Circular Binary Segmentation [CBS], Olshen *and others*, 2004), a Gaussian model-based approach (Gain and Loss Analysis of DNA, Hupe *and others*, 2004), hierarchical tree-style clustering (Clustering Along Chromosomes, Wang *and others*, 2005), wavelet approximation (Hsu *and others*, 2005), a Bayes regression approach (Wen *and others*, 2006), and latent variable approaches using Gaussian mixture models (Engler *and others*, 2006; Broët and Richardson, 2006; Guha *and others*, 2006). Most of these methods approach the 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 the determination of the number of segments, which is sometimes referred to as the smoothness of the segmentation. Information-based model selection (Picard *and others*, 2005; Zhang and Siegmund, 2006) has been proposed as a guideline to this issue. The reviews by Willenbrock and Fridlyand (2005) and Lai, Johnson *and others* (2005) independently survey the effectiveness of existing methods on simulated 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 break points 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 latent variable model and associated inference framework similar to that proposed by Engler *and others* (2006), Broët and Richardson (2006), and Guha *and others* (2006). However, our model has key differences, which give it attractive statistical and computational properties, from these previous models. In particular, our model assumptions allow explicit computation of the posterior distributions of the latent variables, whereas previous methods rely on pseudolikelihood or Monte Carlo approximations. 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 diploid cells should be 2 on all autosomal chromosomes, 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 specificity. This is accomplished for our method through the assumption of a baseline state. Whether nonbaseline states with close mean levels should be merged is questionable. Inhomogeneity and microevolution 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, which is a quantity that arises naturally from our model, 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.

The Bayesian segmentation model contains certain hyperparameters. Their estimation is considered in Appendix C (see supplementary material available at *Biostatistics* online), where other implementation issues are also discussed. In Section 3, we apply our method to several real array-CGH data sets and illustrate the usefulness of confidence assessments for different scenarios. Section 4 evaluates the performance of our method on simulated data that are generated from our and other models. Some concluding remarks are given in Section 5, 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 nonzero state; when the next jump occurs, the signal may move back to the baseline or jump to another nonzero state. Suppose the log fluorescence ratios y_t follow the model

$$y_t = \theta_t + \sigma \epsilon_t, \quad \epsilon_t \sim N(0, 1), \quad (2.1)$$

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

$$P = \begin{pmatrix} 1-p & \frac{1}{2}p & \frac{1}{2}p \\ c & a & b \\ c & b & a \end{pmatrix}. \quad (2.2)$$

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 nonzero state to a different nonzero state, we simply assume that the process can jump from the baseline state with probability $p/2$ to either of the 2 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 , 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 that the 3-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 $K_t = \max\{s \leq t: \theta_s = \dots = \theta_t, \theta_{s-1} \neq \theta_s\}$ denote the nearest change-point at a location less than or equal to t . Let $\mathcal{Y}_t = (y_1, \dots, y_t)$ and $\mathcal{Y}_{i,j} = (y_i, \dots, y_j)$. Define

$$p_t = P(\theta_{K_t} = 0 | \mathcal{Y}_t) = P(\theta_t = 0 | \mathcal{Y}_t), \quad q_{i,t} = P(\theta_{K_t} \neq 0, K_t = i | \mathcal{Y}_t), \quad (2.3)$$

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

$$v_{i,j} = \left(\frac{1}{v} + \frac{j-i+1}{\sigma^2} \right)^{-1}, \quad \mu_{i,j} = \left(\frac{\mu}{v} + \sum_{k=i}^j \frac{y_k}{\sigma^2} \right) v_{i,j}, \quad (2.4)$$

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:

$$\theta_t | \mathcal{Y}_t \sim p_t \delta_0 + \sum_{i=1}^t q_{i,t} N(\mu_{i,t}, v_{i,t}), \quad (2.5)$$

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 $y_t = \theta_t + \sigma \epsilon_t$, Appendix A (supplementary material available at *Biostatistics* online) show that the conditional probabilities p_t and $q_{i,t}$ can be determined by the recursions

$$\begin{aligned} p_t &\propto p_t^* := (1-p)p_{t-1} + cq_{t-1}, \\ q_{i,t} &\propto q_{i,t}^* := \begin{cases} (pp_{t-1} + bq_{t-1})\psi/\psi_{i,t}, & i = t, \\ aq_{i,t-1}\psi_{i,t-1}/\psi_{i,t}, & i < t, \end{cases} \end{aligned} \quad (2.6)$$

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, $p_t = p_t^* / [p_t^* + \sum_{i=1}^t q_{i,t}^*]$ and $q_{i,t} = q_{i,t}^* / [p_t^* + \sum_{i=1}^t q_{i,t}^*]$. By (2.3) and (2.5),

$$P(\theta_t = 0 | \mathcal{Y}_t) = p_t, \quad E(\theta_t | \mathcal{Y}_t) = \sum_{i=1}^t q_{i,t} \mu_{i,t}. \quad (2.7)$$

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 (2.5):

$$\theta_{t+1} | \mathcal{Y}_{t+1,n} \sim \tilde{p}_{t+1} \delta_0 + \sum_{j=t+1}^n \tilde{q}_{j,t+1} N(\mu_{t+1,j}, v_{t+1,j}), \quad (2.8)$$

in which the weights \tilde{p}_s and $\tilde{q}_{j,s}$ can be obtained by backward induction using the time-reversed counterpart of (2.6):

