Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 3

A Bayes Regression Approach to Array-CGH Data

Chi-Chung Wen, *National Health Research Institutes, Taiwan*
Yuh-Jenn Wu, *National Health Research Institutes, Taiwan*
Yung-Hsiang Huang, *National Health Research Institutes, Taiwan*
Wei-Chen Chen, *National Health Research Institutes, Taiwan*
Shu-Chen Liu, *National Health Research Institutes, Taiwan*
Shih Sheng Jiang, *National Health Research Institutes, Taiwan*
Jyh-Lyh Juang, *National Health Research Institutes, Taiwan*
Chung-Yen Lin, *National Health Research Institutes, Taiwan*
Wen-Tsen Fang, *National Health Research Institutes, Taiwan*
Chao Agnes Hsiung, *National Health Reserach Institutes, Taiwan*
I-Shou Chang, *National Health Research Institutes, Taiwan*

Copyright ©2006 The Berkeley Electronic Press. All rights reserved.

A Bayes Regression Approach to Array-CGH Data*

Chi-Chung Wen, Yuh-Jenn Wu, Yung-Hsiang Huang, Wei-Chen Chen, Shu-Chen Liu, Shih Sheng Jiang, Jyh-Lyh Juang, Chung-Yen Lin, Wen-Tsen Fang, Chao Agnes Hsiung, and I-Shou Chang

Abstract

This paper develops a Bayes regression model having change points for the analysis of array-CGH data by utilizing not only the underlying spatial structure of the genomic alterations but also the observation that the noise associated with the ratio of the fluorescence intensities is bigger when the intensities get smaller. We show that this Bayes regression approach is particularly suitable for the analysis of cDNA microarray-CGH data, which are generally noisier than those using genomic clones. A simulation study and a real data analysis are included to illustrate this approach.

KEYWORDS: change point problem, comparative genomic hybridization, DNA copy number imbalance

---

*Part of this work is supported by NSC grant NSC 94-3112-B-400-002-Y. The last author is the corresponding author; the first and the second authors contribute equally to this work. We are grateful to Dr. A. Olshen for helpful conversations during the preparation of this paper. We are also grateful to two referees, whose comments led to improvement of the paper.

1 INTRODUCTION

The application of comparative genomic hybridization(CGH)s to metaphase spreads enabled genome-wide analysis of gross DNA copy number imbalance, which are key genetic events in the development and progression of human cancers. It makes use of the ratio of the fluorescence intensities detected from the differentially labeled 'test' and 'reference' DNA, which are competitively hybridized to normal metaphase chromosomes on a glass slide, to study DNA copy number imbalances. Although array-CGH is based on the same principle as metaphase-CGH, except that the targets are mapped genomic clones or cDNA clones instead of whole chromosomes, it provides more reliable measurements of DNA copy number aberrations and has received much attention in the past years. A review of the current array-CGH technology is given in Mantripragada et al. (2004).

Several methods have been proposed to study DNA copy number changes based on array-CGH data. For example, Hodgson et al. (2001) proposed a three-component normal mixture model and a maximum likelihood approach, Pollack et al. (2002) devised a threshold method and provided an estimate of the false discovery rate (FDR) for the result, and Cheng et al. (2003) described a regression-based statistical method to test for the changes in copy numbers, among others. A common drawback of these methods is that they do not take into consideration of the fact that DNA copy number changes occur in contiguous regions of a chromosome, which often cover multiple markers or even whole chromosome.

Recently, there appeared methods that utilize the above mentioned spatial structure of the genomic alterations. For example, Jong et al. (2003) and Autio et al. (2003) proposed to divide the genome to regions and then assign each region to one of the three states: normal, gain, and lose, Olshen et al. (2004) developed the circular binary segmentation method to translate noisy intensity measurements into regions

Produced by The Berkeley Electronic Press, 2006

of equal copy number, Fridlyand et al. (2004) developed an unsupervised hidden Markov model approach to study the number and types of DNA copy number aberrations, and Wang et al. (2005) proposed an algorithm 'Cluster along chromosomes' (CLAC) for the analysis of array CGH data.

The purpose of this paper is to develop a Bayes regression approach for the analysis of array-CGH data by utilizing not only the underlying spatial structure of the genomic alterations but also the observation that the noise associated with the ratio of the fluorescence intensities is larger when the intensities get smaller. We indicate that this Bayes regression approach is particularly relevant for the analysis of cDNA microarray-CGH data, which are generally noisier than those using genomic clones. Although it is desirable to study the gain or loss at each locus in the test sample, the focus of this paper is on the easier problem of looking for change points, where the DNA copy number on the right of the point is different from that on the left.

This paper is organized as follows. Section 2 presents a Bayes regression model in which the regression function is a step function, and a Markov chain Monte Carlo method for its inference. Section 3 examines this Bayes method in the simulation studies and reports that it is indeed beneficial to exploit the noise structure when existing. Sections 4 illustrates the method in the analysis of a real array-CGH dataset and reports that most of the jump points suggested by the Bayes method are validated by QRT-PCR studies. Section 5 is a discussion on future investigations.

2 A BAYES REGRESSION MODEL HAVING CHANGE POINTS

2.1 The model

We fix one chromosome and consider all the clones on this chromosome. For the ith clone, let X_i denote its location on this chromosome, where the location is measured in physical distance, Y_i denote the logarithm of the ratio of its fluorescence intensity

of the test sample to that of the reference sample, and A_i denote the product of its fluorescence intensities of the test sample and the reference sample.

