Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Xing and Cai, J Biomet Biostat 2015, 6:1
DOI: 10.472/2155-6180.1000221
Research Article
Open Access

A Stochastic Segmentation Model for Recurrent Copy Number Alteration Analysis

Haipeng Xing* and Ying Cai

Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794, USA

Abstract

Recurrent DNA copy number alterations (CNAs) are key genetic events in the study of human genetics and disease. Analysis of recurrent DNA CNA data often involves the inference of individual samples' true signal levels and the cross-sample recurrent regions at each location. We propose for the analysis of multiple samples CNA data a new stochastic segmentation model and an associated inference procedure that has attractive statistical and computational properties. An important feature of our model is that it yields explicit formulas for posterior probabilities of recurrence at each location, which can be used to estimate the recurrent regions directly. We propose an approximation method whose computational complexity is only linear in sequence length, which makes our model applicable to data of higher density. Simulation studies and analysis of an ovarian cancer dataset with 15 samples and a lung cancer dataset with 10 samples are conducted to illustrate the advantage of the proposed model.

Keywords: Categorical states; Hidden Markov models; Multiple change-points; Recurrent CNAs

Introduction

Copy number alterations (CNAs) are key genetic events in the development and progression of numerous human diseases. Recent advances in high density microarray technologies enable high-throughput genome-wide profiling of DNA copy number; see [1,2]. Using the array-based comparative genomic hybridization (array-CGH) technology, the average genomic DNA copy number at thousands of locations linearly ordered along the chromosomes can be quantitatively measured [3]. Since cancer genes are more likely to be found in common or recurrent regions in the sequence of CNAs across patients of the same cancer [4], one is more interested in finding recurrent CNA regions that consist of continuous probes and show evidence of alteration in some samples [5].

During the past years, a large number of computational and statistical methods have been developed to locate the recurrent CNA regions across samples, see reviews and comparisons of these methods in [5,6]. Most of these methods involve a two-step procedure, in which the first step is to identify the gain and loss regions in individual samples and the second step is to make inference on recurrent regions based on a threshold for occurrence frequencies. Examples of these approaches can be found in [7-14]. As the first steps of these approaches require segmentation of individual samples, they may strengthen or weaken some important information in recurrent regions. In contrast to two-step methods, one-stage approaches analyze raw data directly and avoid the information change in the two-step process. Recently, several statistical approaches have been proposed along this line, including score-based approach [15,16] hierarchical hidden Markov model [17], Bayesian hidden Markov model [18] kernel smoothing methods [19,20] analysis of variance approach, and likelihood-based test for simultaneous change-points [21]. Most of these methods involve a hard segmentation procedure. However, for complex alteration profiles across samples, identified recurrent CNA regions vary greatly. This indicates that hard segmentation procedures may be difficult for identification of recurrent regions, and instead, an inference procedure on the probability of recurrent regions might be necessary.

In this paper, we propose for the analysis of recurrent CNAs a stochastic segmentation model and associated inference framework. The proposed model has a hierarchical hidden Markov structure which make the inference framework associated to our model possess attractive statistical and computational properties. The hierarchical hidden Markov structure in our model is similar to that in Shah et al. [7], but our model allows different "quantitative" states conditional on a given "categorical" state, while the model in Shah et al. [7], assumes all "quantitative" states are same for a "categorical" state. Specifically, we assume a finite state hidden Markov chain for (categorical) states of recurrent regions across samples, and then conditional on the categorical state, signal levels (or "quantitative states") of CNAs in each sample follow a sample-specific continuous state hidden Markov chain. As a working model, although these assumptions seem to complicate for obtaining an inference procedure, they actually provide us more flexibility to model the non-simultaneity feature of break points across samples and yields explicit recursive formulas for posterior distributions of hidden "categorical" states (i.e., the recurrent CNA region) and sample-specific "quantitative" states (i.e., the signal levels of CNAs in individual samples) at each probe, whereas the model in Shah et al. [7], has to rely on Monte Carlo simulations.

Our stochastic segmentation model assumes that the log fluorescence ratios y_{lt} for sample $l \in \{1, \ldots, L\}$ measured through the array-CGH technology follow that $y_{lt}=\theta_{lt} + \sigma_l \in_{lt} (t=1, \ldots, T)$, in which θ_{lt} are independent standard normal random variables, and θ_{lt} are piecewise constant whose values at location t follow a prior distribution that depends on a hidden Markov chain s_t with three categorical states (gain, baseline 0 or loss). In this specification, θ_{it} and s_t represent the true signal levels of CNAs in sample l and the gain-loss states across the L samples at location t. When s_t shifts from one categorical state to another, signal levels (or quantitative states) θ_{lt} in sample l jump to a new state, whose prior distribution depends on s_t, hence θ_{lt} may be

***Corresponding author:** Haipeng Xing, Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794, USA, E-mail: xing@ams.sunysb.edu

Received February 06, 2015; **Accepted** May 11, 2015; **Published** May 18, 2015

Citation: Xing H, Cai Y (2015) A Stochastic Segmentation Model for Recurrent Copy Number Alteration Analysis. J Biomet Biostat 6: 221. doi:10.4172/2155-6180.1000221

different from the quantitative states whose corresponding categorical states are same as s_t. Making use of this specific hierarchical model structure, we compute the posterior distributions of θ_{lt} and s_l based on explicit recursive formulas that are derived using forward and backward filters of the hidden Markov model. The forward-backward filters in our model can be considered as a non-trivial extension of the Baum-Welch algorithm and similar to those developed by [22-24]. The difference of our forward-backward filters from previous work is that the hidden categorical and quantitative states in our model have a hierarchical structure and the top layer hidden states become a finite-state Markov chain. As the problem of locating recurrent CNA regions intrinsically involves much computation, to reduce the computational complexity of our inference procedure, we further consider a bounded complexity mixture approximation scheme so that the computational complexity becomes linear. Another discussion we have made is that, since all model hyper parameters are unknown in real applications, we propose to estimate all hyper parameters by an expectation-maximization (EM) algorithm.