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

where $\tilde{q}_{s+1} = \sum_{j=s+1}^n \tilde{q}_{j,s+1} = 1 - \tilde{p}_{s+1}$. 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 (2.8) and the reversibility of $\{\theta_t\}$ that

$$\theta_t | \mathcal{Y}_{t+1,n} \sim [(1-p)\tilde{p}_{t+1} + c\tilde{q}_{t+1}]\delta_0 + (p\tilde{p}_{t+1} + b\tilde{q}_{t+1})N(\mu, v) + a \sum_{j=t+1}^n \tilde{q}_{j,t+1} N(\mu_{t+1,j}, v_{t+1,j}). \quad (2.9)$$

We can use Bayes' theorem to combine the forward filter (2.5) with its backward variant (2.9) 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:

$$\theta_t | \mathcal{Y}_n \sim \alpha_t \delta_0 + \sum_{1 \leq i \leq t \leq j \leq n} \beta_{ijt} N(\mu_{ij}, v_{ij}). \quad (2.10)$$

In particular, by Bayes' theorem

$$\begin{aligned} \alpha_t &= P(\theta_t = 0 | \mathcal{Y}_n) \propto P(\theta_t = 0 | \mathcal{Y}_t) P(\theta_t = 0 | \mathcal{Y}_{t+1,n}) / \pi(0) \\ &= p_t [(1-p)\tilde{p}_{t+1} + c\tilde{q}_{t+1}] / [c/(p+c)]. \end{aligned} \quad (2.11)$$

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

$$\begin{aligned} \alpha_t &= \alpha_t^* / A_t, \quad \beta_{ijt} = \beta_{ijt}^* / A_t, \quad A_t = \alpha_t^* + \sum_{1 \leq i \leq t \leq j \leq n} \beta_{ijt}^*, \\ \alpha_t^* &= p_t [(1-p)\tilde{p}_{t+1} + c\tilde{q}_{t+1}] / c, \\ \beta_{ijt}^* &= \begin{cases} q_{i,t} (p\tilde{p}_{t+1} + b\tilde{q}_{t+1}) / p, & i \leq t = j, \\ a q_{i,t} \tilde{q}_{j,t+1} \psi_{i,t} \psi_{t+1,j} / (p \psi_{i,j}), & i \leq t < j. \end{cases} \end{aligned} \quad (2.12)$$

From (2.10), it follows that

$$P(\theta_t = 0 | \mathcal{Y}_n) = \alpha_t, \quad E(\theta_t | \mathcal{Y}_n) = \sum_{1 \leq i \leq t \leq j \leq n} \beta_{ijt} \mu_{ij}. \quad (2.13)$$

2.4 Inference on segmentation and parameter subsequences

The α_t and β_{ijt} in (2.12) are posterior probabilities that are useful for inference. As shown in (2.13), $\alpha_t = P(\theta_t = 0 | \mathcal{Y}_n)$. Moreover, the derivation of (2.12) in Appendix A shows that, for $i \leq t \leq j$,

$$\beta_{ijt} = P(C_{ij} | \mathcal{Y}_n), \quad \text{where } C_{ij} = \{\theta_i = \dots = \theta_j \neq 0, \theta_i \neq \theta_{i-1}, \theta_j \neq \theta_{j+1}\}. \quad (2.14)$$

For the problem of classifying location t as 0 (no copy number change or normal), G (copy number gain), or L (copy number loss), 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 the ‘‘normal’’ class usually includes a margin w beyond which the location is considered aberrant due to copy number gain or loss. The reason is that θ_t is unobservable and small changes of θ_t from 0 are of little biological interest. Furthermore, the posterior probability α_t is more sensitive to hyperparameter settings than the posterior distribution of the means $\hat{\theta}_t$. Thus, we have found a decision rule based on $\hat{\theta}_t$ to be more robust. Specifically, location t is considered as G if $\theta_t > w$, as L if $\theta_t < -w$, and as 0 if $|\theta_t| \leq w$. The choice of w , therefore, 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 $\{|\theta_t| \leq w\}$.

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 prespecified precision. Then, we can use $P(\bigcup_{(i',j'): |i-i'| \leq k, |j-j'| \leq k} C_{i'j'}|\mathcal{Y}_n)$ 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 $k \leq \lfloor (j-i)/2 \rfloor$, these events $C_{i'j'}$ are disjoint and therefore

$$\sum_{(i',j'): |i-i'| \leq k, |j-j'| \leq k} P(C_{i'j'}|\mathcal{Y}_n) = P\left(\bigcup_{(i',j'): |i-i'| \leq k, |j-j'| \leq k} C_{i'j'}|\mathcal{Y}_n\right). \quad (2.15)$$

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 $\{\theta_t: 1 \leq t \leq n\}$ given \mathcal{Y}_n . It is shown in Appendix B (see supplementary material available at *Biostatistics* online) that this posterior distribution is that of an inhomogeneous Markov chain whose initial distribution is π and whose transition probabilities are given by

$$\theta_t|\theta_{t-1}, \mathcal{Y}_n \sim a_t \delta_0 + c_t \mathbf{1}_{\{\theta_{t-1} \neq 0\}} \delta_{\theta_{t-1}} + \sum_{j=t}^n b_{jt} N(\mu_{t,j}, v_{t,j}), \quad (2.16)$$