Let $[0, \tau]$ denote the length of the chromosome. Then X_i is in $[0, \tau]$. We assume that the DNA copy number of the chromosome in the reference sample is always two and that in the test sample is a step function with finitely many jump points, defined on $[0, \tau]$. Hence the log ratio of the DNA copy number in the test sample to that in the reference sample is also a step function. This unknown step function represents the DNA copy number imbalances and the purpose of this paper is to propose a Bayesian estimate of this unknown step function, based on the data $\{(Y_i, A_i)|i = 1, \cdots, n\}$.

We first parametrize a space of step functions and then introduce a prior on it. For $m \geq 1$, let $B_m = \{b_m = (\tau_1, ..., \tau_m; \lambda_0, ..., \lambda_m) \mid 0 = \tau_0 < \tau_1 < ... < \tau_m < \tau, (\lambda_0, ..., \lambda_m) \in \mathbf{R}^{m+1}\}$ and $\mathcal{B} = \bigcup_{m=1}^{\infty}(\{m\} \times B_m)$. For each (m, b_m) in $\mathcal{B}$, we define the step function

$$F(b_m; x) = \sum_{i=0}^{m-1} I_{(\tau_i, \tau_{i+1}]}(x)\lambda_i + I_{(\tau_m, \tau]}(x)\lambda_m, \tag{1}$$

which may be abbreviated as $F_{(m,b_m)}(x)$ or $F_{b_m}(x)$.

We assume that there are a known transformation η and an $i.i.d.$ sequence $\epsilon_1, \epsilon_2, \cdots$ with known density g such that, given the parameter value (m, b_m), the random variables $(Y_1, A_1), (Y_2, A_2), \cdots$ satisfy

$$Y_i = F_{(m,b_m)}(X_i) + \frac{\epsilon_i}{\eta(A_i)}, \tag{2}$$

for every $i = 1, 2, \cdots$. We also assume that $\{A_i|i = 1, 2, \cdots\}$ and $\{\epsilon_i|i = 1, 2, \cdots\}$ are independent.

Let $\mathcal{S}$ denote the space of all step functions on $[0, \tau]$ having finitely many jump points. A prior on $\mathcal{S}$ can be easily introduced through the mapping from $\mathcal{B}$ to $\mathcal{S}$ defined by (1) and a probability distribution on $\mathcal{B}$. We study only priors introduced in this way and also call a probability distribution on $\mathcal{B}$ a prior. We note that a

natural way to introduce a prior π on $\mathcal{B}$ is to specify $p(m) = \pi(\{m\} \times B_m)$, with $p(m) \geq 0$ and $\sum_{m=1}^{\infty} p(m) = 1$, and the conditional density of π on $\{m\} \times B_m$, denoted by $\pi_m(\cdot \,|\{m\} \times B_m) = \pi_m(\cdot)$. The conditional density $\pi_m(\cdot)$ can be regarded as a density on B_m.

Thus the sampling distribution for the data $\{(Y_i, A_i) \,|\, i = 1, \cdots, n\}$, given the parameter value (m, b_m), is

$$\prod_{i=1}^{n} g(\eta(A_i)(Y_i - F_{b_m}(X_i))),$$

and the posterior density ν at the parameter (m, b_m) given the data is proportional to

$$\left[\prod_{i=1}^{n} g(\eta(A_i)(Y_i - F_{b_m}(X_i)))\right]\left[\pi_m(b_m)p(m)\right]. \tag{3}$$

We note that, although g is assumed known in this paper, the method of this paper can be easily extended to treat the case that g has certain parametric form with a prior on the parameters.

We note that if A_i is a constant not varying with $i = 1, 2, \cdots$ and the function F_{b_m} is not necessarily constant between the jumps, then (2) is the regression model with discontinuities studied by Korostelev (1987), Yin (1988), Muller (1992), Wu and Chu (1993), Oudshoorn (1995), Wang (1995), and Spokoiny (1998), among others.

We also note that if the data is so indexed that $X_i \leq X_j$ for $i \leq j$, then (2) is related to the change-point problems studied by Chernoff and Zacks (1964), Yao (1984), Barry and Hartigan (1992, 1993), and Lee (1998), among others. Although some of these authors also adopted Bayesian approach, our modeling and analysis method are specifically motivated by array-CGH studies and take advantage of both the spatial structure and the intensity dependent noise structure in the data.

2.2 Bayesian inference

Viewing the prior as a probability distribution defined on $\mathcal{B}$, which is the union of spaces of different dimension, we propose to adapt the reversible jump Metropolis-Hastings algorithm (Green, 1995) to generate the posterior distribution (3) for inference purpose and consider the posterior mode the estimate.

In order to illustrate this algorithm, we consider the following simple prior. Using the above notations, we assume that $p(m) \propto \alpha^m/(m!)^2$ for $m \geq 1$ and $\pi_m(\tau_1, \cdots, \tau_m, \lambda_0, \lambda_1, \cdots, \lambda_m) = m!(\frac{1}{\tau})^m \cdot \prod_{j=0}^{m} \frac{exp(-(\lambda_j-\mu)^2/2d)}{\sqrt{2\pi d}}$, with known α, μ and d. We note that this can be viewed as a continuous analogue of the product partition model of Barry and Hartigan (1992). The reason that we consider $(m!)^2$, instead of $(m!)$, is to make sure the posterior mode exists.