The rest of the paper is organized as follows. Section 2 provides the model details and develops an inference procedure. It also discusses some computational issues and proposes a bounded complexity mixture approximation scheme and a hyper parameter estimation algorithm. Section 3 shows the performance of our model and associated inference procedure via intensive simulation studies. Section 4 analyzes two groups of CNA data, one is on ovarian cancer based on 15 samples and the other is on lung cancer based on 10 patients. We identify the recurrent CNA regions related to those cancers and demonstrate that our result is consistent with that in current medical studies. Section 5 provides some conclusive remarks.

A Stochastic Segmentation Model

Model specification

We consider the problem of analyzing DNA copy number profiles from multiple distinct biological samples $\{y_{lt}: 1 \le l \le L,\ 1 \le t \le T\}$, where y_{lt} is the observed log fluorescence ratio at location t of sample l, T is the number of probes, and L is the number of samples. To estimate recurrent signals, we assume the following model for y_{lt}:

$$y_{lt}=\theta_{lt} + \sigma_l \epsilon_{lt}, \qquad (1)$$

in which ϱ_{lt} are independent Gaussian random errors with mean 0 and variance 1, σ_l^2 are sample-specific error variances and $\{\theta_{lt}\}$ is the true signal level of CNA of sample l at location t. Since we want to find recurrent regions across all L samples, we assume that recurrent regions can be represented by a "master" sequence of categorical states $\{s_t\}$, where $s_t \in \{G, O, S\}$ (gain, baseline 0, loss) is an irreducible hidden Markov chain with probability transition matrix $P=(p_{ij})$ and stationary distribution π. Then given the master sequence $\{s_t\}$, the dynamics of θ_{lt} in sample l is expressed as

$$\theta_{lt} = 1\{s_t = s_{t-1}\}\theta_{l,t-1} + 1\{s_t \ne s_{t-1}\} z_{lt}, \qquad (2)$$

in which z_{lt} are independent normal variables with mean $z^{(l,st)}$ and covariance $v^{(l,st)}$.

In the above model assumption, the existence of stationary distribution π could define us a reversed chain for $\{s_t\}$, and it further implies that the Markov chain $\{\theta_{lt}\}$ has a stationary distribution. Moreover, if we further assume that θl0 is initialized at the stationary distribution, $\{\theta_{lt}\}$ become a reversible Markov chain, which provides substantial simplification for the smoothing estimates of $\{\theta_{lt}\}$ and s_t. We should note that this assumption is to simplify the estimation procedure. It may not reflect the real situation since the probability of amplification or deletion might be different across the whole chromosome.

The assumption that the master sequence of states is common across all the samples may not be necessarily true in practice, and it is only an approximation for the fact that most samples share a unique profile signature. This assumption is used in some models to obtain an estimation procedure with reasonable computational complexity for identifying recurrent CNAs; see Shah et al. The assumption also implies that the model is not suitable for a class of samples that consists of several disease subclasses with each subclass having a unique profile signature. For sample with disease subclass, we need to know the information about disease subclasses before applying the above model. Furthermore, assumption (2) indicates that signal levels θ_{lt} with same categorical states could be different.

Filtering estimate

Let $J_t^{(k)} = max\{i \le t : s_{i-1} \ne s_i = \cdots = s_t = k\}$ be the most recent location where s_t switches to state k from other states prior or equal to t. Denote $\xi_t^{(k)} = P(s_t = k \mid Y_{1,t})$ and $\xi_{i,t}^{(k)} = P(J_t^{(k)} = i \mid Y_{1,t})$ for $1 \le i \le t$ and $1 \le k \le K=3$, in which $Y_{i,j} = (\tilde{y}_i, \ldots, \tilde{y}_j)$ and $\tilde{y}_i = (y_{1i}, \ldots, y_{Li})'$. By definition, we have $\xi_{i,t}^{(k)} = \sum_{i=1}^{t} \xi_{i,t}^{(k)}$. Then given $Y_{1,t}$ and the most recent switching location $J_t^{(k)} = i$, the conditional distribution of θ_{lt} is given by $N(z_{i,j}^{(l,k)}, v_{i,j}^{(l,k)})$, where for $j \ge i$,

$$v_{i,j}^{(l,k)} = \left(\frac{1}{v^{(l,k)}} + \frac{j-i-1}{\sigma_1^2}\right)^{-1},\quad z_{i,j}^{(l,k)} = v_{i,j}^{(l,k)}\left(\frac{z^{(l,k)}}{v^{(l,k)}} + \frac{1}{\sigma_1^2}\sum_{u=i}^{j} y_{lu}\right) \qquad (3)$$

conditional distribution of θ_{lt} given $Y_{1,t}$ becomes a mixture of normal distributions:

$$\theta_{lt}|Y_{1,t} \sim \sum_{k=1}^{k}\sum_{i=1}^{t} \xi_{i,t}^{(k)}\, N(z_{i,j}^{(l,k)}, v_{i,j}^{(l,k)}) \qquad (4)$$