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

$$B_t = a_t^* + c_t^* \mathbf{1}_{\{\theta_{t-1} \neq 0\}} + \sum_{j=t}^n b_{jt}^*,$$

$$a_t^* = \phi_{0,\sigma^2}(y_t) [(1-p) \mathbf{1}_{\{\theta_{t-1}=0\}} + c \mathbf{1}_{\{\theta_{t-1} \neq 0\}}] [(1-p) \tilde{p}_{t+1} + c \tilde{q}_{t+1}] / c,$$

$$c_t^* = a \phi_{\theta_{t-1}, \sigma^2}(y_t) \left\{ (p \tilde{p}_{t+1} + b \tilde{q}_{t+1}) + a \sum_{j=t+1}^n \tilde{q}_{j,t+1} \frac{\phi_{\mu_{t+1,j}, v_{t+1,j}}(\theta_{t-1})}{\phi_{\mu,v}(\theta_{t-1})} \right\} / p,$$

$$b_{jt}^* = [p \mathbf{1}_{\{\theta_{t-1}=0\}} + b \mathbf{1}_{\{\theta_{t-1} \neq 0\}}] \phi_{0,\sigma^2}(y_t) \tilde{q}_{j,t}^* / p,$$

using the same notation as that in (2.12).

Making use of the transition probabilities (2.16) of the inhomogeneous Markov chain, we can use the following recursive procedure to sample from the joint posterior distribution of the parameters $\theta_{t_1}, \dots, \theta_{t_2}$ (given \mathcal{Y}_n) in a segment $[t_1, t_2]$. Initialize at location $t = t_1$ by sampling θ_t from the distribution (2.10) for $\theta_t|\mathcal{Y}_n$. At location $t_1 < t \leq t_2$, if $\theta_{t-1} = 0$, sample θ_t from $N(\mu_{t,j}, v_{t,j})$ with probability b_{jt} for

$t \leq j \leq n$, 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 $t \leq j \leq n$. The posterior distribution of $(\theta_{t_1}, \dots, \theta_{t_2})$ given \mathcal{Y}_n , evaluated from a large number of simulated trajectories sampled from it can be used for statistical inference on the segment $[t_1, t_2]$. Some specific applications are given in Section 3.2, in which the special case $t_1 = 1$ and $t_2 = n$ covers an entire genome.

2.5 Bounded complexity mixture approximations

Although the Bayes filter (2.5) uses a recursive updating formula (2.6) for the weights $q_{i,t}$ ($1 \leq i \leq t$), the number of weights increases with t , resulting in rapidly increasing computational complexity and memory requirements in estimating θ_t as t keeps increasing. A simple idea to lower the 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, Liu, and Xing (2005) who consider the case without a baseline state, we keep the most recent m weights $q_{i,t}$ (with $t - m < i \leq t$) and the largest $k - m$ of the remaining weights, where $1 \leq m < k$. Specifically, the updating formula (2.6) 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 $\mathcal{K}_{t-1} \supset \{t - 1, \dots, t - m\}$. At stage t , define $q_{i,t}^*$ by (2.6) for $i \in \{t\} \cup \mathcal{K}_{t-1}$ and let i_t be the index not belonging to $\{t, t - 1, \dots, t - m + 1\}$ such that

$$q_{i_t,t}^* = \min\{q_{i,t}^* : j \in \mathcal{K}_{t-1} \text{ and } j \leq t - m\}, \quad (2.17)$$

choosing i_t to be the one farthest from t if the minimizing set in (2.17) has more than one element. Define $\mathcal{K}_t = \{t\} \cup (\mathcal{K}_{t-1} - \{i_t\})$ and let

$$p_t = p_t^* / \left(p_t^* + \sum_{j \in \{t\} \cup \mathcal{K}_{t-1}} q_{j,t}^* \right),$$

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

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 $\tilde{\mathcal{K}}_{t+1}$ for the backward filter at stage t . The BCMIX approximation $\alpha_t \delta_0 + \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \tilde{\mathcal{K}}_{t+1}} \beta_{ijt} N(\mu_{ij}, v_{ij})$ to (2.10) is defined by

$$\alpha_t = \alpha_t^* / A_t, \quad \beta_{ijt} = \beta_{ijt}^* / A_t, \quad A_t = \alpha_t^* + \sum_{i \in \mathcal{K}_t, j \in \{t\} \cup \tilde{\mathcal{K}}_{t+1}} \beta_{ijt}^*,$$

$$\alpha_t^* = p_t [(1 - p) \tilde{p}_{t+1} + c \tilde{q}_{t+1}] / c,$$

$$\beta_{ijt}^* = \begin{cases} q_{i,t} (p \tilde{p}_{t+1} + b \tilde{q}_{t+1}) / p, & i \in \mathcal{K}_t, j = t, \\ a q_{i,t} \tilde{q}_{j,t+1} \psi_{i,t} \psi_{t+1,j} / (p \psi \psi_{i,j}), & i \in \mathcal{K}_t, j \in \tilde{\mathcal{K}}_{t+1}. \end{cases}$$

3. APPLICATIONS TO REAL DATA SETS

We now illustrate our method and examine its performance on several real array-CGH data sets. In Section 3.1, we consider the bacterial artificial chromosome (BAC) array hybridizations of the Coriel cell lines from Snijders *and others* (2001), which are taken from 15 primary breast tumors. Out of these