Reversible jump Metropolis-Hastings Algorithm

Let $B_{(m)} = \{m\} \times B_m$. Let the current state $x^{(t)} = (m, \tau_1, \cdots, \tau_m; \lambda_0, \cdots, \lambda_m) \in B_{(m)}$ be given. We describe the transition from $x^{(t)} \in B_{(m)}$ to a new point $x^{(t+1)}$ as follows.

Randomly select one of the three types of moves H, H^+, or H^-. Here H is a transition of element in $B_{(m)}$, H^+ a transition of element from $B_{(m)}$ to $B_{(m+1)}$, and H^- a transition of element from $B_{(m)}$ to $B_{(m-1)}$. Denote by P_H^m, $P_{H^+}^m$, and $P_{H^-}^m$ respectively the probabilities of selecting the three different types of moves H, H^+, and H^- when the current state of the Markov chain is in $B_{(m)}$. We set $P_{H^-}^1 = 0$, $P_{H^+}^m = c\min\{1, \frac{p(m+1)}{p(m)}\}$, $P_{H^-}^m = c\min\{1, \frac{p(m-1)}{p(m)}\}$, and $P_H^m = 1 - P_{H^+}^m - P_{H^-}^m$, where c is a sample parameter. Let q denote the density of the normal random variable with mean μ and variance d.

If the move of type H is selected, then

1. select k uniformly from $\{0, 1, 2, \cdots, m\}$ and generate $\lambda' \sim q$.
2. let y be the vector $x^{(t)}$ with λ_k replaced by λ'.

3. set the next state

$$x^{(t+1)} = \begin{cases} y & \text{, with prob.} \quad \min\{1, \dfrac{\nu(y)}{\nu(x^{(t)})}\}, \\ x^{(t)} & \text{, o.w.} \end{cases}$$

If the move of type H^+ is selected, then

1. generate $\tau' \sim U[0, \tau]$; suppose $\tau_{k-1} \leq \tau < \tau_k$.

2. generate $\lambda', \lambda'' \sim q$.

3. let $y = (m+1, \tau_1, \cdots, \tau_{k-1}, \tau', \tau_k, \cdots, \tau_m; \lambda_0, \lambda_1, \cdots, \lambda_{k-2}, \lambda', \lambda'', \lambda_k, \cdots, \lambda_m) \in B_{(m+1)}$.

4. set the next state

$$x^{(t+1)} = \begin{cases} y & \text{, with prob. } \rho_1 \text{ ,} \\ x^{(t)} & \text{, o.w.,} \end{cases}$$

where $\rho_1 = \min\{1, \frac{\nu(y) \cdot p(m) \cdot (\tau \cdot q(\lambda_{k-1}))}{\nu(x^{(t)}) \cdot p(m+1) \cdot (m+1) \cdot q(\lambda') \cdot q(\lambda'')}\}$.

If the move of type H^- is selected, then

1. select k uniformly from $\{1, \cdots, m\}$.

2. generate $\lambda' \sim q$

3. let $y = (m-1, \tau_1, \cdots, \tau_{k-1}, \tau_{k+1}, \cdots, \tau_m; \lambda_0, \lambda_1, \cdots, \lambda_{k-2}, \lambda', \lambda_{k+1}, \cdots, \lambda_m) \in B_{(m-1)}$.

4. set the next state

$$x^{(t+1)} = \begin{cases} y & \text{, with prob. } \rho_2, \\ x^{(t)} & \text{, o.w.,} \end{cases}$$

where $\rho_2 = \min\{1, \frac{\nu(y) \cdot p(m) \cdot m \cdot q(\lambda_{k-1}) \cdot q(\lambda_k)}{\nu(x^{(t)}) \cdot p(m-1) \cdot (\tau \cdot q(\lambda'))}\}$.

Remarks. When applying the above Bayes method to analyze array-CGH data, we need to know the transformation η and the density g, in addition to determining the parameters α, μ and d in the prior. Using the spatial structure, we will indicate a

method to look for η and a method to approximate g, under the assumption that $\{\epsilon_1, \epsilon_2, \cdots\}$ and $\{X_1, A_1, X_2, A_2, \cdots\}$ are independent or the reformulation that all the statements are conditional on $\{X_1, A_1, X_2, A_2, \cdots\}$. The method for specifying g is illustrated in the simulation studies and that for looking for η is indicated in the analysis of a real data set.

3 SIMULATION STUDIES

Using the notations in (1), we define a step function F_0 on $[0,1]$ by setting $(\tau_0, \tau_1, \cdots, \tau_7) = (0, 0.13, 0.2, 0.28, 0.42, 0.51, 0.53, 0.6)$, $(\lambda_0, \lambda_1, \cdots, \lambda_7) = (0.5, -0.2, 0.3, -0.3, 0.1, 0.3, 0.1, 0)$. We now generate data (Y_i, A_i) according to

$$Y_i = F_0(X_i) + \frac{\epsilon_i}{A_i},$$

with X_i being *i.i.d* uniform $[0,1]$ and ϵ_i being *i.i.d.* $N(0, \sigma^2)$. Here $\sigma^2 = (0.1)^2, (0.5)^2$, or 1; $i = 1, 2, \cdots, 1000$; A_i being *i.i.d.* with either $A_i \equiv 1$, or lognormal$(0.2, 0.6)$; $\{\epsilon_1, \epsilon_2, \cdots\}$ and $\{X_1, A_1, X_2, A_2, \cdots\}$ are independent. We note that the number of jump points in F_0 is much smaller than the sample size, which is in line with the spatial structure of the genomic alteration.