Let $g_{i,t}^{(l,k)}(y)$ and $g_{0,0}^{(l,k)}(y)$ denote the density function of the $N(z_{i,j}^{(l,k)}, v_{i,j}^{(l,k)})$ and $N(z^{(l,s_t)}, v^{(l,s_t)})$ distributions at y, respectively. Making use of $\sum_{k=1}^{k}\sum_{i=1}^{t} \xi_{i,t}^{(k)} = 1$ and model assumption (1), Web Appendix A show that the conditional probabilities $\xi_{i,t}^{(k)}$ can be determined by the following recursions

$$\xi_{i,t}^{(k)} \propto \xi_{i,t}^{(k)*} := \begin{cases} (\sum_{r \ne k} \xi_{t-1}^{(r)} p_{rk}) \psi_{0,0}^{(k,k)} / \psi_{t,t}^{(k,k)} & \text{if } i=t, \\ p_{kk}\, \xi_{i,t-1}^{(k)} \psi_{i,t-1}^{(k)} / \psi_{i,t}^{(k)} & \text{if } i<t, \end{cases} \qquad (5)$$

where $\psi_{0,0}^{(k)} = \prod_{l=1}^{L} g_{0,0}^{(l,k)}(0)$ and $\psi_{i,j}^{(k)} = \prod_{l=1}^{L} g_{i,j}^{(l,k)}(0) = \left[g_{i,j}^{(l,k)}(0)\right]^L$. Specifically $\xi_{i,t}^{(k)} = \xi_{i,t}^{(k)*} / \sum_{K=1}^{K}\sum_{j=1}^{t} \xi_{j,t}^{(k)*}$. Then by (4),

$$P(s_t = k|Y_{1,t}) = \sum_{i=1}^{t} \xi_{i,t}^{(k)},\quad E(\theta_{lt}|Y_{1,t}) = \sum_{k=1}^{k}\sum_{i=1}^{t} \xi_{i,t}^{(k)} z_{i,t}^{(l,k)}. \qquad (6)$$

Smoothing estimate

Our model assumptions imply that the stationary distribution of θ_{lt} exists and is given by $\sum_{k=1}^{k} \pi_k N(z^{(l,k)}, v^{(l,k)})$. This indicates that, if θ_{lt} is initialized at its stationary distribution, its time-reversed Markov chain can be defined. This substantially simplifies the smoothing estimates of θ_{lt}. Actually, it further implies that $\{\theta_{lt}\}$ is a reversible Markov chain,

so we can reverse time and obtain a backward filter that is analogous to (4):

$$\theta_{l,t+1}\,|\,y_{t+1,T} \sim \sum_{K=1}^{K}\sum_{j=t+1}^{T}\eta_{t+1,j}^{(k)}N(z_{t+1,j}^{(1,k)},v_{t+1,j}^{(l,k)}), \tag{7}$$

Where the mixture weight $\eta_{t+1,j}^{(k)}$ is given by $\sum_{K=1}^{K}\sum_{j=t+1}^{T}\eta_{t+1,j}^{(k)}=1$ and

$$\eta_{i,t,j}^{(k)}\,\alpha\,\eta_{t+1,j}^{(k)*} := \begin{cases} (\sum_{r\neq k}\eta_{t+2}^{(t)}\tilde{p}_{rk})\psi_{0,0}^{(k)}/\psi_{t+1,t+1}^{(k)} & \text{if } j=t+1, \\ \tilde{p}_{kk}\eta_{t+2,j}^{(k)}\psi_{t+2,j}^{(k)}/\psi_{t+1,j}^{(k)} & \text{if } j>t+1, \end{cases} \tag{8}$$

in which $\tilde{p}_{rk} = P(s_t = k|s_{t+1} = r) = p_{kr}\pi_k / \pi_r$ is the transition probability of the reversed chain of $\{s_t\}$. Since $p(\theta_{tt} \in A|Y_{t+1,T}) = \int p(\theta_{tt} \in A|\theta_{l,t+1})dP(\theta_{l,t+1}|Y_{t+1,T})$, it follows from (8) and the reversibility of $\{\theta_t\}$ that

$$\theta_{lt}|Y_{t+1,T} \sim \sum_{k=1}^{k}\left\{\tilde{p}_{kk}\sum_{j=t+1}^{T}\eta_{t+1}^{(k)}N(z_{t+1,j}^{(l,k)},v_{t+1,j}^{(l,k)}) + (\sum_{r\neq k}\tilde{p}_{rk}\eta_{t+1}^{(t)})\,N(z^{(l,k)},v^{(l,k)})\right\} \tag{9}$$

Next, we shall use Bayes' theorem to combine the forward filter (4) with its backward variant (9) to estimate θ_{lt} given Y_T $(1 \le t < T)$, $(1 \le l < J)$, which is expressed as the following mixture of normal distributions

$$\theta_{lt}|Y_T \sim \sum_{k=1}^{k}\sum_{1\le i\le t\le j\le T}\alpha_{ij,t}^{(k)}N(z_{t,j}^{(l,k)},v_{t,j}^{(l,k)}), \tag{10}$$

In which the mixture weights $\alpha_{ij,t}^{(k)}$ are posterior probabilities explained below. Consider the $C_{ij}^{(k)} = \{s_i = ... = s_j = k, s_i \neq s_{i-1}, s_j \neq s_{j+1}\}$, Web Appendix A shows that, for $i \le t \le j$, $\alpha_{ij,t}^{(k)} = P\left(c_{ij}^{(k)}|Y_{1,T}\right)$ and $\alpha_{ij,t}^{(k)}$ an be calculated recursively as follows:

$$\alpha_{ij,t}^{(k)} = \alpha_{ij,t}^{(k)*}/D_t, \quad D_t = \sum_{k=1}^{k}\sum_{1\le i\le t\le j\le T}\alpha_{ij,t}^{(k)*},$$

$$\alpha_{ij,t}^{(k)*} = \begin{cases} \xi_{i,t}^{(k)}(\sum_{r\neq k}\eta_{t+2}^{(t)}p_{kr}/\pi_r) & i\le t=j, \\ p_{kk}\xi_{i,t}^{(k)}\eta_{t+1}^{(k)}\psi_{i,j}^{(k)}\psi_{0,0}^{(k)}/\left(\pi_k\psi_{i,t}^{(k)}\psi_{t+1,j}^{(k)}\right) & i\le t<j. \end{cases} \tag{11}$$