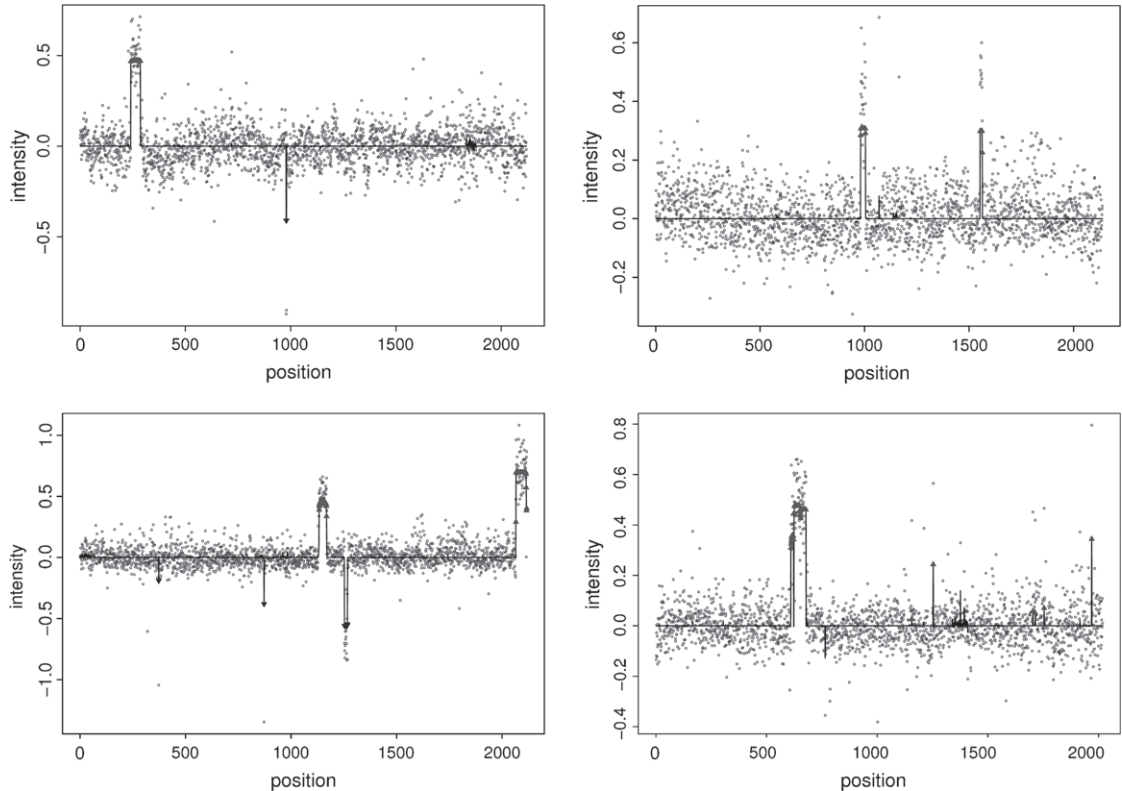


Fig. 1. L_1 distance between true and estimated signals for Coriel cell lines using SCP, CBS, and HMM.

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 3.2, we use the BAC array hybridization of the BT474 cell line, taken from Snijders *and others* (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 Figure 3 and see supplementary Figure 1 available at *Biostatistics* online. These more challenging data sets do not reveal obvious segmentations, and thus a framework for inference becomes crucial.

Finally, in Section 3.3 we evaluate the performance of our method on Glioblastoma multiforme (GBM) cell lines from Bredel *and others* (2005). These data have been used by Lai, Johnson, *and others* (2005) to evaluate 11 different methods, which can be directly compared to the performance of our method.

3.1 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 called the stochastic change-point (denoted by SCP) 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 Appendix C (see supplementary material available at *Biostatistics* online), limiting the global search to $10^{-4} < p < 0.005$ and $10^{-4} < c < 0.05$. For such simple cell lines, we recommend setting the hyperparameters around reasonable values as done above, since with so few aberrant regions there may be not enough data to estimate some of the hyperparameters, such as b . However, one could potentially pool data across samples to estimate the hyperparameters via EM.

Figure 1 plots the posterior means $E(\theta_t | \mathcal{Y}_n)$, $1 \leq t \leq n$, for 4 of the 9 cell lines. Our method yields correct estimates for the other 5 cell lines, which have plots similar to those in the top panel. For this simple data set, the true copy numbers are known. As can be seen from the plots, our method perfectly distinguishes the signals from the noise for all cell lines except GM05296 and GM07081. For GM05296 there were 2 false-positive and for GM07081 there were 5 false-positive aberration calls.