To make the data analysis more realistic, we assume ϵ_i being $N(0, \sigma^2)$ with unknown σ^2 and use the following observations to approximate σ^2. Let X_i and X_j be two points with $F_0(X_i) = F_0(X_j)$. Then, since $E(\epsilon_i | A_i, A_j) = E(\epsilon_i \epsilon_j | A_i, A_j) = 0$, we know

$$\begin{aligned} Var(Y_i - Y_j) &= EVar(Y_i - Y_j | A_i, A_j) + VarE(Y_i - Y_j | A_i, A_j) \\ &= EVar(Y_i - Y_j | A_i, A_j) \\ &= \sigma^2 E(\frac{1}{A_i^2} + \frac{1}{A_j^2}). \end{aligned} \tag{4}$$

Since there are 1000 design points in $[0,1]$ and the true regression function F_0 does not have many discontinuities, it is highly likely that (4) holds when X_i and X_j

are neighboring points. Thus, we use (4) with X_i and X_j being adjacent to find approximations to σ^2.

In applying the Bayes method in subsection 2.2, we need to specify α, μ, and d in the prior and the sample parameter c in the reversible jump Metropolis-Hastings algorithm. We set $c = 0.45$, so that there is at least 0.1 probability to explore step functions with the same number of jump points when updating the Markov chain. Since $E\epsilon_i = 0$, we know $EY_i = F_0(X_i)$ and hence we set μ to be the sample mean of $\{Y_1, \cdots, Y_{1000}\}$. Since $Var(Y_i) = EVar(Y_i|F(X_i)) + VarE(Y_i|F(X_i)) = E\epsilon_i^2 + Var(F(X_i)) = \sigma^2 + d$, the sample variance of $\{Y_1, \cdots, Y_{1000}\}$ minus an estimate of σ^2 could be used for d. In this paper, we set d to be the sample variance of $\{Y_1, \cdots, Y_{1000}\}$, which makes the prior flatter and hence maybe less stringent. We set $\alpha = 3$, which implies that the mean number of jump points of the regression function under the prior is about 1.69. Although we know there are 7 jump points in the true regression, we deliberately set $\alpha = 3$ to see if this method still performs satisfactorily.

We run for $1,000,000$ updates and the last $950,000$ realizations are used to obtain the posterior distribution, whose mode is the estimate $\hat{F}_B$.

For the purpose of comparison, we also analyze the same data by the circular binary segmentation method of Olshen et al. (2004). Denote the step function resulted from this estimation by $\hat{F}_S$. We report in Table 1 the performance of these two estimates by presenting the errors in terms of the L_2-norms $||\hat{F}_B - F_0||_2$ and $||\hat{F}_S - F_0||_2$. It is clear from Table 1 that when the noise does not depend on A_i, $\hat{F}_S$ performs comparably with $\hat{F}_B$, but when the noise does depend on the intensity, $\hat{F}_B$ outperforms and the amount of improvement becomes more conspicuous as the noise gets larger.

The reason we choose the method of Olshen et al. (2004) for comparison is that their model assumptions are more similar to ours than other methods and both methods seem to focus the study on the locations of the change points. We note that

Table 1: Performance of the Bayes estimate $\hat{F}_B$ and the estimate $\hat{F}_S$ based on circular binary segmentation.

Table 1a. $A_i \equiv 1$

σ^2	$\|\|\hat{F}_B - F_0\|\|_2$	$\|\|\hat{F}_S - F_0\|\|_2$
$(0.1)^2$	0.0703	0.0756
$(0.5)^2$	0.1334	0.1188
1	0.1469	0.1658

Table 1b. A_i is lognormal(0.2,0.6)

σ^2	$\|\|\hat{F}_B - F_0\|\|_2$	$\|\|\hat{F}_S - F_0\|\|_2$
$(0.1)^2$	0.0662	0.0842
$(0.5)^2$	0.0818	0.1664
1	0.0883	0.2354

we need to choose parameters in implementing the method of Olshen et al. (2004), and we have tried several sets of parameters, including the default one, and only report those that give the least errors.

In addition to the posterior mode, we also calculated the probability of being a jump point based on the posterior distribution. For a design point $X_{(i)}$, we count, among the 950,000 updates, the number of updates that have a jump point in the interval $(X_{(i-5)}, X_{(i+5)})$. Here $X_{(1)}, X_{(2)}, \cdots, X_{(1000)}$ is the order statistic of $X_1, X_2, \cdots, X_{1000}$ and $X_{(i)} = X_{(1)}$ if $i \leq 1$, $X_{(i)} = X_{(1000)}$ if $i \geq 1000$. Denote $\lambda(X_{(i)})$ the quotient of this number over 950,000. Let λ be the continuous function with value $\lambda(X_{(i)})$ at $X_{(i)}$ and linear on $(X_{(i)}, X_{(i+1)})$ for every $i = 1, \cdots, 999$. The plots of λ are shown by thick solid lines in Figure 1. Places having higher λ value are indicative of the discontinuities of the regression function. To make the presentation short, we only report the case that A_i is lognormal(0.2,0.6). In Figure 1, the dashed lines are for the true regression function and solid lines are those for the posterior modes. Note that places at which $\hat{F}_B$ has large jump size often have large λ value. In Figure 1, the left vertical axes indicate the values for Y_i and the right vertical axes indicate the

