Copy Number Detection Using Self-weighted Least Square Regression

Xiaorong Yang
College of Statistics & Mathematics
Zhejiang Gongshang University
Hangzhou, 310018, China
Email: xryang@mail.zjgsu.edu.cn

Ke-Ang Fu
College of Statistics & Mathematics
Zhejiang Gongshang University
Hangzhou, 310018, China
Email: fukeang@hotmail.com

***Abstract*—In this article, an efficient algorithm to detect the breakpoints in DNA copy number alterations is considered. In view of the influence of the heavy noises, the self-weighted least square estimation is adopted to downweight the covariance matrix of the wild observations (outliers), which ensure the convergence between the estimated parameters and the true values. The proposed approach makes use of the most of the data itself to reduces the complexity of the model, and presents an insightful discussion for discovery of copy number alterations.**

I. Introduction

DNA copy number alterations characterize the potential biomarkers of the human diseases. Amount of techniques contribute to the structural detection in the expression of gene microarrays. With the development of genome-wide analysis of DNA copy number, various high resolution platforms emerge, such as comparative genomic hybridization (CGH) array and single nucleotide polymorphism (SNP) arrays. Related literatures always focus on the identification of the change points of the amplified or deleted segments. For instance, [1] developed a partition method called circular binary segmentation (CBS), which seeks for two break points at a time by considering the segment as a circle. It is a modification of the traditional binary segmentation. [2] used an unsupervised hidden Markov model (HMM) approach to sort the chromosome into different states representing different copy numbers. Later on, dynamic programming developed by [3] was used to search the change points. This approach was further improved by [4] with a penalized likelihood being combined. [5] tried to denoising the data by wavelet smoothing method in the detection process. [6] proposed a LASSO based least regression with L_1 penalty to access DNA copy number alterations.

The goal of the data analysis is to obtain the accurate structure of the sequence, and the cost of the computation time is also of important. Some existing approaches with complicated models or nuisance parameters are only feasible theoretically. Moreover, the resolution of arrays is much higher than before, the data itself contains heavy noises due to all kinds of mistakes generated during the experiments or manipulations. The expression data is usually modeled as regressions plus residuals. Traditional methods always restricted the models with Gauss noises or noises with finite moments. Although current methodological advancements in bioinformatics or computational biology, models with Gaussian assumption in stochastic processes are still widely assumed in various cases, it does not mean the assumptions are unfeasible in practical problems. As mentioned, data is usually contaminated or presents some dependent properties, therefore, new approaches which have the capability to deal with heavy noise are desired.

Among the frequently used model, the regressive process models the observed value as a true copy number at a specific marker plus a random noise. The ordinary regressive estimation such as the least square method gives the same weight to the wild observations which cannot reduce the effect of these points on the covariance matrix and make the convergence between the estimated parameters and the true values no longer hold. Motivated by practical applications, our interest is to reduce the effect of the wild points in the sequence. We employ a self-weighted regression estimation to deal with the models involving heavy noises including infinite variance residuals. This method is distribution free for residuals. In a sense, the proposed method can be widely used.

The rest of this paper is organized as follows. Section 2 shows the models and methods. Section 3 describes the results of data analysis. And the discussion is presented in Section 4.

II. Models and Methods

In this section, we describe the model fitted the copy number alterations, and state the statistical hypothesis test. The change points detection problem is formulated as an optimization, which could use a recursive algorithm to get a quick reply.

A. The Regression Model

For a CGH profile or a SNP array, we assume Y_i is the log2 ratio of intensity of marker i on a chromosome. Y_i can be realized by a true relative copy number μ_i at marker i plus a random noise (we call it residual), see the following formulation

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

2011 IEEE International Conference on Systems Biology (ISB)
978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Zhuhai, China, September 2–4, 2011

where n is the number of markers on a given chromosome. Since the copy number data are ordered by the locations of the markers and have spatial dependence due to physical dependence, the intensities of any adjacent markers are very close to each other.