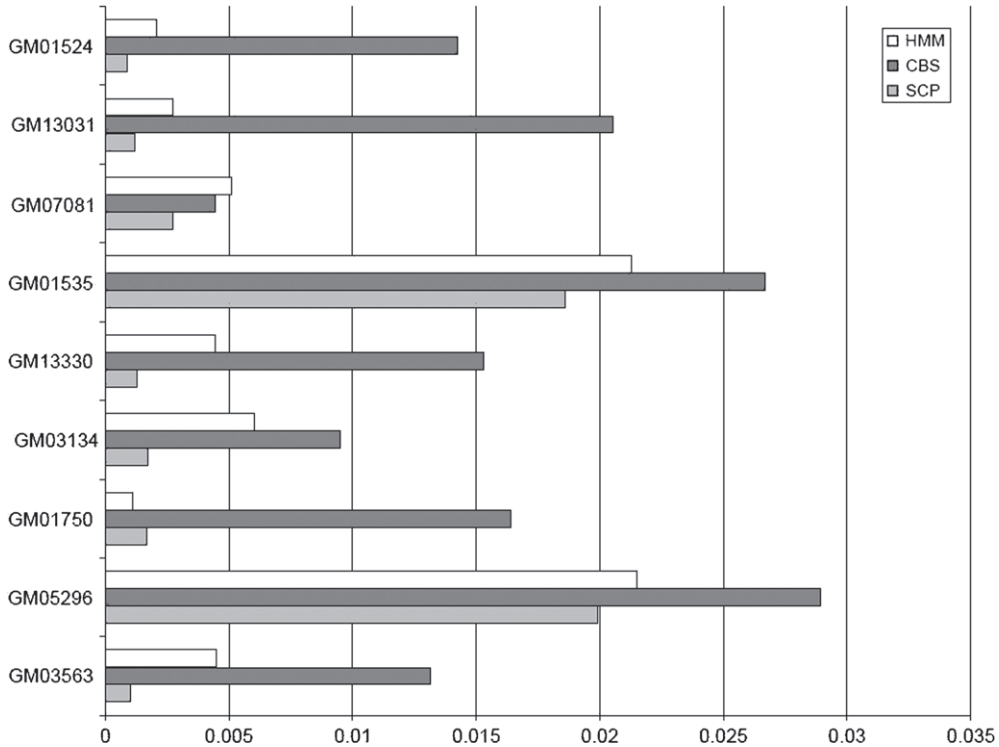


Fig. 2. Genome-wide DNA copy number variation for Coriel breast cancer cell lines. Due to lack of space, only 4 cell lines are plotted. The other 4 cell lines have results similar to GM03563 and GM01750 (top panel). Data are plotted in gray, estimated signals in black. The aberration calls are given by the arrows (arrows pointing up indicate amplification, arrows pointing down indicate deletion).

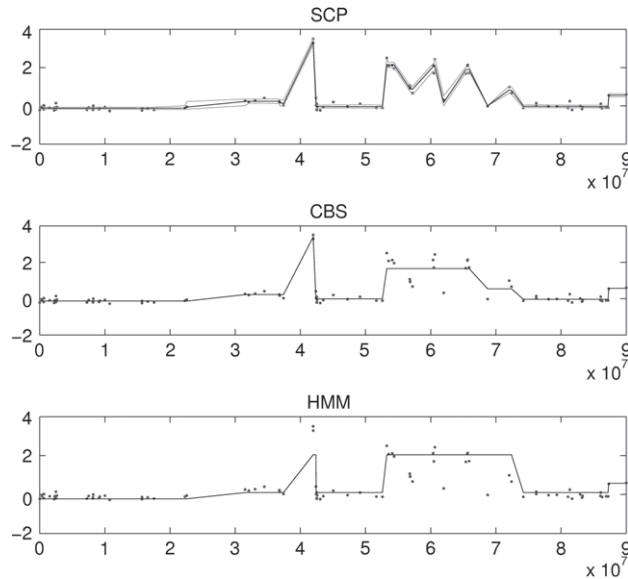


Fig. 3. BAC array-CGH profile for chromosome 17 in cell line BT474. The lines are the signal levels estimated using SCP (top plot), HMM (middle plot), and CBS (bottom plot). Note that SCP does not smooth out the sawtooth pattern in this region. Also shown in the top plot are the 2.5 and 97.5% quantiles (gray lines) of the posterior distribution of θ_t estimated by SCP.

To give another assessment of the performance of our method, we calculated the L_1 distance $\sum_{t=1}^n |\hat{\theta}_t - \theta_t|$ for the estimated signal produced by a given method. The methods that we choose for comparison are the HMM-based algorithm (HMM) of Fridlyand *and others* (2004) and the CBS algorithm of Olshen *and others* (2004). For HMM, we start with 5 states in the HMM and apply the state-merging step with a merging threshold of 0.25 as recommended in Fridlyand *and others* (2004). For the CBS algorithm, we used the default parameters for the “dnacopy” software in Bioconductor. Figure 2, 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.

3.2 Breast cancer cell line BT474

For the cell line BT474, we use the EM algorithm in Appendix C (see supplementary material available at *Biostatistics* online) to estimate the hyperparameters. With initial values of (p, a, b) at $(0.05, 0.995, 0.0025)$ and (μ, ν, σ^2) at $(0.065, 0.087, 0.020)$, the EM algorithm stops after 7 iterations according to a convergence criterion, yielding $\hat{p} = 0.7196$, $\hat{a} = 0.9147$, $\hat{c} = 0.0662$, $\hat{\mu} = 0.3063$, $\hat{\nu} = 0.5668$, and $\hat{\sigma}^2 = 0.0152$.

The top plots of Figure 3 and supplementary Figure 1 available at *Biostatistics* online show respectively 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. The q th 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 (2.10), is obtained by solving the equation

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

Because of the complexity of this cell line, previous methods disagree widely on the correct segmentation, as can be seen by comparing the signal estimates given by CBS, HMM, and our method on these 2 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 3, 50- to 70-Mb region, and supplementary Figure 1 available at *Biostatistics* online, 40- to 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 *and others* (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

Table 1. *Ranking of the aberrations in BT474 cell line by the posterior probability $P(C_{ij}|\mathcal{Y}_n)$ (truncated list)*

Chromosome number	AugKB region	Posterior probability	Posterior mean	Chromosome number	AugKB region	Posterior probability	Posterior 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	X	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

The corresponding posterior means $(j - i + 1)^{-1} \sum_{t=i}^j E(\theta_t|\mathcal{Y}_n)$ are also shown.

the right boundary j , for all $i < j$, on the same chromosome arm. In Table 1, the aberrations detected in BT474 are ranked by $P(C_{ij}|\mathcal{Y}_n)$. Comparing Table 1 with Figures 3 and supplementary Figure 1 available at *Biostatistics* online, 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 contains 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 *and others*, 1999; Pinkel *and others*, 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.

