The Korean Journal of Applied Statistics (2009)
22(1), 115–127

A Penalized Spline Based Method for Detecting the DNA Copy Number Alteration in an Array-CGH Experiment

Byung-Soo Kim[1] · Sang-Cheol Kim[2]

[1]Dept. of Applied Statistics, Yonsei University; [2]Dept. of Applied Statistics, Yonsei University

(Received August 2008; accepted November 2008)

Abstract

The purpose of statistical analyses of array-CGH experiment data is to divide the whole genome into regions of equal copy number, to quantify the copy number in each region and finally to evaluate its significance of being different from two. Several statistical procedures have been proposed which include the circular binary segmentation, and a Gaussian based local regression for detecting break points(GLAD) by estimating a piecewise constant function. We propose in this note a penalized spline regression and its simultaneous confidence band(SCB) approach to evaluate the statistical significance of regions of genetic gain/loss. The region of which the simultaneous confidence band stays above 0 or below 0 can be considered as a region of genetic gain or loss. We compare the performance of the SCB procedure with GLAD and hidden Markov model approaches through a simulation study in which the data were generated from AR(1) and AR(2) models to reflect spatial dependence of the array-CGH data in addition to the independence model. We found that the SCB method is more sensitive in detecting the low level copy number alterations.

Keywords: DNA copy number alteration, gastric cancer, penalized spline, simultaneous confidence band.

1. Introduction

The DNA copy number at a location in a genome is the number of copies of DNA. The normal copy number is two for the autosomal chromosome in human. Many defects in human development are due to gains and losses of chromosomes and chromosomal segments. DNA copy number alternations occurring in somatic cells are frequent contributors to cancer (Pinkel and Albertson, 2005). Therefore, studying them is a way of identifying and validating important cancer genes (Mestre-Escorihuela *et al.*, 2007). There are various techniques for assessing DNA copy number variations. Among them the array comparative genomic hybridization(array-CGH) has proven its value over the past several years for analyzing DNA copy number variations. The majority of array-CGH platforms use large-insert genomic clones such as bacterial artificial chromosomes(BAC), cDNA's

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2006-312-C00086).

[1]Corresponding author: Professor, Dept. of Applied Statistics, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-749, S. Korea. E-mail: bskim@yonsei.ac.kr

or oligonucleotide (oligo) (Pinkel and Albertson, 2005). It is expected that a gradual transition will be made toward the use of oligo from the initial use of BAC through cDNA's (Ylstra *et al.*, 2006).

In the array-CGH experiment DNA from both reference and test genomes are differentially labeled with fluorescent dyes and competitively hybridized to DNA targets arrayed on a glass slide. The hybridization slide is then scanned and the relative hybridization intensity of the test and the reference signals is ideally equal to the relative copy number of these sequences in the test and the reference genomes. When there is no gain or loss, the relative intensity would be close to one, which is referred to as "normal" in this note. In the case of chromosomal gain or loss the ratio becomes greater than 1 or less than 1. The outcome variable of an array-CGH experiment is $M = \log_2(R/G)$, where R and G represent fluorescent intensities of the reference and the test samples, respectively.

In the microarray experiment it is important to perform the data pre-processing which consists of filtering bad spots, normalization, determining the no missing proportion(NMP) and imputation of missing values. Normalization removes the systematic bias due to variations in experimental conditions and it has been one of hot issues in statistics on microarray (Fan and Niu, 2007; Rigaill *et al.*, 2008). The NMP is defined as the proportion of valid observations out of the total number of arrays (Kim *et al.*, 2005). There are several methods for imputing the missing values including the k-nearest neighbor (Scheel *et al.*, 2005; Kim *et al.*, 2005). However, we don't discuss these issues on the pre-processing in this note.

The purpose of statistical analyses of array CGH experiment data are to divide the whole genome into regions of equal copy number, to quantify the copy number in each region and finally to evaluate its significance of being different from two. Several statistical procedures have been proposed which include the circular binary segmentation(CBS, Olshen *et al.*, 2004; Venkatraman and Olshen, 2007) and a Gaussian based local regression for detecting break points by estimating a piecewise constant function(GLAD, Hupé *et al.*, 2004). Lai *et al.* (2005) conducted a comparison of eleven approaches that included the above two methods. Chari *et al.* (2006) reviewed twenty three software packages for the statistical analysis and visualization of the array-CGH data.

The DNA copy number of a human gene takes a positive integer and hence, the relative copy number, R/G, assumes 1, 0.5, 1.5 and 2 for diploid, monoploid, triploid and tetraploid, respectively. Even though the underlying biological process is of a step function of taking one of 1, 0.5, 1.5, 2 and *etc.* in a segment of a genome, the observed outcome variable shows deviation from the step function (Figure 4.3a) due to various biological and experimental factors. In contrast to the gene expression microarray experiment, array-CGH showed spatial dependence among sequences along the chromosome. That is, if a copy number occurs at a certain locus, we expect that the same copy number will be extended along segments around the locus. Smoothing is a natural way of using spatial dependence (Eilers and de Menezes, 2005).

We propose in this note using a penalized spline regression function to fit the observed data. We then obtain a simultaneous confidence bands(SCB) for the regression function to evaluate the statistical significance of regions of genetic gain/loss. The region for which the SCB stays over zero is deemed to be the region of gain. The region of loss is similarly defined. Small scale simulation results are reported for comparing the proposed method with two other commonly used procedures together with the application of our approach to a gastric cancer data set.