Therefore, from (10), it follows that

$$E(\theta_{tt}|Y_{1,T}) = \sum_{k=1}^{k}\sum_{1\le i\le t\le j\le T}\alpha_{ij,t}^{(k)}z_{i,j}^{(l,k)} \tag{12}$$

$$f(\theta_{lt}|Y_{1,t})\,f(\theta_{lt}|Y_{t+1,n})\,f(\theta_{lt}|Y_{tt}) \tag{13}$$

Figure 1 provides a schematic explanation for the above algorithm. For location t, the algorithm first decomposes conditional distributions $f(\theta_{tt}|Y_{1,t})$ and $f(\theta_{tt}|Y_{t+1,n})$ based on the most recent switching location of s_t before and after t, then use Bayes theorem to combine these two distributions and obtain $f(\theta_{tt}|Y_{tt})$.

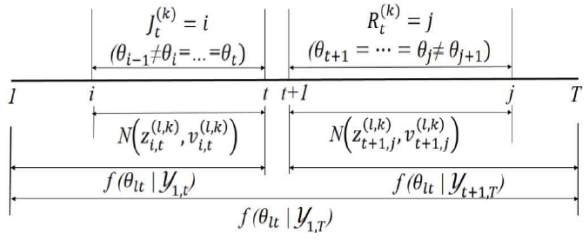

Figure 1: A schematic plot of forward-backward algorithm.

Bounded complexity mixture approximations and hyper parameter estimation

The number of mixture weights in the above discussion increases dramatically with t (or n), resulting in rapidly increasing computational complexity and memory requirements in estimating θ_{lt} as t keeps increasing. To address the issue of computational efficiency, we follow Lai and Xing and use a bounded complexity mixture (BCMIX) approximation procedure with linear computational complexity; see Web Appendix B for details of the algorithm.

The inference procedures involve the hyper parameters p, probability transition matrix P, and $\{z^{(l,k)}, v^{(l,k)}, \sigma_l^2; 1 \le k \le K, 1 \le l \le L\}$. In practice, these *[(K − 1) K + (2K + 1) L + 1]* parameters are unknown and should be replaced by their estimates. We consider an EM algorithm to estimate all hyper parameters with the details given in Web Appendix C.

In practical applications, we should also notice that the three categorical states in the above model are exchangeable; hence the categorical states s_t could be very difficult to identify. A remedy for this is to replace the normal priors for θ_{lt} by truncated priors, then the filtering and smoothing formulas in Sections 2.2 and 2.3 needs to be modified somewhat. Specifically, the normal distribution in conditional distribution (10) needs to be replaced by corresponding truncated normal distributions. Another way to mitigate the identification issue is to put constraints on hyper parameters. For example, a prior normal distribution with smaller variances could limit the estimated quantitative signals staying around its prior mean, so the distinction between the categorical states becomes clearer.

Simulation Studies

We now perform intensive simulations to evaluate the performance of the proposed model and inference procedure from frequentist and Bayesian viewpoints. To demonstrate the performance, we consider two measures for different purpose. One measure is mean square error (MSE), which provides the mean errors between the estimates $\hat{\theta}_{lt}$ and the true θ_{lt}, i.e., MSE:= $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{l=1}^{L}(\theta_{lt}-\hat{\theta}_{lt})^2$ and the other is mean identification rate (IR), which evaluates the accuracy of our inference on the hidden states s_t. As our model only computes the posterior probability of s_t given $Y_{1,T}$, we estimate the state of location t as the one which maximizes the posterior probability of being in categorical state k, i.e., $\hat{s}_t := \text{argmax}_k\{P(s_t = k|Y_{1,T})\}$ With the above estimate, we define the mean identification ratio as IR:=$\frac{1}{T}\sum_{t=1}^{T}1_{\{s_t=\hat{s}_t\}}$.

We first evaluate the performance of BCMIX estimates in the frequentist setting by considering the following four cases of hidden state $\{s_t\}$ with K=3.

Case 1. $s_t = 1.1_{\{1\le t\le 200\}} + 2.1_{\{201\le t\le 400\}} + 3.1_{\{401\le t\le 1000\}}$.
Case 2. $s_t = 1.1_{\{1\le t\le 500\}} + 2.1_{\{501\le t\le 700\}} + 3.1_{\{701\le t\le 1000\}}$.
Case 3. $s_t = 2.1_{\{1\le t\le 200,401\le t\le 600,801\le t\le 1000\}} + 1.1_{\{201\le t\le 400\}} + 3.1_{\{601\le t\le 800\}}$.
Case 4. $s_t = 2.1_{\{1\le t\le 200,401\le t\le 600,801\le t\le 1000\}} + 3.1_{\{201\le t\le 300,601\le t\le 700\}} + 1.1_{\{301\le t\le 400,701\le t\le 800\}}$.