Set $\mu_0 = 0$, and define $\beta_j = \mu_j - \mu_{j-1}$, then β_j can be regarded as as the jump between the $(j-1)$th and jth markers. Then model (1) can be reformulated into the following one

$$Y_i = \sum_{j=1}^{i} \beta_j + \varepsilon_i, \quad , i = 1, 2, \cdots, n.$$

The above model can be rewritten as

$$Y_i = I^T \beta^i + \varepsilon_i, \quad , i = 1, 2, \cdots, n, \tag{2}$$

where I is a $i \times 1$ unit vector, and $\beta^i = (\beta_1, \beta_2, \cdots, \beta_i)^T$. If there is no copy number alteration, $\beta^i = (\mu_0, 0, \cdots, 0)^T = \beta^0$. Consequently, we aim at testing the following null hypothesis that β^i remains constant, i.e.

$$H_0 : \beta^i = \beta^0$$

against the alternative one that at least one component of β^i varies over positions.

In practical application, m change points are allowed to appear in a chromosome. Then β^i shifts from one stable state to different ones. Altogether, there are $m + 1$ segments in which the components are constants. We consequently use the partition index $I_m = \{i_1, i_2, \cdots, i_m\}$ to denote the set of change points, hence model (2) is equivalent to

$$\begin{aligned} &Y_i = I^T \beta^j + \varepsilon_i, \\ &i = i_{j-1}, i_{j-2} \cdots, i_j; \quad j = 1, 2, \cdots, m+1, \end{aligned} \tag{3}$$

and obviously, $i_0 = 0$, and $i_{m+1} = n$.

B. The Self-weighted Least Square Estimation

Usually, the parameters β^j in model(3) is derived by ordinary least square estimation, i.e.

$$\hat{\beta}^j = \arg\min \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{i} \beta_j)^2,$$

see [6] for example. Asymptotic normality and weak convergence of the estimators are required in statistical inference, or else the obtained estimators are meaningless. However, to get these properties, finite moment assumptions on residual sequence $\{\varepsilon_i\}$ are usually needed. As mentioned, in bio-data analysis, noises heavily influence the results. Large positive or negative values of residuals produce the response variables to be outlier. Sometimes, the same error produces many leverage points such that response variables have heavy tails, which results in the existence of large deviation. To ensure the convergence between estimated parameters and the true value, the we modify the objective function as

$$\hat{\beta}^j = \arg\min \sum_{i=1}^{n} \omega_i (Y_i - \sum_{j=1}^{i} \beta_j)^2, \tag{4}$$

where the weight ω_i is a function of $\{Y_1, \cdots, Y_n\}$, and that is why we call it "self-weighted least square estimation".

By the proposed approach, even if the residuals $\{\varepsilon_i\}$ have infinite variance, i.e. the data have heavy noises, we still obtain a good estimation which approximates of the true values well. The weights used in our method are analogues to the influence function in [7]:

$$w_t = \begin{cases} 1, & \text{if} \quad a_t = 0, \\ C^3/a_t^3, & \text{if} \quad a_t \neq 0, \end{cases} \tag{5}$$

where $a_t = |Y_t| I_{(|Y_t| \geq C)}$ and $C > 0$ is a constant.

C. Structural Changes

We pay close attention to the structural change of the DNA copy number, therefore, a statistical test is stated here. For the simplest case, if there is only one change point at position j, i.e., $m = 1$, the F test statistic can be calculated by

$$F_j = \frac{\hat{\varepsilon}^T \hat{\varepsilon} - \hat{\varepsilon}(j)^T \hat{\varepsilon}(j)}{\hat{\varepsilon}(j)^T \hat{\varepsilon}(j)/(n-2k)}, \tag{6}$$

where $\hat{\varepsilon}(j)$ is the ordinary least square residual at position j, and $\hat{\varepsilon}$ is the residual from the unsegmented model. The above defined F statistics are then computed for $i = n_h, \cdots, n - n_h (n_h > k)$, where $n_h = [nh]$ is a trimming parameter chosen by the practitioner. We reject the null hypothesis H_0 if their supremum is too large.

If the partitions I_m is given, the least square estimates for β^j can be easily obtained. The residual sum of squares (RSS) is given by

$$RSS(i_1, i_2, \cdots, i_m) = \sum_{j=1}^{m+1} rss(i_{j-1} + 1, i_j), \tag{7}$$

where $rss(i_{j-1} + 1, i_j)$ is the usual minimal residual sum of squares in the jth segment. However, in practice, the change points are rarely given exogenously but are unknown, the purpose is to find the estimation $\hat{i}_1, \cdots, \hat{i}_m$ that minimize the objective function

$$(\hat{i}_1, \cdots, \hat{i}_m) = \arg \min_{(i_1, i_2, \cdots, i_m)} RSS(i_1, i_2, \cdots, i_m). \tag{8}$$