Figure 1: Performance of the Bayes regression method with A_i being lognormal (0.2,0.6). Dashed line is for the true regression function; solid line is for the posterior mode $\hat{F}_B$; thick solid line indicates the probability that a point is a change point based on the posterior distribution. Dots are the simulated data (X_i, Y_i). Places at which the posterior mode has large jump size usually have large λ values. The left vertical axes indicate the log ratio of the intensities and the right vertical axes indicate the λ values.

Figure 1*a*. $\sigma^2 = (0.1)^2$

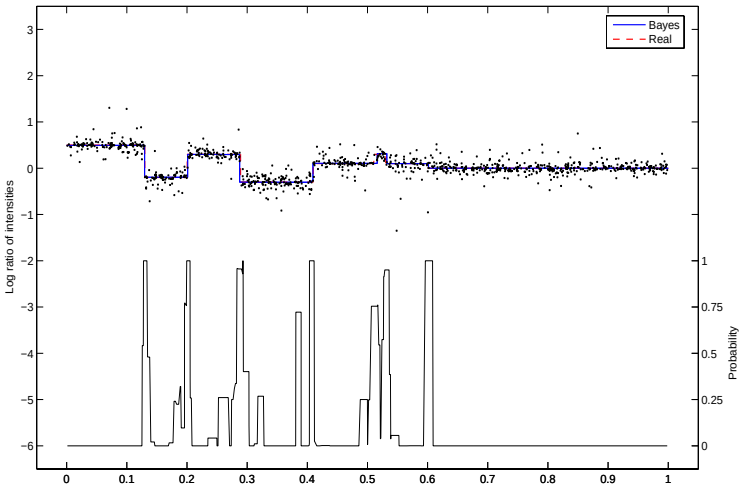


Figure 1*b*. $\sigma^2 = (0.5)^2$

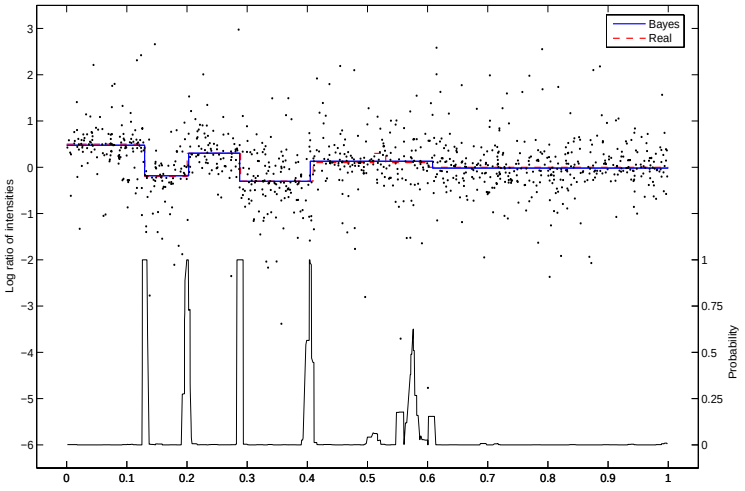


Figure 1*c*. $\sigma^2 = 1$

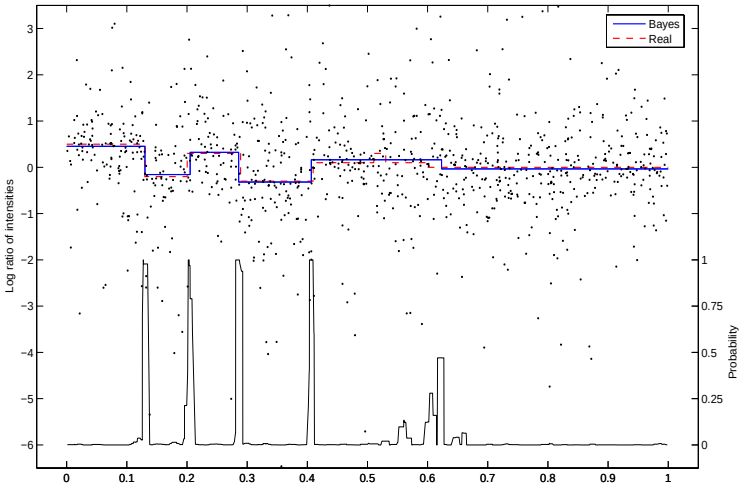
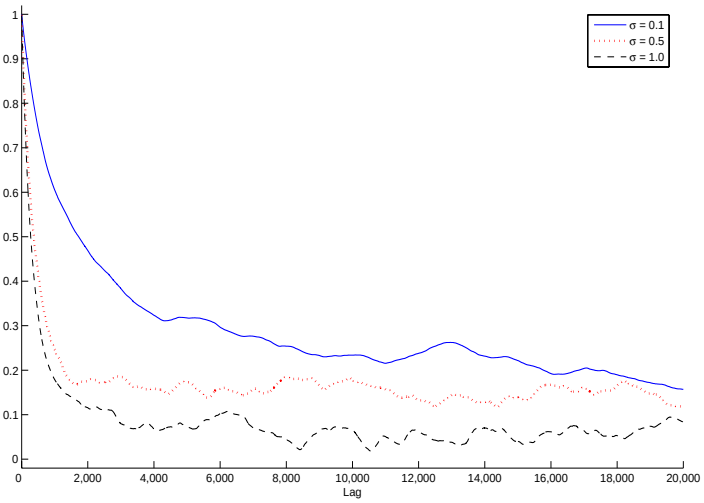


Figure 2: Autocorrelation plots for the reversible jump Metropolis-Hastings algorithm. Solid line for $\sigma^2 = (0.1)^2$, dotted line for $\sigma^2 = (0.5)^2$, and dashed line for $\sigma^2 = 1$.



values for λ.