Given the above $\{s_t\}$, we generate θ_{lt} by assuming $\theta_{lt} \sim N(z^{(l,st)}, v^{(l,st)})$ with $(z^{(l,1)}, z^{(l,2)}, z^{(l,3)}) = (1, 0, -1)$, $v^{(l,1)} = v^{(l,2)} = v^{(l,3)} = 0.22$ (hence the standard deviation is about 0.47). We further assume L=10, T=1000, and generate observations y_{lt} by (1) and $\sigma_l^2 = 1\ (l = 1, \ldots, L)$. We then use the EM

algorithm to estimate the hyper parameters and compute the BCMIX estimates with *(M, m)=(10, 5), (20, 10), (30, 15) and (40, 20)*. For comparison purpose, we also compute oracle estimate which assume $\{s_t\}$ is known. We then run such simulation for each case 500 times, and summarize the MSE of two estimates and corresponding standard errors (in parentheses) in Web Table 1. We can see that the oracle and BCMIX estimates are quite similar, and the difference among BCMIX estimates with different *(M, m)* are quite small. Therefore, we will use BCMIX estimate with *(M, m)=(20, 10)* in the following discussion.

We then evaluate the performance of the inference procedure under our model assumption. Let K= 3, $(z^{(l,1)}, z^{(l,2)}, z^{(l,3)})=(1, 0, -1)$, $v^{(l,1)}=v^{(l,2)}=v^{(l,3)}=0.16$, and $\sigma_l^2 = 1$ for $1 \le l \le L=10$. The probability transition matrix $\left(p_{ij}\right)_{1\le i,j\le K}$ of $\{s_t\}$ is assumed to follow nine scenarios. Specifically, for Scenarios S_k, $k = 1, 2, \ldots, 5$, we let $p_{ij}=0.001\times 2^{k-1}$ for $1 \le i \ne j \le 3$. For Scenario S_6, $(p_{12}, p_{13}, p_{21}, p_{23}, p_{31}, p_{32})=(0.002, 0.001, 0.002, 0.002, 0.001, 0.002)$. For Scenario S_7, $(p_{12}, p_{13}, p_{21}, p_{23}, p_{31}, p_{32})=(0.004, 0.001, 0.004, 0.004, 0.001, 0.004)$. For Scenario S_8, $(p_{12}, p_{13}, p_{21}, p_{23}, p_{31}, p_{32})=(0.001, 0.002, 0.001, 0.001, 0.001, 0.001)$. For Scenario S_9, *(p12, p13, p21, p23, p31, p32)=(0.001, 0.004, 0.001, 0.001, 0.001, 0.001)*. For each scenario, we first generate samples of different lengths with *T=3000, 5000, 7000,* then use the proposed EM algorithm to estimate the hyper parameters and estimate θ_{it} and $P(s_t|Y_{1,T})$. Web Tables 2 and 3 summarizes the MSE and IR of our estimate, and also provided in parentheses are the corresponding standard errors based on 500 simulations in each cell. We can see that the MSE are very small, and the IR is all larger than 84%.

Since data generation procedures in above studies do not deviate from our model assumption too much, to show the variability of our model, we also evaluate the performance of our algorithm on the data in Willenbrock and Fridly and [25], which are generated from a completely different procedure, and compute the MSE between the estimates $\hat{\theta}_{lt}$ and the true signals θ_{lt}. The MSE of 100 datasets with 20 samples and each sample with 500 clones on Chromosome 1 is 0.011 with standard error *5.89e-4,* indicating the estimates for signals in individual samples is still pretty good. Web Figure 1 demonstrates a randomly selected simulated y_{lt} and $\hat{\theta}_{lt}$ for 20 samples.

We next compare our model to the hierarchical hidden Markov model (HMM) in Shah et al. [7]. Specifically, we estimate all hyper parameters by the EM algorithm, and then fit the hierarchical HMM model to the simulated data generated in Scenarios S_1-S_9. Since Shah et al. assume the signal levels θ_{lt} of individual CNAs follow a normal distribution with the mean and variance depending on the hidden state s_t directly; it implies that the individual signal levels are fixed within the same segment of recurrent CNA regions. This is different from our model which allows signal levels θ_{lt} of sample l have different values at different locations *t* even if their categorical states s_t are same, which is more realistic in practice. Furthermore, our algorithm avoids the use of Markov Chain Monte Carlo algorithm, hence computationally is very fast. We run all simulations on a desktop with Intel core *i5-3210M* and 4G memory, for each simulation of 10 samples with sample length *T=3000, 4000, 5000 and 6000,* our algorithm takes about 2.8-6 seconds to obtain the estimates for θ_{lt} and s_l, while the hierarchical HMM model takes over 10-20 minutes to get its estimates. Web Table 3 summarizes the identification ratios and the corresponding standard errors (in parentheses) using Shah et al.'s model for different settings. Each cell is based on 100 simulations. We shall note that the hidden states s_t in our setting are very close to each other due to the large signal-to-noise ratios, hence it is not easy to make a correct state calling. The identification ratios of the hierarchical HMM are very good (all their ratios are about 70%), but they are typically smaller than ours.

Real data studies

Analysis for Ovarian cancer data

Ovaries are reproductive organs that produce eggs in women, and ovarian cancer is the fifth leading cause of cancer death in women. Ovarian cancers display a high degree of complex genetic variations. The existing literature show that the most frequently affected chromosomes in ovarian cancer are chromosome 1, 8, and 17. We use our model to analyze the copy number alteration (CNA) data for Ovarian serous cystadenocarcinoma (OV) based on Array based-CGH technology. The data in our analysis, consisting of the CNAs from 15 OV cancer patients, were published on April 1st, 2010 in the Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/). We analyze the whole 23 chromosomes. Since the existing literature shows that the most frequently affected chromosomes in ovarian cancer are chromosomes 1, and 17, we only present our result on these two chromosomes.