The global minimizers in (8) is not easy to derive, if we execute a grid search, the computing complexity would be of order $O(n^m)$. For any $m > 2$, the computation time will out of affordability. To overcome the problem, many hierarchical algorithms have been proposed, like the recursive partitioning by [8], the joining subsampling method by [9], etc. These approaches are much more efficient than global search, and are of order $O(n^2)$ for any m. In a maximum likelihood framework, [10] discussed the change points estimation problems, which extended the early work in [11]. Following the idea "Bellman's principle": the optimal segmentation satisfies the recursion in (9) (see below), [12] presented a dynamic

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

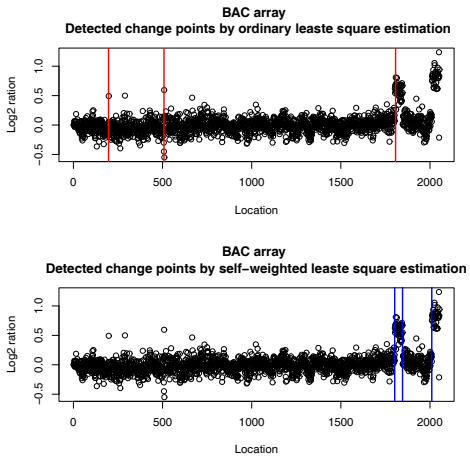

Fig. 1. Detected change points for BAC array. The data spots are the log2 ratios of the observations. The vertical lines indicate the predicted change points of the array.

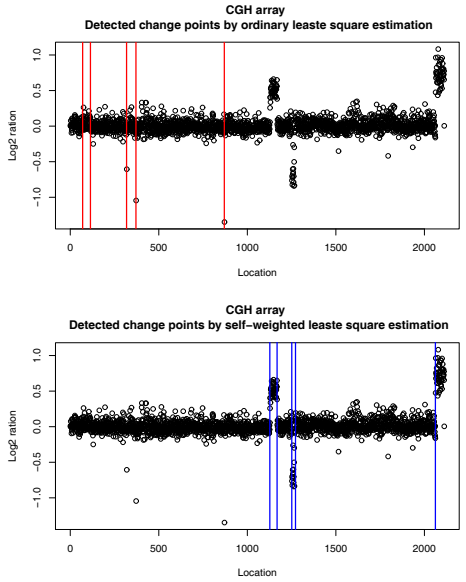

Fig. 2. Detected change points for CGH array. The data spots are the log2 ratios of the observations. The vertical lines indicate the predicted change points of the array.

programming algorithm for pure and partial structural change models in an ordinary least square regression context.

$$RSS(I_{m,n}) = \min_{mn_h \le j \le n-n_h} [RSS(I_{m-1}, j) + rss(j+1, n)]. \quad (9)$$

It implies that we only need to know the "optimal previous partner" for each breakpoint j in an m-partition. The computation cost will decrease significantly since the recursive relation $rss(i,j) = r(i,j-1) + r(i,j)^2$ holds, where $r(i,j)$ is the recursive residual at time j of a sample starting at i. One can refer [12] and [13] for details.

III. Results

We applied our approach to some public datasets. Simulations of heavy noise dataset will be generated to test the performance of the method. Also, high resolution SNP array will be included to see the efficiency of the algorithm (see section IV for details).

A. Application to public dataset

To evaluate our method, we first apply the algorithm to two public datasets. The first one is BAC array (http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html). The dataset was used by [1], [2], [5], [14] and others to evaluate their methods. This BAC array contains measurements for 2053 BACs spotted (spots with no expression values were deleted) in triplicates. The second one is a CGH array used in [6] (http://bioinformatics.med.yale.edu/DNACopyNumber). The detection results by both ordinary least square estimation and self-weighted least square estimation are visualized in Fig. 1 and Fig. 2 for BAC array and CGH array, respectively. The weights used is defined as (5) with the cutoff value $C = 0.35$, which corresponds to the critical value of a copy number of three or one, see also [6].

The vertical lines in both Fig. 1 and Fig. 2 are the locations where the predicted change points are. Both the top subfigures indicate that the detection results by ordinary least square estimation will be interfered by noise because we give the equivalent weights to each observation. Therefore, the segmentations are quite different with the truth. However, if we used the proposed method to modify the weights of outlier, the copy number alterations can be inferred correctly. It implies that our approach is a powerful tool in the applications for contaminated dataset, as we make good use of the data itself.