To assess the behavior of the algorithm, we also present the autocorrelations and effective sample sizes for the L_1-norm of the step function. These quantities seem to suggest that the algorithm behaves nicely. We only report the case that A_i is lognormal(0.2,0.6). The effective sample sizes are respectively 81.3, 137.8, and 279.7 for the cases $\sigma^2 = (0.1)^2, (0.5)^2$ and 1; the autocorrelation plots are contained in Figure 2. The concepts of autocorrelation and effective sample size for Monte-Carlo Markov chains can be found, for example, in Liu (2001).

4 ANALYSIS OF A REAL DATASET

4.1 Data analyses

We now illustrate the method by analyzing a dataset from a cDNA microarray-based comparative genomic hybridization experiment conducted in National Health Research Institutes, Taiwan. The genomic DNA from a cultured NPC cancer cell line and that from peripheral blood mononuclear cells of a healthy individual were extracted and labeled with fluorescent dye Cy5(red) and Cy3(green) respectively. Equal amount of these labeled DNA were then mixed and co-hybridized onto a cDNA microarray containing more than $46,000$ human cDNA probes. The DNA sample from the cancer cell line is referred to as the test sample, and that from the healthy individual is referred to as the reference sample. In this paper, we only consider data for probes derived from chromosome 3. In this cDNA array, there are in total 1204 non-redundant cDNA probes with known positions on chromosome 3, which has a length about 200 Mb.

To apply the method in Section 2, we need to specify η. Since there are 1204 design points and we think there are not many discontinuities of the true regression function, the following exploratory arguments seem to be useful in specifying an appropriate η. Let X_i and X_j be neighboring points and A_i and A_j do not differ much from each

other. Then $\eta(A_i)Y_i - \eta(A_j)Y_j$ should approximate $\epsilon_i - \epsilon_j$ in distribution.

Considering neighboring X_i and X_j with $| \sqrt{A_i} - \sqrt{A_j} | < 0.05$, we present the Q-Q plot of the distribution of these $\sqrt{A_i}Y_i - \sqrt{A_j}Y_j$ against a normal distribution in Figure 3. Since the Q-Q plot is approximately linear, it seems that the simple transformation $\eta(A_i) = \sqrt{A_i}$ is a good choice, although a more sophisticated transformation might give a better Q-Q plot.

Using $\alpha = 3$, $\mu = 0.9542$, $d = 0.2366$, and $c = 0.45$ to set the prior and algorithm parameters, the posterior mode $\hat{F}_B$ is obtained and represented by the solid line in Figure 4. We note that the sample mean and sample variance of $\{Y_i | i = 1, \cdots, 1024\}$ are respectively 0.9542 and 0.2366. The rationals for choosing these parameters are similar to those in Section 3. The Markov chain is updated 1,000,000 times and the last 950,000 updates are used to obtain the posterior distribution. Figure 5 is the autocorrelation of the L_1-norm of the step functions in the chain; the corresponding effective sample size is 160.4. The dashed line in Figure 4 is the graph of the estimate $\hat{F}_S$ using circular binary segmentation method of Olshen et al. (2004), with default parameters. We can see that our method exhibits more jump points.

4.2 Validation of change point using real-time-PCR

The purpose of this study is to check by quantitative real-time PCR (QRT-PCR) if the change points suggested by $\hat{F}_S$ in Figure 4 are supported by QRT-PCR experiments, which measure the relative DNA copy number of amplicons derived from DNA fragments flanking a putative change point. For each change point to be validated, two pairs of primers were designed for the production of two amplicons corresponding to two flanking genomic regions within certain distance from the change point. For each region, two sets of real-time PCR reactions, one with DNA template from the NPC cancer cell line and the other with DNA template from normal cells, were run on a LightCycler real-time PCR machine to obtain the numbers of amplicons derived

Figure 3: Q-Q plot of the distribution of $\sqrt{A_i}Y_i - \sqrt{A_j}Y_j$, with $\mid \sqrt{A_i} - \sqrt{A_j} \mid < 0.05$ and $Y_i = Y_j$, against standard normal distribution.