There are 55,274 and 20,009 probes on chromosomes 1 and 17. Let *k=1, 2 and 3* denote amplification, baseline and deletion, respectively. We first use the proposed model and inference procedure to estimate the hyper parameters, and then the signal levels θ_{lt} and probability of s_t for chromosomes 1 and 17. We run our model on a desktop with Intel core *i5-3210M* and 4G memory, and it takes 223 and 109 seconds for chromosomes 1 and 17, respectively. The results are summarized in Web Figures 2 and 3, respectively. We can see that our procedure analyzes all samples and estimates signal θ_{lt} for each sample simultaneously, which avoid the weakness of two-stage analysis. As our interest here is the recurrent CNA region, we now focus on the estimated probabilities of s_t, which are plotted in Figure 2. Those estimated probabilities indicate that the recurrent copy number amplifications involve regions *1p34.2, 1p12, 1q23.2 and 1q42.3,* and deletions involve regions *1p36.33, 1p36.21, 1p36.13, 17p11.2 and 17p12.* Well known tumor suppressor genes *TP73 (1p36.33), TP53 (17p13.1), BRCA1 (17q21),* oncogene *MYCL1 (1p34.2),* and transcription factors *RAI1 (17p11.2), SREBF1 (17p11.2)* are found recurrent regions of copy number variants. Our results are consistent with earlier studies [26-28]. It is important to note that for chromosome 17, we focus on detecting the recurrent regions of

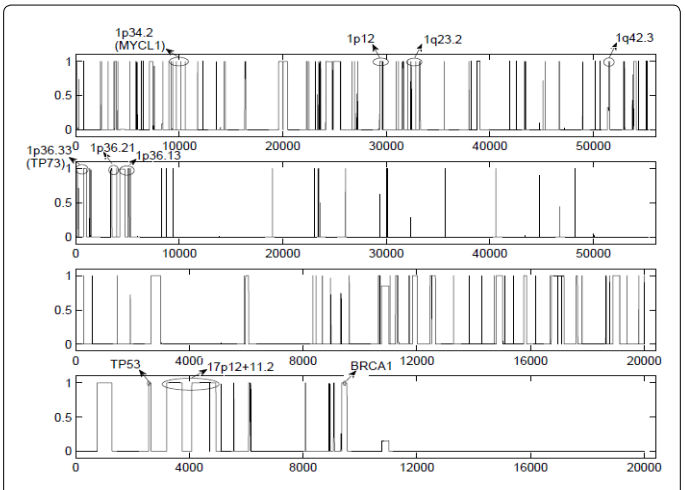

Figure 2: Estimated P (s_t=k|$Y_{1,T}$) of chromosomes 1 (The top two) and 17 (The bottom two) for k=amplification (The 1st and 3rd panels) and deletion (The 2nd and 4th panels).

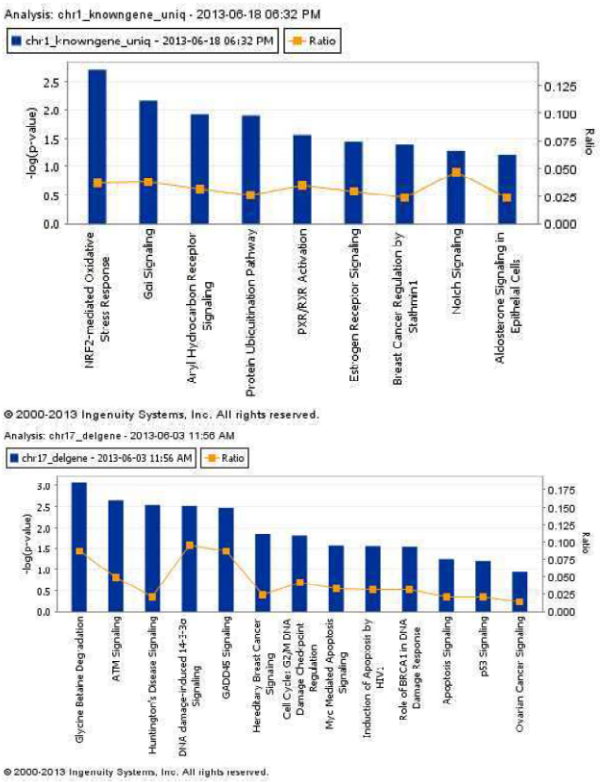

Figure 3: Canonical pathway analysis of detected genes on chromosomes 1 (top) and 17 (bottom).

copy number deletions, since the most common alterations for serous histology of OV cancer are deletions of 17p [26,28,29].

There are totally 178 and 136 unique known genes involving in recurrent CNA regions for chromosomes 1 and 17 respectively. These known genes are subjected to pathway exploration using the Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, Redwood City, CA). Significantly enriched pathways with Fishers exact P-values less than 0.05 are listed in this bar plot as shown in Figure 3. Yellow square line in the figure represents ratio which is the number of focus genes in the pathway divided by the total number of genes that make up that pathway. For chromosome 1, most of the pathways are related to cancer. Notably, as listed on the 7th, the breast cancer signaling was found enriched. Furthermore, a few hormone metabolism pathways are involved, which includes PXR/RXR activation, Estrogen receptor signaling and Aldosterone signaling in Epithelial Cells. This is consistent with current knowledge of disrupted hormone metabolism pathways as important causal factors in breast cancer [30-32]. In addition, a few important cellular pathways are revealed: The NRF2-mediated oxidative stress response turns out to be most significantly changed in the list, which has been related to breast cancer. The G-protein signaling pathway, which is well-known to be related to cancer, is listed on the second. A basic transcription factor related pathway is ranked on the 3rd. And listed on the 4th, Ubiquitination regulates degradation of cellular proteins by the ubiquitin proteasome system, controlling a proteins half-life and expression levels. A change of ubiquitination activity is associated with ovarian tumorigenesis, so the protein ubiquitination pathway might be involved in breast ovarian progression. Finally, one of the most important developmental pathway in mammals, Notch signaling also known to play a role in cancer [33]. For chromosome 17, the pathway enrichment result reveals some biological mechanisms and pathway changes involved in ovarian cancer. First obviously, the ovarian cancer signaling pathway was found enriched. Particularly, the GADD45 and p53 signaling pathways are enriched. Both these two factors, especially p53, are well established tumor suppressor proteins. More importantly, almost half of the pathways are basic and critical cellular processes such as DNA repair, cell cycle regulation and apoptosis. Changes in these pathways indicate severe disruptions of normal cellular functions. This could be either the cause or the result of cancer.