B. Simulation

The simulation data used in this paper is referred as that of [6]. However, we modify the residuals as random variables generated from Cauchy distribution, which have infinite variance (i.e. heavy tail noises).

We assume that the true log2 ratios of 1000 markers follow model (1) with $\{\mu_i\}$ defined in Table I.

The residuals $\{\varepsilon_i\}$ follow the Cauchy distribution with location parameter 0 and scale parameter 1. Both the mathematical expectation and the variance of Cauchy distributed variable are infinite, which mean the residuals are heavy tailed.

Since the residuals has no mathematical expectation, the absolute values of the observations are pretty large, some

TABLE I
SIMULATION DATA: LOG2 RATIO OF 1000 MARKERS

i	1-100	101-150	151-450
μ_i	0	1	0
i	451-600	601-900	901-1000
μ_i	0.585	0	-1

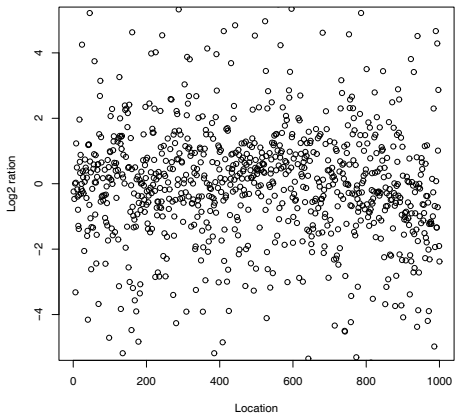

Fig. 3. Visualization of simulation data with residuals' range from -5 to 5. The data spots are the log2 ratios of the observations.

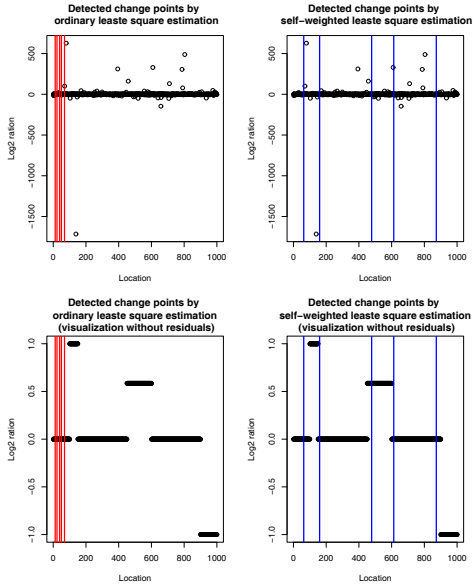

Fig. 4. Detected change points for simulation data. The data spots are the log2 ratios of the observations. The vertical lines indicate the predicted change points of the array.

are great than 1500. We visualized all 1000 observations in Fig. 3 with the range of vertical axes range from -5 to 5. From Fig. 3, we notice that the scatter plot of the data exhibits disorderly and unsystematic, we can hardly catch the true copy number alterations. The detection problem becomes troublesome. The detected change points by both ordinary least square estimation and self-weighted least square estimation are visualized in Figure 4. Cauchy distribution may generates extremely large values, therefore, the scale of vertical axes varies from -1500 to 500 in the top windows in Fig. 4. It makes the breakpoints invisible. In order see the performance of our method, we visualized the detected change points of the data again in the bottom of Fig. 4, but we remove the residuals such that the original copy number of each marker could be clearly displayed. From the comparison, we find that our approach eliminate the influence of the noises. However, the traditional least square estimation is thoroughly inefficient.

IV. Discussion and Conclusion

Various statistical approaches for analyzing the copy number data were developed in the past few years. Our algorithm is a kind of regression based method in essence. While the realization of the detection based on a dynamic programming with the complexity of order $O(n^2)$. An analogous approach is a LASSO based penalized least square estimation in [6]. The LASSO based method fails with default parameters if we modify the residuals from a Cauchy distribution. Moreover, we once tried to apply this method to a 250K SNP array, it took more than 6 hours to run the algorithm for each sample array, but nothing could be found. The failure might because the LASSO based method is time-consuming for high resolution data. We also apply our algorithm to each chromosome of the same array, the average response time is 25 minutes.