Finer scale confidence assessments targeted at a specific genome region also arise naturally from our framework. We illustrate this with the data from chromosome 20 in BT474 (Supplementary Figure 1). 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). Supplementary Figure 1 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- to 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 (2.15), with $k^* = 2$, for a single segment in 40- to 50-MB region proposed by the HMM procedure is 0.000, whereas

$$P\{[i, j] \text{ contains a subsegment } [i', j'] \text{ such that } \theta_{i'} = \dots = \theta_{j'} \neq 0,$$

$$\theta_{i'} \neq \theta_{i'-1}, \theta_{j'} \neq \theta_{j'+1}, i' - 1 \geq i, j' + 1 \leq j\} \geq \max_{i < i' \leq j' < j} P(C_{i'j'}|\mathcal{Y}_n),$$

which exceeds 0.997, 0.999, and 1, respectively, for the segments $[i', j'] = A, B, C$ within the 40- to 50-Mb region in Supplementary Figure 1. Thus, the probability of a spike within the 40- to 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 *and others*, 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 HMM, 103 change-points are found if the state-merging step proposed by Fridlyand *and others* (2004) is not taken; after the merging of states, the complete procedure from Fridlyand *and others* (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

$$\kappa = \sum_{i=1}^{n-1} \mathbf{1}_{\{|\theta_{i+1} \neq 0, |\theta_i - \theta_{i+1}| > \delta\}}, \quad (3.18)$$

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 (2.16). Supplementary Figure 2 available at *Biostatistics* online shows the histogram of κ , in which we set $\delta = \sqrt{v}$ in (3.18), calculated for 5000 simulated sequences; recall that our model assumes Gaussian jumps with variance v . The mean and 95% confidence interval of κ based on these 5000 simulations are 60.24 and [60.14, 60.35].

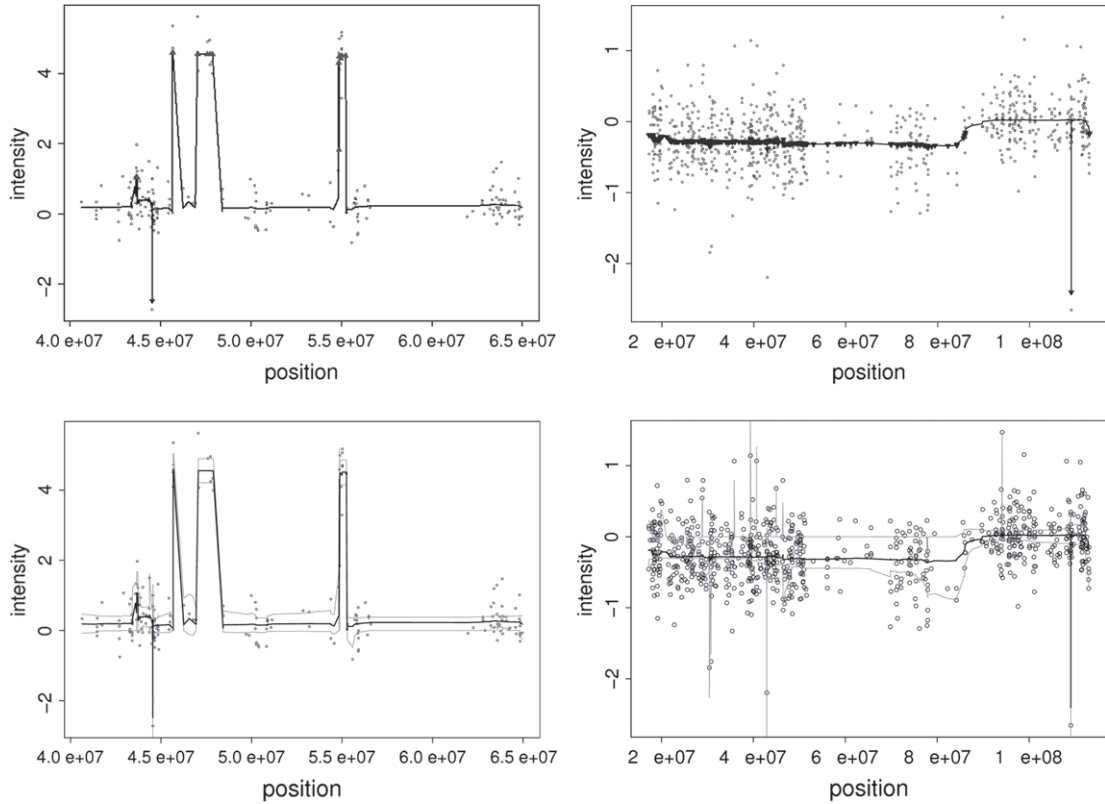


Fig. 4. GBM29 chromosomes 7 and 13. Upper plots show the fitted signals and the aberration calls. Lower plots show the 2.5 and 97.5% confidence bands (gray lines).

3.3 GBM data

To facilitate comparison with other methods on real array-CGH data, we also analyzed the GBM data of Bredel *and others* (2005). This data set has been used in the comparative analysis of Lai, Johnson, *and others* (2005), comparing the estimates of the signal for 11 methods that are different from ours. Guha *and others* (2006) also used these data to assess their method. Figure 4 shows the results of our method on GBM29 chromosome 7 and GBM31 chromosome 13. All analyses are done with EM hyperparameter estimation using default settings. GBM29 chromosome 7 is typical of midsize high amplitude aberrations. Our method correctly calls 3 regions of amplification. Figure 4 also shows the confidence bands. The confidence band for GBM29 chromosome 7 is relatively narrow compared to the jumps, indicating a very strong signal. On the other hand, GBM31 chromosome 13 contains a low amplitude deletion that was missed by many of the previous methods (Lai, Johnson, *and others*, 2005). Our algorithm is able to call this deletion correctly, as shown by Figure 4. However, the confidence bands are rather wide, indicating that this is a relatively difficult data set.