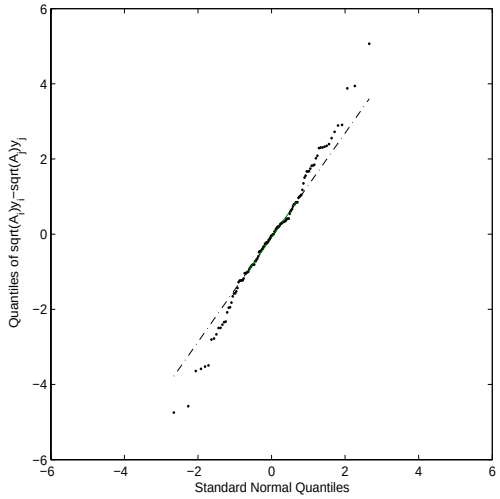


Figure 4: Horizontal axis represents loci in chromosome 3, vertical axis on the left indicates log ratio of the intensities, vertical axis on the right indicates the probability. Dots are the data. Solid line is the estimate $\hat{F}_B$ of this paper; dashed line is the estimate $\hat{F}_S$ obtained by the method of Olshen et al. (2004); thick solid lines indicate the probability that a point is a change point based on the posterior distribution.

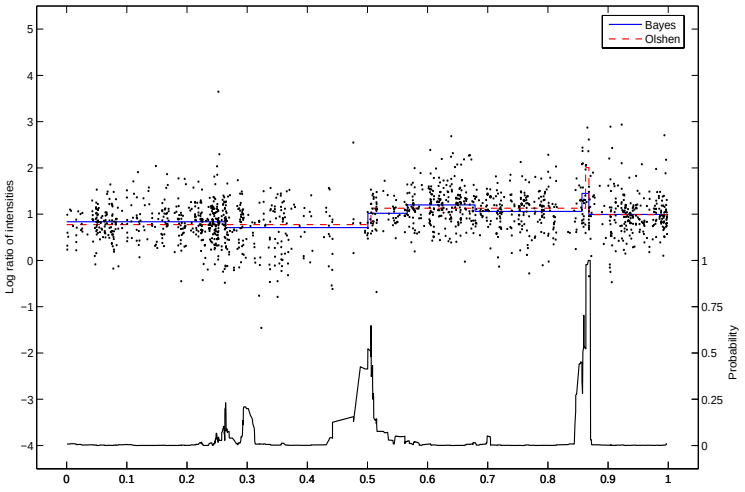
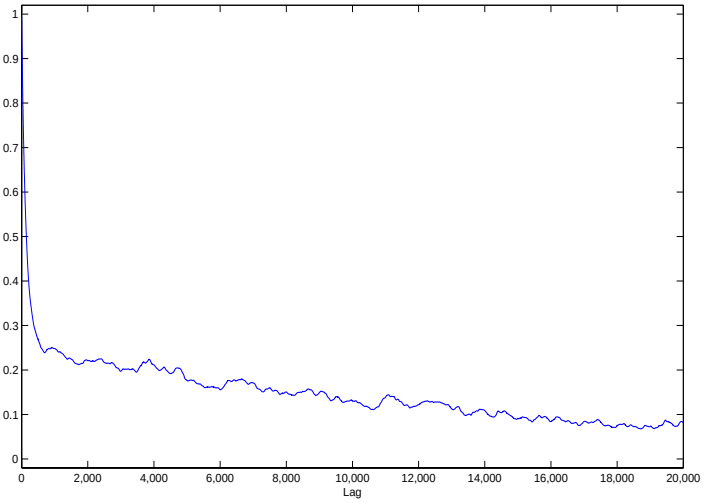


Figure 5: Autocorrelation plot for the reversible jump Metropolis-Hastings algorithm in real data analysis.



respectively from cancer and normal cells under specific experimental conditions, including the number of cycling. Since we assume that the copy number of genomic DNA of normal cells is constant throughout a chromosome within a cell, the number of amplicons derived from cancer cell can be normalized by the number of the corresponding amplicons derived from normal cells. The relative DNA copy number between the two flanking regions of a change point is defined to be the ratio of normalized copy number of the amplicons on one arm to that on the other arm. A ratio different from 1 represents a change in DNA copy number upon crossing the point. We also note that since QRT-PCR is sensitive and the quality of DNA sample varies from preparation to preparation, the relative DNA copy number can only be found consistent when DNA samples are prepared from the same batch and used freshly. Since, in our experiment, DNA samples from the same preparation was held only for validation of a single change point, comparison of DNA copy number derived from different change points are not suggested in this study.

It follows from Figure 4 that $\hat{F}_B$ has six change points, and they are ordered and denoted by $J_1, J_2, \cdots$, and J_6, according to the physical distance; $\hat{F}_S$ has three change points, and they are ordered and denoted by S_1, S_2, and S_3, also according to the physical distance. Since S_1 is close to J_2, S_3 is close to J_6, and J_1 has smaller jump size, we conducted QRT-PCR experiments for each of J_2, J_3, J_4, J_5, S_2, and J_6 as the validation study. The results are contained in Table 2. In Table 2, the second column gives the position, in terms of physical distance from left end of the chromosome, of the change point; the third column gives the intervals representing the two amplicons flanking the change point. The fourth column and fifth are respectively the left/right and right/left ratios. These results indicate that among the change points J_2, J_3, J_4, J_5, and J_6, all, except J_4, are supported by QRT-PCR; the change point S_2, suggested by Olshen et al. (2004) but not by our method, seems not supported by QRT-PCR. Here we follow the tradition in laboratory to decide if QRT-PCR suggests a change point. These seem to indicate that our method performs quite satisfactorily. Note that points having larger jump sizes in $\hat{F}_B$ often have ratio farther away from 1, reported in Table 2.

Table 2: Validation of change point using real-time-PCR.