Analysis for lung cancer data

There are two main types of lung cancer, small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC) [34]. NSCLC is the most common type of lung cancer, accounting for about 85% of total lung cancers. NSCLC is mainly comprised of adnenocarcinoma, squamous cell carcinoma and large cell carcinoma. About 30% of lung cancers are squamous cell carcinoma. Previous cancer studies have revealed that multiple tumor suppressor genes are involved in deletions at multiple chromosomal regions in lung carcinogenesis, and the most frequent deletions in lung cancer tissues are at chromosome 3, 13 and 17.

We use our model to analyze the CNA data for Lung squamous cell carcinoma (one type of non-small cell lung cancer) based on Array based-CGH technology. The data used in our study, consisting of the CNAs from 10 cancer patients, were published on October 22nd, 2010 in the Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/). We analyze the whole 23 chromosomes. Since the existing literature shows that the most frequently affected chromosome in this type of lung cancer is chromosomes 17, we present our result on chromosome 17.

There are totally 13,575 probes on chromosome 17. Let k=1, 2 and 3 denote the amplification, baseline and deletion. We first use the proposed model and inference procedure to estimate hyper parameters and then fit the model to the data. We run our model on a desktop with Intel core i5-3210M and 4G memory, and it takes 30 seconds for chromosome 17. The estimated signal levels θ_{lt} for chromosomes 17 are summarized in Web Figure 4. We can see that our procedure analyzes

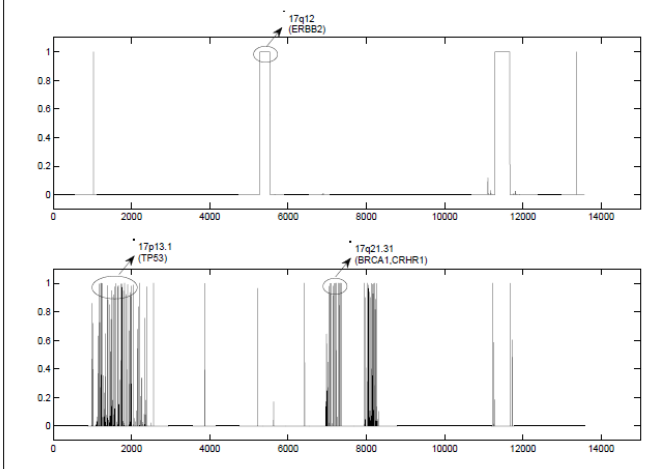

Figure 4: Estimated P (s_t=k|$Y_{1,T}$) of chromosomes 17 for k=amplification (The 1st panels) and deletion (The 2nd and 4th panels).

all samples and estimates signal θ_{lt} for each sample simultaneously, which avoids the weakness of two-stage analysis. As our interest here is the recurrent CNA region, we now focus on the estimated probabilities of s_t, which are plotted in Figure 4. Those estimated probabilities indicate that the recurrent copy number amplifications at long arm of chromosome 17, which contains the oncogene ERBB2 at 17q12. Two regions of deletion can be found at short arm and long arm of chromosome 17 respectively, which contain the well-known tumor suppressor genes TP53 at 17p13.1, BRCA1 and CRHR1 at 17q21.31. Our results are consistent with earlier studies [35,36].

Conclusions

We have developed a stochastic segmentation model and an associated inference procedure for recurrent CNA data. The model implies explicit recursive formulas for both the posterior distribution of individual samples' signal levels and the probabilities of the cross-sample recurrent events at each probe. This further suggests the estimate of the recurrent states of CNAs. To speed up the computation for practical purpose, an approximation to the exact explicit formulas is developed, and the computational complexity is reduced to linear order. Estimation of hyper parameters involves an explicit EM algorithm which is described in the Web Appendix D.

In Section 4, we have analyzed two real datasets to illustrate the application of our model. In particular, we identify the recurrent CNAs regions using the copy number data for ovarian serous cystadenovarcinoma and non-small lung cancer carcinoma that are produced by the array-CGH technology. The estimated CNA regions by our model are consistent with the biological discovery in medical study. For ovarian serous cystadenovarcinoma, we further perform a canonical pathway analysis to evaluate our result, and find our pathway enrichment results yield significant pathways and most of them are cancer related pathways. Our result based on chromosomes 1 and 17 already reveals certain biological mechanisms and pathway changes involved in ovarian cancer. These facts demonstrate that our model can successfully capture recurrent CNA regions and generate promising results in biological context.

Acknowledgment

The first author was supported by National Science Foundation DMS-0906593 and DMS-1206321. We thank gratefully the associate editor and two anonymous referees for their constructive comment on how to improve this paper.