4. 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 $y_t = \theta_t + \epsilon_t$, $1 \leq t \leq n$,

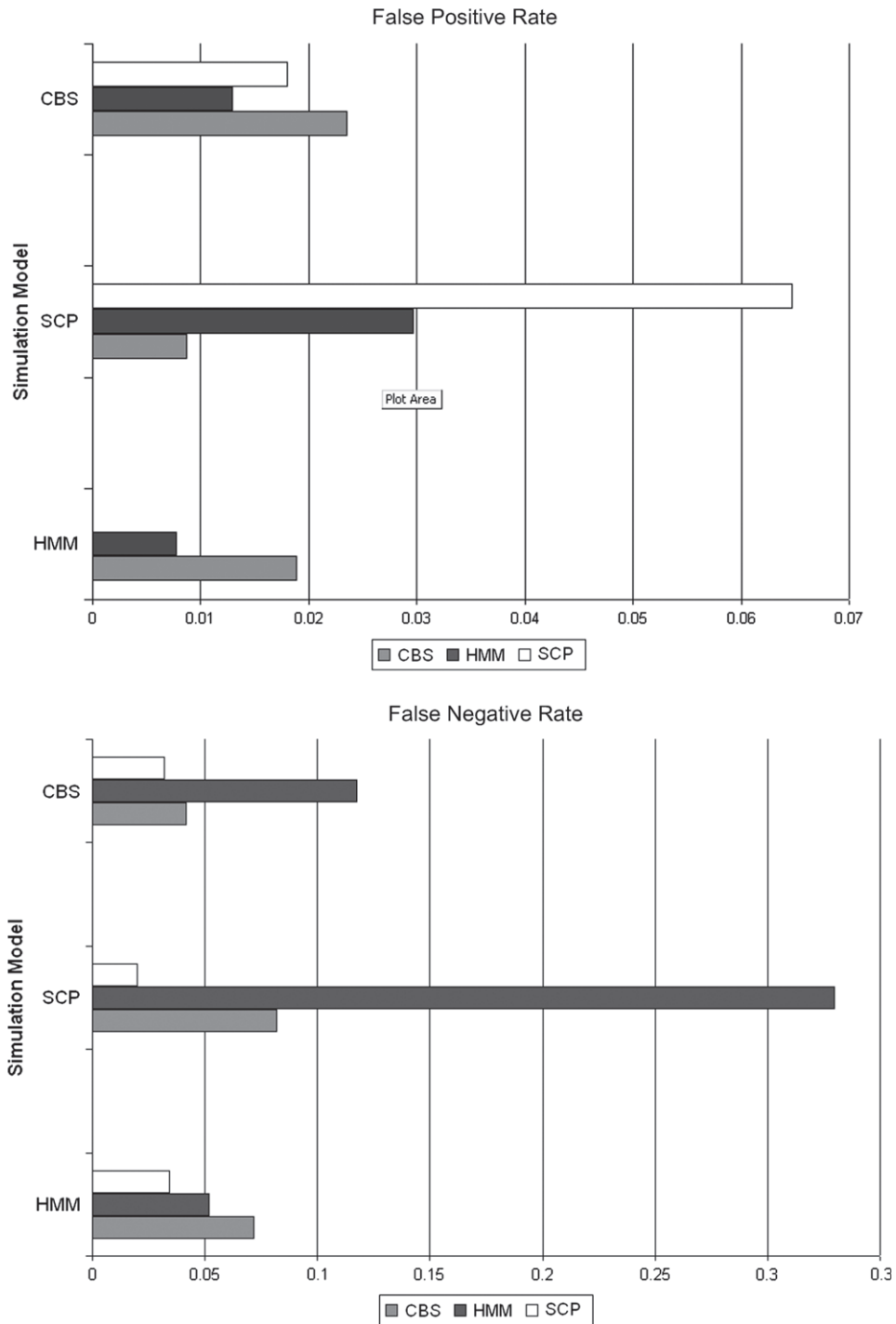


Fig. 5. Misclassification rates for CBS, HMM, and SCP compared to the simulation data generated by the HMM of Fridlyand *and others* (2004), the SCP model, and the frequentist model (CBS) of Olshen *and others* (2004).

where ϵ_t are i.i.d. $N(0, \sigma^2)$ and one of the following 3 models is used to generate θ_t : (a) the HMM of Fridlyand *and others* (2004), in which $\{\theta_t\}$ is a finite-state Markov chain; (b) the SCP model described in Section 2; (c) the frequentist model considered by Olshen *and others* (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, 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 Akaike information criterion. 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 3.2. For the frequentist model, the θ_t are the estimated mean levels for BT474 using the CBS algorithm. We used $n = 2056$ for all 3 models, which is the same length as the complete BT474 data set without missing values. Supplementary Figure 3 available at *Biostatistics* online shows an example of a simulation data series generated from each of the 3 models. The series generated from the CBS model looks most similar to “real” array-CGH data because it keeps the fitted means from BT474 and only simulates the errors. The series generated from the HMM involves more randomization: it keeps the estimated state means and variances from BT474, but simulates the locations of changes as well as the errors. The series generated from the SCP model appears to be most unlike the original BT474 data series: it only keeps the hyperparameters estimated from BT474 and randomly simulates the state means, transition points, and errors. Thus, this variety of simulation models provides a fair assessment of our method.

We simulated 100 data series from each model for this study. 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.