	Location of change point	Location of the flanking amplicons	$\frac{Left}{Right}$	$\frac{Right}{Left}$
J_2	99835714	(88288246, 88288454) (101947763,101947964)	0.435	2.299
J_3	112923710	(109259900,109260102) (113798402,113798606)	0.665	1.504
J_4	135365576	(135095349,135095559) (135665773,135666026)	0.925	1.081
J_5	170825134	(170285375,170285629) (171154991,171155304)	0.395	2.532
S_2	172134364	(171154991,171155304) (172859690,172859755)	0.832	1.202
J_6	173070748	(172859690,172859755) (180593954,180594040)	3.000	0.333

In fact, because QRT-PCR validation is costly, we decided to carry out the experiments sequentially, starting from points having larger jump sizes and stopping

the experiments once a change point is not confirmed by QRT-PCR. Although we examined amplicons near the change point, some of the amplicons are not as near to the change point as we wish, because the success of a QRT-PCR experiment depends on the availability of amplicons having large exon and usable primers.

5 DISCUSSION

We have presented a Bayes regression approach to array-CGH data analysis. This approach takes into account not only the spatial structure of the genomic alterations but also the intensity dependent noise structure in the array data. Our simulation studies indicate that methods utilizing intensity dependent noise structure do have advantage over methods not using this structure, and this advantage gets noticeable when the structure exists and the data gets noisy. Since array-CGH data based on cDNA microarray is much noisier than those based on genomic clones like BACs, it seems advisable to try our method when analyzing cDNA microarray based CGH data.

This paper assumes the mean of the log ratio of DNA copy number in the test sample to that in the reference sample is a step function on the chromosome. Because of the heterogeneity in the cells of a sample, the ratio may not be exactly an integral multiple of 1/2. This suggests that, in array-CGH data analysis, the study of the change points is easier than the study of the gain or loss in DNA copy number at a given locus. We note that the present paper does not address the latter problem.

Our treatment of regression model with data dependent noise structure is preliminary. It is desirable to treat the case that the transformation η in (1) is not assumed known. The η in our model is a function of the product of the intensities, and we may use other functions of the intensities to adjust the error term in the model for array-CGH data. These extensions will be pursued in a later study.

Produced by The Berkeley Electronic Press, 2006

REFERENCES

Autio, R., Hautaniemi, S., Kauraniemi, P., Yli-Haraja, O., Astola, J., Wolf, M. and Kallioniemi, A. 2003. CGH-Plotter: MATLAB toolbox for CGH-data analysis. Bioinformatics 19, 1714-1715.

Barry, D. and Hartigan, J.A. 1992. Product partition models for change-point problems. Ann. Stat. 20, 260-279.

Barry, D. and Hartigan, J.A. 1993. A Bayesian analysis for change point problems. J. Amer. Statist. Assoc. 88, 309-319.

Cheng, C., Kimmel, R., Neiman, P. and Zhao, L.P. 2003. Array rank order regression analysis for the detection of gene copy-number changes in human cancer. Genomics 82, 122-129.

Chernoff, H. and Zacks, S. 1964. Estimating the current mean of a normal distribution which is subject to changes in time. Annals of Mathematical Statistics 35, 999-1018.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. and Jain, A.N. 2004. Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 90, 132-153.

Green, P.J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711-732.

Hodgson, G., Hager, J.H., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D.G., Pinkel D., Collins, C., Hanahan, D. and Gray, J.W. 2001. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. Nature Genetics 29, 459-464.

Jong, K., Marchiori, E., Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. 2003. Chromosomal breakpoint detection in human cancer. In Applications of Evolutionary Computing. EvoBIO: Evolutionary Computation and Bioinformatics. Springer LNCS 2611, 54-65.

Korostelev, A. 1987. On minimax estimation of a discontinuous signal. Theory Probab. Appl. 32, 727-730.

Lee, C.B. 1998. Bayesian estimation of the number of change points. Statistica Sinica 8, 923-939.

Liu, J.S. 2001. Monte Carlo Strategies in Scientific Computing. Springer-Verlag.

Mantripragada, K.K., Buckley, P.G., deStahl, T.D. and Dumanski, J.P. 2004. Genomic microarrays in the spotlight. Trends Genet. 20(2):87-94.

Muller, H. 1992. Change-points in nonparametric regression analysis. Ann. Statist. 20, 737-761.

Olshen, ADAM B. and Venkatraman, E.S. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5, 4, 557-572.

Oudshoorn, C. 1995. Minimax estimation of a regression funciton with jumps: attaining the optimal constant. Technical Report 934, Dept. Mathematics, Univ. Utrecht.

Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L. and Brown P.O. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Processdings of the National Academy of Scienses USA 99, 12963-12968.

Spokoiny, V.G. 1998. Estimaiton of a function with discontinuities via local polynomial fit with an adaptive window choice. Ann. Statist. 26, 4, 1356-1378.

Wang, Y. 1995. Jump and sharp cusp detection by wavelets. Biometrika 82, 385-397.

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, 1, 45-58.

Wu, J. and Chu, C. 1993. Kernel type estimators of jump points and values of a regression function. Ann. Statist. 21, 1545-1566.

Yao, Y.C. 1984. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. Ann. Statist. 12, 1434-1447.

Yin, Y. 1988. Detection of the number, locations and magnitudes of jumps. Comm. Statist. Stochastic Models 4, 445-455.