References

1. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, et al. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 20: 207-211.
2. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. Genome Research 16: 1136-1148.
3. Pinkel D, Albertson DG (2005) Array comparative genomic hybridization and its applications in cancer. Nature Genetics 37: 11-17.
4. Rueda OM, Diaz-Uriarte R (2007) Flexible and accurate detection of genomic copy number changes from aCGH. PLoS Computational Biology 3: e122.
5. Rueda OM, Diaz-Uriarte R (2010) Finding Recurrent Copy Number Alteration Regions: A Review of Methods. Current Bioinformatics 5: 1-17.
6. Shah SP (2009) Computational methods for identification of recurrent copy number alteration patterns by array CGH. Cytogenetic and Genome Research 123: 343-351.
7. DeLeeuw RJ, Davies JJ, Rosenwald (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. Human Molecular Genetics 13: 1827-1837.
8. Garnis C, Lockwood WW, Vucic E, Ge Y, Girard L, et al. (2006) High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. International Journal of Cancer 118: 1556-1564.
9. Rouveirol C, Stransky N, Hup´Ep, LaRosa P, Viara E (2006) Computation of recurrent minimal genomic alterations array-CGH data. Bioinformatics 22: 849-856.
10. Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. Genome Research 16: L1149-1158.
11. Guttman M, Mies C, Dudycz-Sulicz K, Diskin SJ, Baldwin DA, et al. (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. PLoS Genetics 3: e143.
12. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, et al. (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to Glioma. Proceedings of the National Academy of Sciences 104: 20007-20012.
13. Morganella S, Pagnotta SM, Ceccarelli M (2011) Finding recurrent copy number alterations preserving within-sample homogeneity. Bioinformatics 27: 2949-2956.
14. Walter V, Nobel AB, Wright FA (2011) DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. Bioinformatics 27: 678-685.
15. Lipson D, Aumann Y, Ben-Dor A, Linial N, Yakhini Z (2006) Efficient calculation of interval scores for DNA copy number data analysis. Journal of Computational Biology 13: 215-228.
16. Liu Q, Zhang H, Smeester L, Zou F, Kesic M (2010) The NRF2-mediated oxidative stress response pathway is associated with tumor cell resistance to arsenic trioxide across the NCI-60 panel. BMC medical genomics 3: 37.
17. Shah SP, Lam WL, Ng RT, Murphy KP (2007) Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics 23: i450-i458.
18. Guha S, Li Y, Neuberg D (2008) Bayesian hidden Markov modeling of array CGH data. Journal of the American Statistical Association 103: 485-497.
19. Klijn C, Holstege H, de Ridder J, Liu X, Reinders M (2008) Identification of cancer genes using a statistical framework for multi experiment analysis of no discretized array CGH data. Nucleic Acids Research 36: e13-e13.
20. VanDyk E, Reinders MJT, Wessels LFA (2013) A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. Nucleic acids research 41: e100-e100.
21. Zhang Q, Ding L, Larson DE, Koboldt DC, McLellan MD (2010) CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. Bioinformatics 26: 464-469.
22. Lai, TL, Xing H, Zhang N (2008) Stochastic segmentation models for array-based comparative genomic hybridization data analysis. Biostatistics 9: 290-307.
23. Lai TL, Xing H (2011) A simple Bayesian approach to multiple change-points. Statistica Sinica 21: 539-569.
24. Xing H, Mo Y, Liao W, Zhang MQ (2012) Genome-wide localization of protein-DNA binding and histone modification by BCP with ChIP-seq data. PLoS Computational Biology 8: e1002613.
25. Willenbrock H, Fridlyand J (2005) A comparison study: applying segmentation toarray CGH data for downstream analyse. Bioinformatics 21: 4084-4091.
26. Dimova I, Orsetti B, Negre V, Rouge C, Ursule L, et al. (2009) Genomic markers for ovarian cancer at chromosomes 1, 8 and 17 revealed by array CGH analysis. Tumori 95: 357.
27. Engler DA, Gupta S, Growdon WB, Drapkin RI, Nitta M, et al. (2012) Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. PloS one 7: 30996.
28. Zhang J, Shi Y, Lalonde E, Li L, Cavallone L (2013) Exome profiling of primary, metastatic and recurrent ovarian carcinomas in a BRCA1-positive patient. BMC cancer 13: 146.
29. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. PLoS One 6: e24709.
30. Davis DL, Telang NT, Osborne (1997) Medical hypothesis: bi functional genetic-

hormonal pathways to breast cancer. Environmental Health Perspectives 105: 571.

31. Beckmann L, Husing A, Setiawan VW, Amiano P, Clavel-Chapelon F, et al. (2011) Comprehensive analysis of hormone and genetic variation in 36 genes related to steroid hormone metabolism in pre-and postmenopausal women from the breast and prostate cancer cohort consortium (BPC3). Journal of Clinical Endocrinology and Metabolism 96: E360-E36.

32. Clendenen T, Zeleniuch-Jacquotte A, Wirgin I, Koenig KL, Afanasyeva Y (2013) Genetic Variants in Hormone-Related Genes and Risk of Breast Cancer. PloS one 8: e69367.

33. Hu Y, Zheng M, Zhang R, Liang Y, Han H (2012) Notch signaling pathway and cancer metastasis. Notch Signaling in Embryology and Cancer 727: 186-198.

34. Pass HI, Mitchell JB, Johnson DH, Turrisi AT, Minna JD (1996) Lung Cancer: Principles and Practice. Lippincott-Raven Publishers, Philadelphia, PA.

35. Clark J, Edwards S, Feber A, Flohr P, John, et al. (2003) High-resolution analysis of DNA copy number using oligonucleotide Genome-wide screening for complete genetic loss in prostate cancer by comparative hybridization onto cDNA microarrays. Oncogene 22: 1247-1252.

36. Kohno T, Yokota J (1999) How many tumor suppressor genes are involved in human lung carcinogenesis? Carcinogenesis 20: 1403-1410.