We consider the performance of our classification procedure described in the second paragraph of Section 2.4. We choose the threshold $w = 2\hat{\sigma}$ for each sequence, in which $\hat{\sigma}$ 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. Figure 5 shows the false-positive and false-negative rates (in classifying no change versus a gain or a loss) for the 3 different methods in each of the 3 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 $w = 2\hat{\sigma}$.

5. DISCUSSION

We have developed a SCP 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. Estimation of the hyperparameters involves an explicit EM algorithm or a hybrid method described in Appendix C. The signal estimates and inference from our method are robust to changes in the hyperparameters within a narrow range of their optimal value. In practice, the performance of our model is better when the training set used for estimation of hyperparameters is large.

In Section 3, we have used several 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 3.1 we give a method for calculating pointwise marginal confidence intervals for the estimated

signal. In Section 3.2, 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 produces a list of aberration calls with associated posterior probabilities, giving the biologist an informed option to investigate further.

A departure of our model from most earlier models is the assumption of a baseline state, which yields a natural classification of genomic regions into “amplified,” “deleted,” and “normal” states. The models recently proposed in Engler *and others* (2006), Guha *and others* (2006), and Broët and Richardson (2006) also give such a classification rule. However, these papers do not provide explicit formulas for computing posterior probabilities and instead rely on pseudolikelihood or Markov chain Monte Carlo approaches, which may require heavy computation to obtain good approximations. We provide explicit formulas for Bayesian analysis and maximum likelihood estimation of the hyperparameters.

Because of our model assumptions, if the data do not have a baseline state at 0, then our method would not perform very well. However, for array-CGH data, if the baseline state is not at 0, then normalization procedures can usually be applied to get a zero baseline state.

We chose to conduct our data analysis at the genome level, rather than at the chromosome level because the interchromosome 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 multichromosome analysis improves sensitivity has also been shown in Engler *and others* (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 change-points exist.

The data sets used in Section 3 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 *and others* 2002), molecular inversion probe technology (Hardenbol *and others* 2003), and oligonucleotide arrays (Komura *and others* 2006). Wen *and others* (2006) have pointed out the need for incorporating possible changes in both the mean and variance in the analysis of cDNA array-CGH data. By making use of the ideas of Lai *and others* (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. This is a topic for future research.

The software used in this paper was written in R and C++ and is freely available at <http://www-stat.stanford.edu/nzhang/SCP/>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

National Science Foundation (DMS-0305749); National Institutes of Health (CA088890) to T. L.

REFERENCES

- BREDEL, M., BREDEL, C., JURIC, D., HARSH, G., VOGEL, H., RECHT, L. AND SIKIC, B. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Research* **65**, 4088–4096.
- BROËT, P. AND RICHARDSON, S. (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911–918.

- ENGLER, D. A., MOHAPATRA, G., LOUIS, D. N. AND BETENSKY, R. A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399–421.
- FABARIUS, A., HEHLMANN, R., AND DUESBERG, P. H. (2003). Instability of chromosome structure in cancer cells increases exponentially with degrees of aneuploidy. *Cancer Genetics and Cytogenetics* **143**, 59–72.
- 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. *Journal of Multivariate Analysis* **90**, 132–153.
- GUHA, S., LI, Y. AND NEUBERG, D. (2006). Bayesian hidden Markov modeling of array CGH data. *Harvard University Biostatistics Working Paper Series*. Working paper 24. Available at: <http://www.bepress.com/harvardbiostat/paper24>.
- HARDENBOL, P., BANER, J., JAIN, M., NILSSON, M., NAMSARAIEV, E.A., KARLIN-NEUMANN, G. A., FAKHRAI-RAD, H., RONAGHI, M., WILLIS, T. D., LANDEGREN, U. and others (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology*, **21**, 673–678.
- HICKS, J., MUTHUSWAMY, L., KRASNITZ, A., NAVIN, N., RIGGS, M., GRUBOR, V., ESPOSITO, D., ALEXANDER, J., TROGE, J., WIGLER, M. and others (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.G, 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.
- HUPÉ, P., STRANSKY, N., THIERY, J., RADVANYI, F., AND BARILLOT, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422.
- KOMURA, D., SHEN, F., ISHIKAWA, S., FITCH, K. R., CHEN, W., ZHANG, J., LIU, G., IHARA, S., NAKAMURA, H., HURLES, M. E. and others (2006). Genome-wide detection of human copy number variations using high density DNA oligonucleotide arrays. *Genome Research*, 10.1101/gr.5629106.
- 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.
- 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. and others (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.
- POLLACK, J. R., SØRLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BØRRESEN-DALE, A., AND BROWN, P. O. (2002). Microarray analysis reveals a major direct role of DNA copy number alternation in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* **99**, 20, 12963–12968.
- SNIJDERS, A. M., FRIDLYAND, J., MANS, D. A., SEGRAVES, R., JAIN, A. N., PINKEL, D. AND ALBERTSON, D. G. (2003). Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379.

- SNIJDERS, A. M., NOWAK, N., SEGRAVES, R., BLACKWOOD, S., BROWN, N., CONROY, J., HAMILTON, G., HINDLE, A. K., HUEY, B., KIMURA K. *and others* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263–264.
- 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 others* (2006). A Bayes regression approach to array-CGH data. *Statistical Applications in Genomics and Molecular Biology* **5**, Article 3. Available at: <http://www.bepress.com/sagmb/vol5/iss1/art3>.
- 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* **63**, 22–32.

[Received October 10, 2006; revised June 4, 2007; accepted for publication July 11, 2007]