2. Materials and Data

We conducted an array-CGH experiment using a cDNA microarray containing ~17,000 human genes. A normal gastric mucosa sample and a tumor sample were obtained as a pair from each of thirty gastric cancer patients during surgical operations at College of Medicine, Yonsei University in 1997–1999. Each of these patients was followed-up for at least five years. Use of these tissues were approved by the Internal Review Board of College of Medicine, Yonsei University. We adopted the indirect design in which a sex-matched placenta from a clinically healthy mother was used a common reference. Thus, the outcome variable is $\log_2(\text{tumor/ref}) - \log_2(\text{normal/ref})$, where "tumor", "normal" and "ref" represent the fluorescent intensities of the tumor, the normal tissue and the common reference, respectively. For the details of the biological aspects of the experiment, one can refer to Yang (2007) and Yang *et al.* (2007). We may note that Yang *et al.* (2007)'s data set was produced under the direct design which interrogated in a single micorarray a normal gastric mucosa sample and a tumor sample obtained from the same patient as a pair. The array-CGH data set that we analyzed in this note used the same patients as Yang *et al.* (2007)'s. However, we employed the indirect design in the experiment which required two microarrays for each patient.

The data set was preprocessed following the procedure described in Kim *et al.* (2005). Briefly, it consisted of intensity-dependent normalization (Yang *et al.*, 2002), deleting genes containing missing values $> 20\%$ of the total number of observations, employing k-nearest neighbor ($k = 10$) method for imputation of missing values and finally averaging values over multiple spots. We ended up with $10{,}585 \times 30$ data matrix after the preprocessing, where 10,585 and 30 stand for the numbers of genes and cases, respectively.

3. Methods

3.1. Statistical issues and the review on the existing methods

The gene expression cDNA microarray and the array-CGH experiments share several things in their experimental technology. Often an array-CGH microarray experiment paralleled a gene expression microarray experiment using the same platform, for example, to measure how much the variation of the gene copy number contributed to the variation of gene expression in tumor cells (Pollack *et al.*, 2002). Statistical issues of these two experiments are, however, quite different.

Outcome variable, denoted by Y, for short, in a gene expression experiment is the relative amount of mRNA transcripts, whereas Y represents the relative ratio of the DNA copy numbers for an array-CGH experiment. The range of Y is mostly in $(-6, 6)$ for the gene expression, while it is mostly in $(-1, 1)$ for the array-CGH. Two adjacent genes along the genome may have quite different Y's for the gene expression experiment. However, DNA copy number alterations occur in contiguous regions of a chromosome and hence Y's are spatially dependent in an array-CGH experiment. It is quite possible to aggregate Y's over individuals for the gene expression. In an array-CGH between-individual variation is sizable. Hence, exploratory analysis based on each individual is usually conducted, particularly in a small sample study. One can report a region at which a gain occurred for at least 3 cases out of 30 patients. Finding recurrent DNA copy number changes constitutes another statistical issue (Shah *et al.*, 2007; Rouveirol *et al.*, 2006). One of the most important goals in a gene expression microarray experiment is to identify differentially expressed genes between prespecified classes, for example, using a two sample t test for the two group comparison. The primary goal of the array-CGH is to partition the genome into regions of equal DNA copy number

Table 3.1. Comparison of statistical characteristics of two cDNA microarray-based experiments: gene expression experiment versus array-CGH experiment.

	Gene Expression	Array-CGH
Target	mRNA transcript	genomic DNA
Outcome variable (Y)	relative amount of mRNA transcripts	relative ratio of DNA copy numbers
Range of Y	mostly in $(-6, 6)$	mostly in $(-1, 1)$
Spatial dependence	Two adjacent genes along the genome may have quite different Y's.	Y's are spatially dependent, because DNA copy number changes occur in contiguous regions of a chromosome.
Aggregation	Possible to aggregate along individuals	Exploratory analysis based on each individual is usually conducted. Finding recurrent DNA copy number alterations poses another statistical problem.
Statistical methods	Two sample t test *e.g.*, for the two group comparison	Partitioning the genome into regions of equal DNA copy number and quantify the copy number in each region.

and quantify the copy number in each region. Statistical methods of these two experiments were geared to solve different problems and hence they were not exchangeable. Comparison of statistical characteristics of these two experiments can be summarized in Table 3.1.

The initial and natural approach of identifying the locations with copy number transition is the segmentation method, which could be formulated as a change point detection problem. According to Barry and Hartigan (1993) change point detection can be formulated as follows: "We supposed that there is an underlying sequence of parameters partitioned into contiguous blocks of equal parameter values; the beginning of each block is said to be a change point." By extending the standard segmentation method Olshen *et al.* (2004) and Venkatraman and Olshen (2007) developed a circular binary segmentation which detected change points (or break points) where neighboring regions of DNA exhibited statistical significance in the copy number. For the other approaches of the segmentation one can refer to Myers *et al.* (2004), Jong *et al.* (2004) and Picard *et al.* (2007).

Spatial dependence of DNA copy number changes can be accommodated by applying some form of smoothing. Eilers and de Menezes (2005) proposed a quantile smoothing through a fused quantile regression model of which the objective function to minimize was the sum of L_1-norm of the deviance and the L_1-norm penalty of the successive differences of the copy numbers. Li and Zhu (2007) also dealt the detection of regions of gains and losses in the fused quantile regression incorporating the physical locations of clones instead of the uniform spacing between neighboring clones. Other approaches of implementing spatial dependence can be found in Huang *et al.* (2005), Wen *et al.* (2006) and Broët and Richardson (2006).

Hupé *et al.* (2004) applied the adaptive weights smoothing(AWS) algorithm to detect the chromosomal breakpoint based on a Gaussian model. The AWS is an iterative, data-adaptive smoothing procedure that was developed for smoothing in regression problems involving discontinuous regression functions. For the other smoothing or regression approaches one can refer to Huang *et al.* (2005), Hsu *et al.* (2005) and Tibshirani and Wang (2008).

Fridlyand *et al.* (2004) applied an unsupervised (discrete-index) hidden Markov models(HMM) which had been familiar in the speech signal processing area (Rabiner, 1989). HMMs consist of three components: a set of probabilities associated with transitions between all states, a set of probability distributions associated with each states, and a distribution of initial states. The hidden states represent the underlying copy number of the colones. Fitting HMMs on array-CGH data is to partition the clones into the states which represent the underlying copy number of the group of the clones. Stjernqvist *et al.* (2007) proposed a continuous-index HMM by extending Fridlyand *et al.* (2004)'s discrete-index HMM.

3.2. A proposed procedure based on a penalized spline

Materials in the section are heavily dependent on Ruppert *et al.* (2003) and Henderson (1973). Let Y_i denote the log ratio of the i^{th} marker on a chromosome and let x_i represent the physical location of the i^{th} marker(kb). We propose a penalized quadratic spline model of (3.1) to fit the array-CGH plot of Figure 1.1.

$$\begin{aligned} Y_i &= f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma_\varepsilon^2), \ i = 1, \ldots, n, \\ f(x) &= E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K} u_k (x - \kappa_k)_+^2, \end{aligned} \tag{3.1}$$

where $(x)_+ = \max(0, x)$, $\{\kappa_k\}_{k=1}^K$ are knots and $\sum_{k=1}^K u_k^2 < C$ for some constant C.

The regression function f in (3.1) can be viewed as a derivative of its cusum, which renders clearer interpretation in statistical inference on the array-CGH data. The normal region in which no DNA copy number changes occur has zero derivative of the cusum, whereas in gain/loss region the derivative of the cusum is equal to the jump size.

The penalized spline model of Equation (3.1) can be estimated through a linear mixed model formulation as is described in Ruppert *et al.* (2003, pp. 138–139). Also $(1-\alpha)100\%$ simultaneous confidence band for $f(x)$ can be obtained by employing simulation-based approach of Ruppert *et al.* (2003, 6.5)

The penalized spline model of Equation (3.1) can be estimated through a linear mixed model formulation of Equation (3.2)

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + Z\underset{\sim}{u} + \underset{\sim}{\varepsilon}, \tag{3.2}$$

where $\underset{\sim}{y} = (y_1, \ldots, y_n)'$, $\underset{n\times 3}{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$, $\underset{n\times K}{Z} = \begin{bmatrix} (x_1-\kappa_1)_+^2 & \ldots & (x_1-\kappa_K)_+^2 \\ \vdots & \ddots & \vdots \\ (x_n-\kappa_1)_+^2 & \ldots & (x_n-\kappa_K)_+^2 \end{bmatrix}$, $\mathrm{Cov}\left(\begin{bmatrix} u \\ \varepsilon \end{bmatrix}\right) = \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix}$, $\underset{\sim}{\beta} = (1, \beta_1, \beta_2)'$, $\underset{\sim}{u} = (u_1, \ldots, u_K)'$ and $\underset{\sim}{y} \sim N\left(X\underset{\sim}{\beta},\ \sigma_u^2 ZZ' + \sigma_\varepsilon^2 I\right)$.

We introduce some more notations to outline the best linear unbiased prediction(BLUP) procedure of f and a simultaneous confidence band of f.

Let $\underset{1\times 3}{X_x} = [1, x, x^2]'$, $\underset{1\times K}{Z_x} = [(x-\kappa_1)_+^2, \ldots, (x-\kappa_K)_+^2]$ and $\tilde{f}(x) \equiv X_x\tilde{\beta} + Z_x\tilde{u}$, where $\tilde{\beta}$ and $\tilde{u}$ are the BLUP of β and u. $\tilde{\beta}$ and $\tilde{u}$ contain parameters σ_u^2 and σ_ε^2, which are typically estimated via maximum likelihood(ML) or restricted maximum likelihood(REML) method. We plug in the ML or

REML estimates of σ_u^2 and σ_ε^2 in $\tilde{\beta}$ and $\tilde{u}$ and then refer to $\hat{\beta}$ and $\hat{u}$ as estimated BLUP(EBLUP). The $\tilde{f}(x)$ is the BLUP of $f(x) \equiv X_x\beta + Z_x u$ and $\hat{f}(x) \equiv X_x\hat{\beta} + Z_x\hat{u}$ becomes the corresponding EBLUP of $f(x)$. Following the procedure of Ruppert *et al.* (2003, 6.4) and Henderson (1975) one can derive the standard deviation of $\hat{f}(x) - f(x)$ as in Equation (3.3) using the unconditional distribution of Y rather than the conditional distribution of $Y|u$.

$$\widehat{\text{st.dev}}\left\{\hat{f}(x) - f(x)\right\} = \hat{\sigma}_\varepsilon \sqrt{C_x \left(C'C + \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2} D\right)^{-1} C_x'}, \tag{3.3}$$

where $\underset{n\times(K+3)}{C} \equiv [X, Z]$, $\underset{1\times(K+3)}{C_x} = [X_x, Z_x]$, $\underset{(K+3)\times(K+3)}{D} = \text{diag}(0, 0, 0, 1, \ldots, 1)$.

Therefore, approximate $(1-\alpha)100\%$ pointwise confidence interval for $f(x)$ is obtained for large n as in Equation (3.4)

$$\hat{f}(x) \pm z_{1-\frac{\alpha}{2}}\ \widehat{\text{st.dev}}\left\{\hat{f}(x) - f(x)\right\}, \tag{3.4}$$

where $z_{1-\alpha/2}$ indicates $(1-\alpha/2)100^{th}$ percentile of the standard normal distribution.

Now, the simultaneous confidence band of f can be obtained straightforwardly using the simulation-based approach. Suppose we would like to have a simultaneous confidence band for f over a grid of M x-values, say $\underset{\sim}{g} = (g_1, \ldots, g_M)$. Let $f_g \equiv (f(g_1), \ldots, f(g_M))'$ be the true function over g and let $\hat{\underset{\sim}{f}}_g$ denote the corresponding EBLUP based on the quadratic penalized spline of Equation (3.1) formulated in the linear mixed model. Following Ruppert *et al.* (2003, 6.5) we define a $(1-\alpha)100\%$ simultaneous confidence band(SCB) for $\underset{\sim}{f}_g$ as in Equation (3.5).

$$\hat{\underset{\sim}{f}}_g \pm m_{1-\alpha} \begin{bmatrix} \widehat{\text{st.dev}}\left\{\hat{f}(g_1) - f(g_1)\right\} \\ \vdots \\ \widehat{\text{st.dev}}\left\{\hat{f}(g_M) - f(g_M)\right\} \end{bmatrix}, \tag{3.5}$$

where $m_{1-\alpha}$ is the $(1-\alpha)^{th}$ quantile of the random variable in Equation (3.6),

$$\sup_{x\in\chi}\left|\frac{\hat{f}(x) - f(x)}{\widehat{\text{st.dev}}\left\{\hat{f}(x) - f(x)\right\}}\right| \cong \max_{1\le l\le M}\left|\frac{\left(C_g \begin{bmatrix} \hat{\beta} - \beta \\ \hat{u} - u \end{bmatrix}\right)_l}{\widehat{\text{st.dev}}\left\{\hat{f}(g_l) - f(g_l)\right\}}\right|, \tag{3.6}$$

where χ denotes the set of x values of interest and $\underset{M\times(K+3)}{C_g} = \left(\underset{\sim}{1}, \underset{\sim}{g}, \underset{\sim}{g^2}, (\underset{\sim}{g} - \kappa_1\underset{\sim}{1})_+^2, \ldots, (\underset{\sim}{g} - \kappa_K\underset{\sim}{1})_+^2\right)$.

The quantile $m_{1-\alpha}$ can be approximated by simulation. One can compute the corresponding values of Equation (3.6) for a large number of time, say $N = 10,000$. The N simulated values are sorted in the ascending order and the one with rank $\lceil(1-\alpha)N\rceil$ is used as $m_{1-\alpha}$, where $\lceil x\rceil$ denotes the smallest integer greater than or equal to x.

4. Results

4.1. Recurrent copy number alterations from regions of gain/loss in gastric cancer

We applied the simultaneous confidence band(SCB) method to the gastric cancer data set together with GLAD procedure. We found that the SCB method was more sensitive in detecting DNA copy

number alterations(CNAs) than the GLAD procedure. A recurrent CNA in a cohort of patients is defined to be a CNA found at the same location in multiple samples (Shah *et al.*, 2007). One simple strategy to identify a recurrent CNA is to infer recurrent CNAs using a threshold frequency of occurrence. Figure 4.1 shows the number of detected CNAs for the SCB and the GLAD procedures as a function of the threshold frequency. For example, when we set four for the threshold of the GLAD procedure there are about 2,000 CNAs which occurred at least four out of 30 gastric cancers. To have approximately the same number of recurrent CNAs under the SCB procedure, we set twelve for the threshold of the SCB. The determination of recurrent CNAs and the biological significance of its finding, in general, and in this data set, in particular, will be discussed in a separate communication.

The two procedures, GLAD and SCB can be compared in terms of overlapping CNAs in a Venn diagram of Figure 4.2. The two procedures jointly picked up approximately 40% of the total CNAs. The area B of Figure 4.2 Venn diagram represents CNAs detected by the SCB method, but not by the GLAD procedure. The area C of Figure 4.2 shows CNAs detected by the GLAD, but not by the SCB. We found that CNAs in the area B of Figure 4.2 had a characteristic of having relatively low copy numbers alterations in absolute value such that the GLAD procedure didn't pick up.

Figure 4.3 shows a typical plot of a array-CGH for which the SCB method picked up quite a few regions of gain, but the GLAD procedure didn't detect any. The dark curve and the light gray band around the curve in Figure 4.3a represent the estimated regression curve and the 95% SCB, respectively. The region for which the SCB stays over the horizonal line at logratio $= 0$ corresponds to the region of genetic gain. The SCB method could pick up low DNA copy number alterations(CNAs) as in Figure 4.3a, but the GLAD procedure didn't detect them as one can see in Figure 4.3b. In Figure 4.3b empty dots represent the normal copy numbers and all genes are represented by the empty dots, which mean that GLAD dose not detect any region of gain/loss. In the GLAD plot regions of gain and loss are represented by dots of black and light gray, respectively, as one can find in Figure 4.4b and a normal region is shown by as in Figure 4.3b an empty dot.

We also observed that in area C for which the SCB procedure was less sensitive than the GLAD procedure clones were rather sparse as one could typically found in Figure 4.4. In Figure 4.4a there were not many clones up to one sixth of the chromosome, which contributed the large variance of the simultaneous confidnence band, where as in Figure 4.4b the GLAD procedure picked up the same region as the region of gain (black dots). The dotted vertical lines in Figure 4.4b show demarcations between regions of different characteristic. The solid horizonal line in each region represent the mean logratio. Black dots show the region of gain, and dots with light gray exhibit the region of loss. We can see in Figure 4.4b that the GLAD picked up two regions of gain and a region of loss.

4.2. Comparison with GLAD and HMM

We conducted a simulation study and generated 20 data sets. Each data set was generated under independence, AR(1) and AR(2) following Huang *et al.* (2005). However, the true log2 ratios of regions of gain/loss have smaller jump sizes in our simulation. We assumed $0.2 \sim 0.25$ for the jump size in chromosome 1 to 11 and $0.1 \sim 0.15$ in chromosome 12 to 23. Table 4.1 shows the false discovery rates(FDR), true positive and false positive rates averaged over the 20 data sets for three procedures, namely GLAD, HMM and SCB procedures.

We note from Table 4.1 that the SCB outperforms the GLAD and the HMM in terms of FDR and

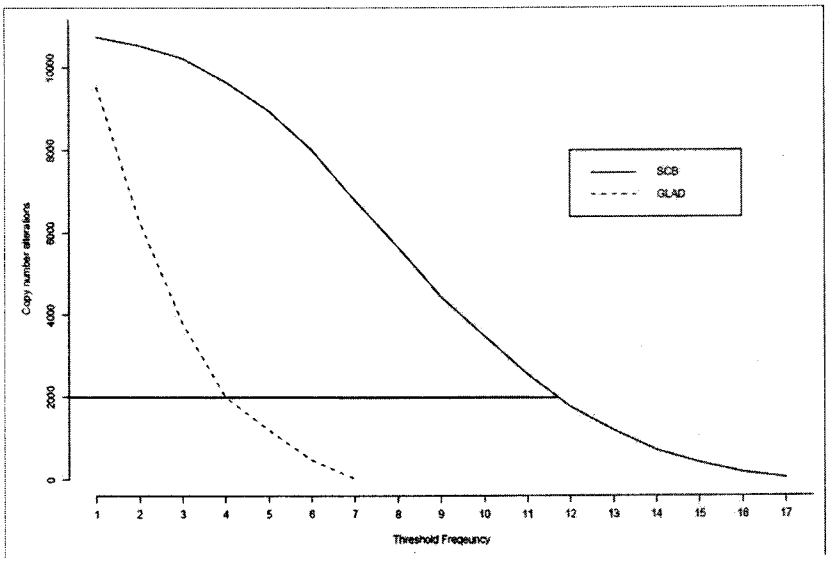


Figure 4.1. The number of detected DNA copy number alterations(CNAs) in multiple samples as a function of the threshold frequency.

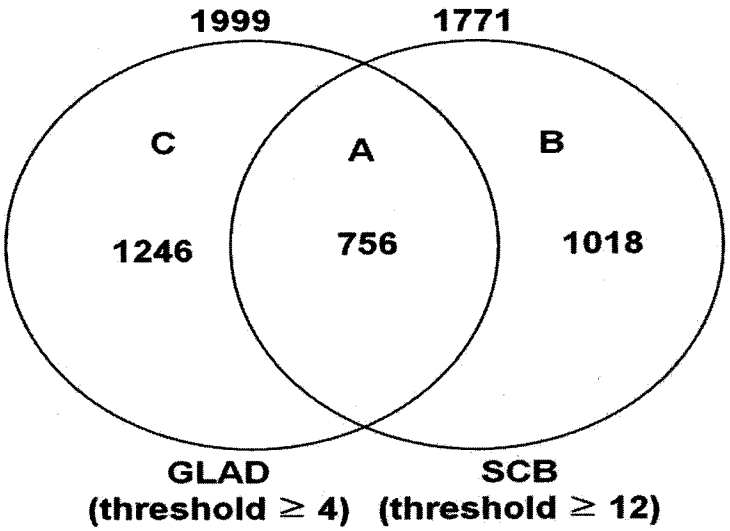
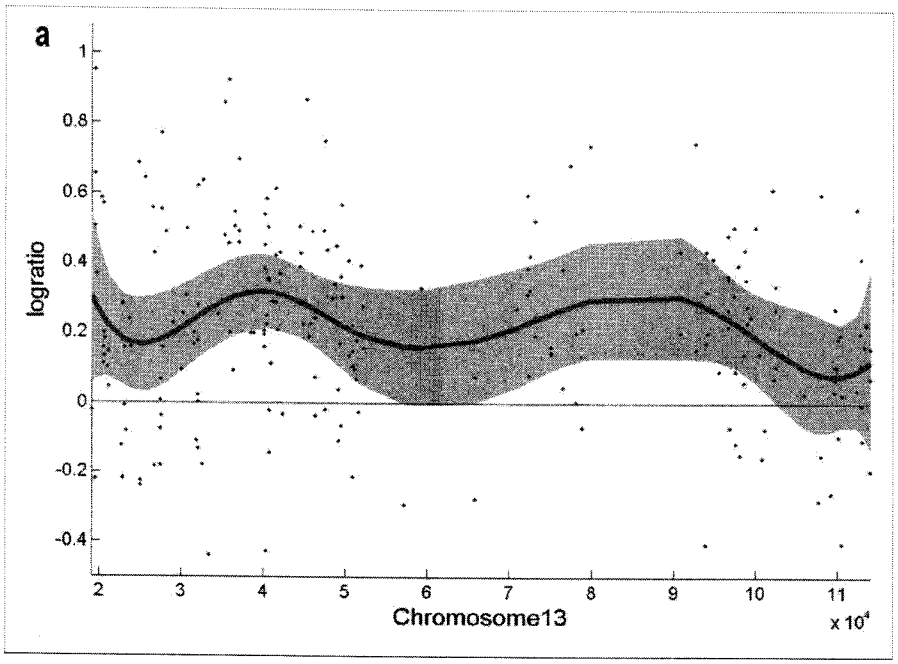


Figure 4.2. The number of DNA copy number alterations(CNAs) that are detected by either GLAD or SCB procedure.

the true positive rates, particularly in chromosomes 12 to 23 which have small jump sizes, namely 0.1∼0.15. We may note that the performance measures(*i.e.*, FDR: true positive rates and false positive rates) of GLAD and SCB methods are rather stable in all three models of independence, AR(1) and AR(2). As a means of implementing spatial dependence along the contiguous regions in a chromosome, we generated simulated data from AR(1) and AR(2). It turned out that the



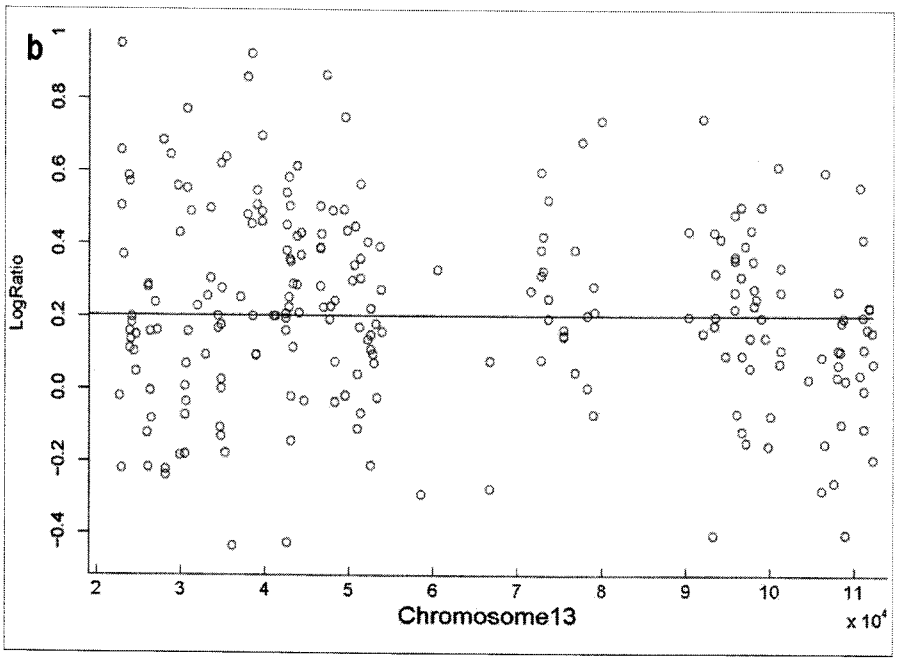


Figure 4.3. Comparison between the SCB method and the glad procedure. The SCB is more sensitive in detecting low DNA copy number alterations. Figure 4.3a shows an estimated regression curve and its 95% SCB. Figure 4.3b represents an output of the GLAD procedure.

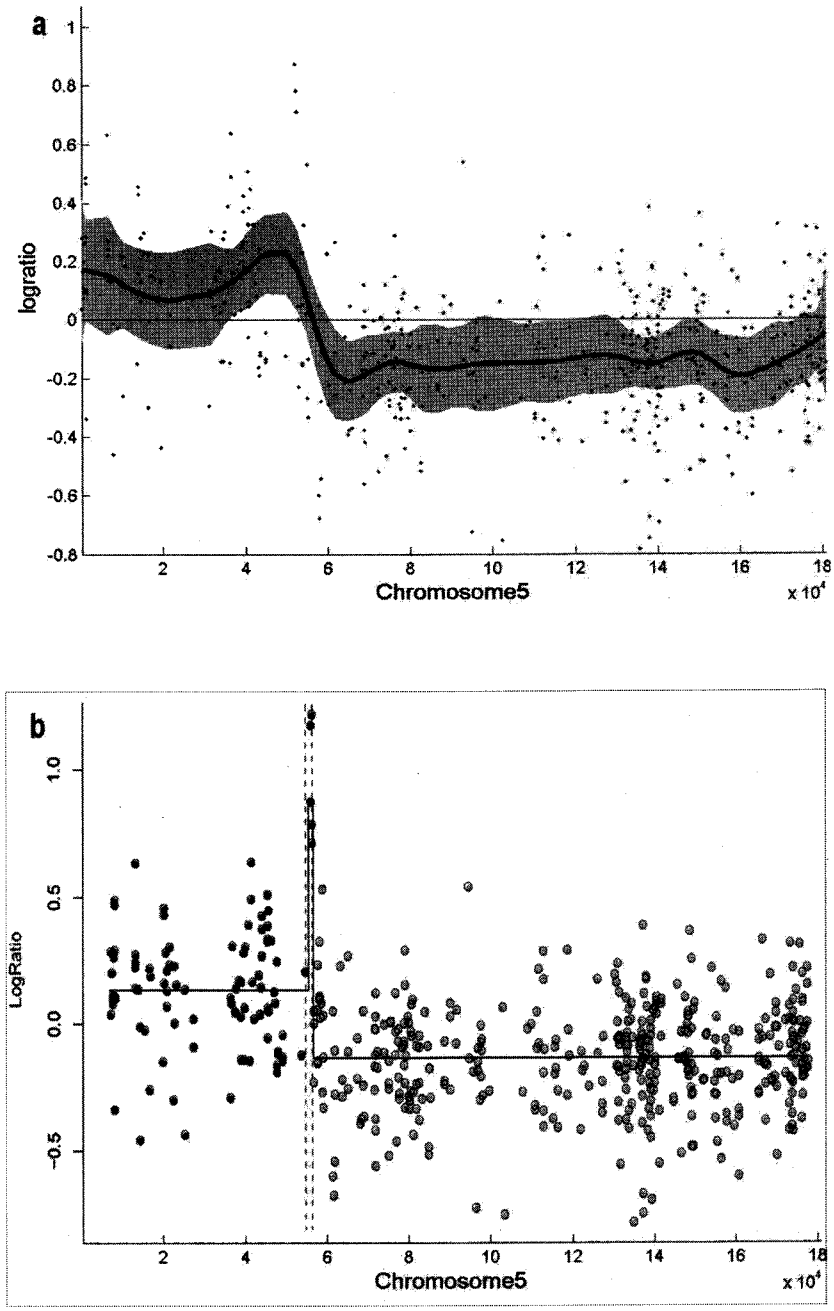


Figure 4.4. Comparison between the SCB method and glad procedure. The SCB method failed to detect the region of gain in an area where not many clones were spotted. Figure 4.4a shows an estimated regression curve and its 95% SCB. Figure 4.4b represents an output of the GLAD procedure.

Table 4.1. Comparison of three procedures: GLAD, HMM and simultaneous confidence band(SCB)

FDR									
	Independence			AR(1)			AR(2)		
	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall
GLAD	0.0049	0.1316	0.0417	0.0061	0.119	0.04	0.0081	0.089	0.0316
HMM	0.2254	0.2272	0.2393	0.2258	0.2392	0.2603	0.4874	0.3708	0.4301
SCB	0.0743	0.087	0.0803	0.076	0.0833	0.0791	0.0671	0.0835	0.0743
true positive rates									
	Independence			AR(1)			AR(2)		
	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall
GLAD	0.6598	0.2092	0.4539	0.7249	0.2204	0.4943	0.6271	0.1896	0.4272
HMM	0.0733	0.232	0.1458	0.0811	0.2007	0.1358	0.3085	0.3124	0.3103
SCB	0.7902	0.6921	0.7453	0.7945	0.6675	0.7365	0.7964	0.6636	0.7357
false positive rates									
	Independence			AR(1)			AR(2)		
	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall	Ch1–Ch11	Ch12–Ch23	Overall
GLAD	0.0024	0.0384	0.0161	0.0031	0.0454	0.0191	0.0036	0.0238	0.0113
HMM	0.0224	0.0722	0.0413	0.0286	0.0676	0.0434	0.1973	0.1674	0.1859
SCB	0.0451	0.0648	0.0525	0.0461	0.0593	0.0511	0.0408	0.0598	0.0481

GLAD and the SCB methods were robust in the transition from independence model to AR(1) and to AR(2) models. We also note that the HMM shows lower performance in term of FDR and true positive rates and the false positive rate of HMM turns out unstable in the transition of AR(1) to AR(2) model. We observe that the type I error rate of the SCB is quite stable around the nominal level of 0.05. For chromosomes 1 to 11 which have jump sizes 0.2∼0.25 the GLAD and the SCB procedures trade off the FDR and the false positive rates, and we may judge that these two procedures are comparable in chromosomes 1 to 11.

5. Discussions

In this paper we proposed a new approach, namely, the penalized spline based simultaneous confidence band(SCB) approach for detecting regions of gain/loss in an array CGH experiment data set. We first applied the simultaneous confidence band-based procedure to array-CGH data of 30 gastric cancer patients. We found that it could detect more regions of gain/loss than GLAD procedure. When we compared the performance of this procedure with GLAD and HMM based procedures, we learned that our procedure turned out to be more sensitive in detecting regions of gain/loss of which the jump sizes are small. Furthermore, it turned out that the SCB method exhibited stable type I error probability in the transition from the independence model to AR(1) and to AR(2), whereas neither the GLAD nor the HMM method enjoyed this property.

The SCB procedure yielded more regions of gain/loss than the GLAD procedure. It was not uncommon that different statistical methods detected different number of regions of gain/loss in the array-CGH experiment. We went around this unequal numbers by equalizing the number of recurrent CNAs of both procedures, which was possible by tuning the threshold frequency. However, determining a recurrent CNA would comprise another important statistical issue and remains a further research topic. In particular to the gastric cancer data set the biological significance of CNA's in area B and C of Figure 4.2 is yet to be evaluated by further research.

References

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association*, **88**, 309–319.

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.

Chari, R., Lockwood, W. W. and Lam, W. L. (2006). Computational methods for the analysis of array comparative genomic hybridization, *Cancer Informatics*, **2**, 48–58.

Eilers, P. H. C. and de Menezes, R. X. (2005). Quantile smoothing of array CGH data, *Bioinformatics*, **21**, 1146–1153.

Fan, J. and Niu, Y. (2007). Selection and validation of normalization methods for c-DNA microarrays using within-array replications, *Bioinformatics*, **23**, 2391–2398.

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

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model, *Biometrics*, **31**, 423–447.

Hsu, L., Self, S. G., Grove, D., Randolf, 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.

Huang, T., Wu, B., Lizardi, P. and Zhao, H. (2005). Detection of DNA copy number alterations using penalized least squares regression, *Bioinformatics*, **21**, 3811–3817.

Hupé, P., Stransky, N., Thiery, J. P., 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.

Jong, K., Marchiori, E., Meijer, G., Vaart, A. V. D. and Ylstra, B. (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data, *Bioinformatics*, **20**, 3636–3637.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y. and Chung, C. H. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer, *Bioinformatics*, **21**, 517–528.

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.

Li, Y. and Zhu, J. (2007). Analysis of array CGH data for cancer studies using fused quantile regression, *Bioinformatics*, **23**, 2470–2476.

Mestre-Escorihuela, C., Rubio-Moscardo, F., Richter, J. A., Seibert, R., Clement, J., Fresquet, V., Beltran, E., Agirre, X., Marugan, I., Marin, M., Rosenwald, A., Sugimoto, K. J., Wheat, L. M., Karran, E. L., Garcia, J. F., Sanchez, L., Prosper, F., Staudt, L. M., Pinkel, D., Dyer, M. J. and Martinez-Climent, J. A. (2007). Homozygous deletions localize novel tumor suppressor gene in B-cell lymphoma, *Blood*, **109**, 271–280.

Myers, C. L., Dunham, M. J., Kung, S. Y. and Troyanskaya, O. G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microaray data, *Bioinformatics*, **20**, 3533–3543.

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., Lebarbier, E. and Daudin, J.-J. (2007). A segmentation/clustering model for the analysis of array CGH data, *Biometrics*, **63**, 758–766.

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

Pollack, J. R., Sørlie, T., Perou, C. M., Rees, C. A., Jeffrey, S. S., Lonning, P. E., Tibshirani, R., Botstein, D., Børresen-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, *Proceedings of the National Academy of Sciences*, **99**, 12963–12968.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, In *Proceedings of the IEEE*, **77**, 257–286.

Rigaill, G., Hupé, P., LaRosa, P., Meyniel, J.-P., Decraene, C., Almeida, A. and Barillot, E. (2008). ITALICS: An algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays, *Bioinformatics*, **24**, 768–774.

Rouveirol, C., Stransky, N., Hupé, P., Rosa, P. L., Viara, E., Barillot, E. and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data, *Bioinformatics*, **22**, 849–856.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, New York.

Scheel, I., Aldrin, M., Glad, I. K., Sørum, R., Lying, H. and Frigessi, A. (2005). The inference of missing value imputation on detection of differentially expressed genes from microarray data, *Bioinformatics*, **21**, 4272–4279.

Shah, S. P., Lam, W. L., Ng, R. T. and Murphy, K. P. (2007). Modeling recurrent DNA copy number alterations in array CGH data, *Bioinformatics*, **23**, i450–i458.

Stjernqvist, S., Rydèn, T., Sköld, M. and Staaf, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data, *Bioinformatics*, **23**, 1006–1014.

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

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

Wen, C-C., Wu, Y-J., Huang, Y-H., Chen, W-C., Liu, S-C., Jiang, S. S., Juang, J. L., Lin, C. Y., Fang, W. T., Hsiung, C. A. and Chang, I. S. (2006). A Bayes regression approach to array-CGH data, *Statistical Applications in Genetics and Molecular Biology*, **5**, Article 3.

Yang, S. (2007). Gene amplifications at chromosome 7 of the human gastric cancer genome, *International Journal of Molecular Medicine*, **20**, 225–231.

Yang, S., Jeung, H. C., Choi, Y. H., Kim, J. E., Jung, J-J., Jeong, H. J., Rha, S. Y., Yang, W. I. and Chung, H. C. (2007). Identification of genes with correlated patterns of variations in DNA copy number and gene expression level in gastric cancer, *Genomics*, **89**, 451–459.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation, *Nucleic Acid Research*, **30**, e15.

Ylstra, B., van der IJssel, P., Carvalho, B., Brakenhoff, R. H. and Meijer, G. A. (2006). BAC to the future! or oligonucleotides: A perspective for micro array comparative genomic hybridization(array CGH), *Nucleic Acid Research*, **34**, 445–450.