The selection of the weights is mild. Under some restrictions, if proper weights is given (for example, the $\mathsf{E}(\omega_t+\omega_t^2) < \infty$, where E denote the mathematical expectation.), it can easily downweights the covariance of the wild observations. There are many different weights can be chosen, such as $w_t = (1 + C\|Y_t\|^2)^{-3/2}$, $I(\max_{1\le i\le p}|Y_{t-1}| \le C)$. A basic principle of selection of the weights is that "the larger the observations are, the smaller the weights should be". The detection results are not sensitive to the selected weights if the above principle satisfied. However, the cutoff of value (i.e. the constant C in weights) should be meaningful (like we use 0.35 because it corresponds to the absolute values of log2 ratio of copy number deletion or amplification). Moreover, different weights will results in different response time of the algorithm. The theory of the weight selection can go through [15], [16] and [17] for details. The weights defined in (5) work more efficient than others.

In this article, we proposed an efficient algorithm to detect the breakpoints in the copy number alteration. To avoid the influence of the heavy noises, we adopt the self-weighted least square to downweight the covariance matrix of the wild observations, which make the convergence between the estimated parameters and the true values hold. The proposed approach makes the most of the data itself and dramatically reduces the complexity of the model. The superiority of our method is that the distribution of the residuals in the model is free, therefore it presents an insightful discussion for copy number alteration discovery.

Acknowledgment

This research is supported by National Natural Science Foundation of China (No. 11026087 & 11071214); Humanities and Social Sciences Foundation of the Ministry of Education of China (No. 10YJC910010).

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

REFERENCES

[1] A. B. Olshen, E. S. Venkatraman, R. Lucito and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," Bioinformatics, vol. 5, pp. 557-572, 2004.

[2] J. Fridlyand, A. M. Stransky, D. Pinkel, D. G. Albertson and A. N. Jain, "Hidden Markow models approach to the analysis of CGH data," J. Multivar. Anal., vol. 90, pp. 132-153, 2004.

[3] T. S. Price, et al., "SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data," Nucleic Acids Res., vol. 33, pp. 3455C3464, 2005.

[4] F. Picard, et al., "A statistical approach for array CGH data analysis," BMC Bioinformatics, vol. 6, pp. 27, 2005.

[5] L. Hsu, et al., "Denoising array-based comparative genomic hybridization data using wavelets," Biostatics, vol. 6, pp. 211-226.

[6] T. Huang, B. Wu, P. Lizardi and H. Zhao, "Detection of DNA copy number alterations using penalized least squares regression," Bioinformatics, vol. 21, pp. 3811-3817, 2005.

[7] P. J. Huber, *Robust Statistical Procedures.* Philadelphia: Society for Industrial and Applied Mathematics, 1977.

[8] J. Bai, "Estimation of a Change Point in Multiple Regression Models," Review of Economics and Statistics, vol. 79, pp. 551-563, 1997.

[9] J. H. Sullivan, "Estimating the Locations of Multiple Change Points in the Mean," Computational Statistics, vol. 17, pp. 289C296, 2002.

[10] D. M. Hawkins, "Fitting Multiple Change-Point Models to Data," Computational Statistics & Data Analysis, vol. 37, pp. 323-341, 2001.

[11] D. M. Hawkins, "Point Estimation of the Parameters of a Piecewise Regression Model," Applied Statistics, vol. 25, pp. 51-57, 1976.

[12] J. Bai and P. Perron, "Computation and Analysis of Multiple Structural Change Models," Journal of Applied Econometrics, vol. 18, pp. 1-22, 2003.

[13] R. L. Brown, J. Durbin, and J. M. Evans, "Techniques for Testing the Constancy of Regression Relationships over Time," Journal of the Royal Statistical Society, vol. B37, pp. 149-163, 1975.

[14] P. Wang, et al., "A method for calling gains and losses in array CGH data," Biostatistics, vol. 6, pp. 45-58, 2005.

[15] X. R. Yang, L. X. Zhang, "A note on self-weighted quantile estimation for infinite variance quantile autoregression models," Statistics and Probability Letters, vol. 78, pp. 2731-2738, 2008.

[16] X. He, D. G. Simpson and G. Y. Wang, "Breakdown points of t-type regression estimators," Biometrika, vol. 87, pp. 675-687, 2000.

[17] X. He, H. Cui and D. G. Simpson, "Longitudinal data analysis using t-type regression," J. Statist. Plann. Inference, vol. 122, pp. 253-269, 2004.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE