Detection of DNA Copy Number Variations Using Penalized Least Absolute Deviations Regression

Xiaoli Gao[1] and Jian Huang[1,2]

[1]Department of Statistics and Actuarial Science, [2]Department of Biostatistics, University of Iowa, Iowa City, IA 52246

December 2007

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 386

Abstract. Deletions and amplifications of the human genomic DNA copy number are the cause of numerous diseases such as various forms of cancer. Therefore, the detection of DNA copy number variations (CNV) is important in understanding the genetic basis of disease. Various techniques and platforms have been developed for genome-wide analysis of DNA copy number, such as array-based comparative genomic hybridization (aCGH) and high-resolution mapping using high-density tiling oligonucleotide arrays. Since complicated biological and experimental processes are involved in these platforms, data can be contaminated by outliers. Inspired by the robustness property of the LAD regression, we propose a penalized LAD regression with the fused lasso penalty for detecting CNV. This method incorporates the spatial dependence and sparsity of CNV into the analysis and is computationally feasible for high-dimensional array-based data. We evaluate the proposed method using simulation studies, which indicate that it can correctly detect the numbers and locations of the true breakpoints while controlling the false positives appropriately. We demonstrate the proposed method on two real data examples.

Key words and phrases. DNA copy number, break points, false discovery rate, lasso, fused lasso, LAD regression.

1 Introduction

Deletions and amplifications of the human genomic DNA copy number are the causes of numerous diseases. They are also related to phenotypic variation in the normal population. Therefore, the detection of DNA copy number variation (CNV) is important in understanding the genetic basis of disease such as various types of cancer. Several techniques and platforms have been developed for genome-wide analysis of

DNA copy number, including comparative genomic hybridization (CGH) (Kallioniemi et al. (1992)), array-based comparative genomic hybridization (aGCH) (Pinkel et al. (1998), Snijders et al. (2001)), commercially available single nucleotide polymorphism (SNP) arrays (Zhao et al. (2004)) and high-resolution mapping using high-density tiling oligonucleotide arrays (HR-CGH) (Urban et al. (2006)). These platforms have been used with microarrays. Each microarray consists of tens of thousands of genomic targets or probes, sometimes referred to as markers, which are spotted or printed on a glass surface. In an aCGH experiment a DNA sample of interest (test sample), and a reference sample are differentially labelled with dyes, typically Cy3 and Cy5, and mixed. The combined sample is then hybridized to the microarray and imaged, which results in test and reference intensities for all the markers. The goal of the analysis of DNA copy number data is to partition the whole genome into segments where copy numbers change between contiguous segments, and subsequently to quantify the copy number in each segment. Therefore, identifying the locations of copy number changes is the key step in the analysis of DNA copy number data.

Several methods have been proposed to identify the breakpoints of copy number changes. Jong et al. (2003) developed a genetic local search algorithm to localize the breakpoints along the chromosome. Olshen et al. (2004) proposed a binary segmentation procedure (CBS) to look for two breakpoints at a time by considering the segment as a circle. Fridlyand et al. (2004) used an unsupervised hidden markov model (HMM) approach to classify each chromosome into different states representing different copy numbers. Wang et al. (2005) proposed a hierarchical clustering algorithm to select interesting clusters by controlling the false discovery rate (FDR). Hsu et al. (2005) used a wavelets approach for denoising the data to uncover the true copy number changes. Lai and Zhao (2005) used a t-test to detect copy number alterations

by aggregating information from replicated arrays.

Of particular relevance to the proposed method is the work of Huang et al. (2005), which is the first to model the problem of CNV detection in the framework of penalized regression. This work used a least squares (LS) regression model with the least absolute penalty on the differences between the copy numbers of the neighboring markers. This method can be recast into a LS regression with the Lasso penalty, and thus is called the Lasso based (LB) method by the authors. The LB method imposes smoothness on the copy numbers along the chromosome. But it does not take into account the sparsity in the copy number variations. Here the sparsity means that there is only a small number of positions where changes occur in the copy numbers. In addition, because the LS regression is not robust, the LB method can be affected by the outliers.

Since complicated experimental processes are involved in a microarray experiment, data generated from such experiment can be contaminated by outliers. Inspired by the robustness property of the least absolute deviations (LAD) approach, we propose a penalized LAD regression with the fused lasso penalty for detecting CNV. We call this method LAD-FL. By use of the LAD loss function, the proposed method is resistant to outliers. By use of the fused Lasso penalty, it incorporates spatial dependence and sparsity of CNV data sets into the analysis.

The remainder of this paper is organized as follows. In Section 2, we describe the LAD-FL method and propose an approach for computing the LAD-FL estimator. In Section 3, we present a method for calculating false discovery rate (FDR) in the context of the LAD-FL method. In Section 4, we use simulations to evaluate the performance of LAD-FL. In Section 5, we analyze two CNV data sets to illustrate the proposed method. Concluding remarks are given in Section 6.

2 LAD regression with the fused Lasso penalty for CNV analysis

Consider an array of CGH profiles. For the ith profile, let y_{ij} be the log2 ratio of the intensities of the red over green channels of marker i on a chromosome, where the red and green channels measure the intensities of the test (e.g. cancer) and reference (e.g. normal) samples. We assume that the intensities have been properly normalized. Following Huang et al. (2005), the observed y_i can be considered a realization of the true relative copy number β_i at marker i plus a random noise,

$$y_i = \beta_i + \epsilon_i, \quad i = 1 \cdots n, \tag{2.1}$$

where n is the number of markers on a given chromosome. Our task is to make inference about β_i's based on the observed y_i's. There are three factors that should be taken into account. First, there may be outliers in the observed y_i's, so a robust procedure is needed. Second, the signals β_i's have the spatial dependence because the true copy numbers of the nearby markers are the same except in the regions where the copy numbers change abruptly. Third, copy number changes only occur at a few locations in the chromosome, most of the β_i's should be zero. Based on these considerations, we propose the criterion

$$\sum_{i=1}^{n} |y_i - \beta_i| + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=2}^{n} |\beta_i - \beta_{i-1}|, \tag{2.2}$$

where λ_1 and λ_2 are two tuning parameters determined by cross validation. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)'$. The estimate of $\boldsymbol{\beta}$ is the value $\widehat{\boldsymbol{\beta}}$ that minimizes (2.2). In this criterion, we use the absolute loss to reduce the influence of outliers. And the penalty we use is fused Lasso (Tibshirani et al. 2005). Therefore, we call $\widehat{\boldsymbol{\beta}}$ the LAD-FL estimator. In (2.2), the term $\sum_{i=2}^{n} |\beta_i - \beta_{i-1}|$ provides a measurement of the smoothness of the

parameters β_i's, which reflects the spatial dependence of the signals. Thus penalizing this smoothness measure forces the estimates of β_i's to be smooth. The term $\sum_{i=1}^{n} |\beta_i|$ is a Lasso penalty. Penalizing this term leads to sparse nonzero estimates.

2.1 Computation

We now describe our approach for computing $\widehat{\boldsymbol{\beta}}$. Let $\mathbf{y} = (y_1, \cdots, y_n)'$. Let $\mathbf{U}_{\lambda_1} = \text{diag}\,(\lambda_1/2, \lambda_1, \cdots, \lambda_1)$ be a $n \times n$ diagonal matrix. Define a $n \times n$ matrix

$$\mathbf{V}_{\lambda_1,\lambda_2} = \begin{bmatrix} \lambda_1/2 & 0 & 0 & \dots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_2 & \lambda_2 \end{bmatrix}.$$

Consider a new response vector $\mathbf{y}^* = (\mathbf{y}', \mathbf{0}', \mathbf{0}')'$ and a new design matrix $\mathbf{X}^* = [\mathbf{I}, \mathbf{U}'_{\lambda_1}, \mathbf{V}'_{\lambda_1,\lambda_2}]'$. Then (2.2) can be written as

$$L(\beta, \lambda_1, \lambda_2) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}|.$$

For every fixed λ_1 and λ_2, this is the objective function of a LAD regression problem. Therefore, we can use the existing programs such as the R `quantreg` package to compute $\widehat{\boldsymbol{\beta}}$.

2.2 Determining the tuning parameters

It is important to choose the tuning parameters λ_1 and λ_2 appropriately, which determine the smoothness and sparsity of the estimates $\widehat{\beta}_i$'s. In one extreme, if $\lambda_1 = 0$ and $\lambda_2 = 0$, then the estimate of β_i is simply y_i. This obviously leads to too many estimated non-zero relative ratios. In the other extreme, if λ_1 and λ_2 are very large, then all the $\widehat{\beta}_i$'s are forced to be zero regardless of the data, which is not reasonable.

We use a cross validation method to select (λ_1, λ_2). We divide the data sequence $\{y_1, \ldots, y_n\}$ into two subsequences. The first subsequence consists of y_i's with odd subscripts, that is, $\{y_1, y_3, \ldots, \}$. The second one consists of $\{y_2, y_4, \ldots\}$. We use the odd/even subsequence as the training set, and the left subsequence as the test set. We start from large enough λ_1 and λ_2, which will get all the 0 coefficients (it means that the breakpoint set is null). In our simulation and data analysis examples, we find that starting from $(\lambda_1, \lambda_2) = (1, 1)$ works well. We decrease the values of the two parameters step by step till $\lambda_1 = \lambda_2 = 0$, in that case there is no penalties. For example, by step length 0.1, we want to find the best combination of λ_1 and λ_2 in a square $[0, 1] \times [0, 1]$ such that the combination minimizes $\sum_{i=1}^{[n/2]} |y_{2i} - \widehat{y}_{2i-1}| + \sum_{i=1}^{[n/2]} |y_{2i-1} - \widehat{y}_{2i}|$, the sum of prediction error for test data.

3 Estimation of FDR

Using λ_1 and λ_2 selected by the cross-validation method described above, we compute $\widehat{\boldsymbol{\beta}}$. Let $\mu_i = \beta_i - \beta_{i-1}$ and $\widehat{\mu}_i = \widehat{\beta}_i - \widehat{\beta}_{i-1}$. Let $B = \{i : \widehat{\mu}_i \neq 0\}$, which is the set of locations where there is change in estimated relative copy numbers. The elements in B are potential breakpoints. However some of the nonzero estimation of the jumps may not be significant and can also lead to false positives. Similar to Huang et al (2005), we first use the stationary bootstrap sampling to test the significance of the elements in B, and then estimate the false discovery rate (FDR). The schemes for stationary bootstrap sampling and calculating FDR are presented in this section.

3.1 Stationary bootstrap sampling

Because of the dependence among the observations, standard bootstrap methods for the independent data are not applicable here. We use the stationary bootstrap sampling method (Politis and Romano (1994)). This method is designed for resampling from a stationary sample. In the present setting, we proceed as follows.

Step 1. We pick one observation randomly from the original observations, say y_{i_1}.

Step 2. With probability q, we select a new observation randomly from the original observation, say y_{i_2}. It could happen that y_{i_1}=y_{i_2}. And with probability $1-q$, we set y_{i_2} to be y_{i_1+1}.

Step 3. Repeat step 2 n times to obtain a bootstrap sample of size n.

Step 4. Repeat Steps 1, 2 and 3 to get N new data sets.

3.2 Estimation of the p-values and FDR

Suppose N stationary bootstrap data sets are generated. For marker i, we denote $\widehat{\mu}^*_{ik}$ as the estimation of μ_i from the kth bootstrap data set, and divide the N bootstrap data sets into three categories, K_{i0}, K_{i1} and K_{i2}, where

$$K_{i0} \equiv \{k:\ |\widehat{\mu}^*_{ik}| = |\widehat{\mu}_i|,\ 1 \le k \le N\},$$

$$K_{i1} \equiv \{k:\ |\widehat{\mu}^*_{ik}| > |\widehat{\mu}_i|,\ 1 \le k \le N\},$$

$$K_{i2} \equiv \{k:\ |\widehat{\mu}^*_{ik}| < |\widehat{\mu}_i|,\ 1 \le k \le N\}.$$

The p-value at marker i is calculated as

$$\widehat{p}_i = \frac{|K_{i0}|}{2N} + \frac{|K_{i1}|}{N},\ i = 1, \cdots, n.$$

The question of whether there is a significant copy number change at a position can be restated as a hypothesis testing problem. The null hypothesis is that marker

i does not belong to any gain/loss region. When all the positions are considered simultaneously, it becomes a multiple test problem. We use the FDR approach to adjust for multiple comparisons (Benjamini and Hochberg (1995)). The FDR is defined as the expectation of the proportion of false positive results. Let p be a cutoff value. We define

$$\begin{aligned}\widehat{\text{FDR}} &= \frac{\text{number of markers picked under null hypothesis}}{\text{number of markers picked in the observed data}} \\ &= \frac{p \times \text{total number of markers}}{\text{number of markers whose p}-\text{values are less than } p}\end{aligned}$$

as the estimator of FDR (Storey (2002) and Efron and Tibshirani (2002)).

In our study, we choose $q = 0.35$ and $p = 0.01$ if there is no other specification.

4 Simulation studies

We evaluate the performance of the LAD-FL method for detecting CNV using three simulation examples. Suppose there are 1000 markers equally spaced along a chromosome. All observed log2 rations are generated from

$$y_i = \beta_{0i} + \epsilon_i, \quad i = 1, \cdots, 1000, \tag{4.1}$$

where β_{0i}'s are the true log2 ratios of these 1000 markers. They are set up to have three altered regions along the chromosome which correspond to quadraploid, triploid and monoploid states, respectively. All random noises ϵ_i's are simulated from three different models as AR(2) model, AR(1) model and independent model.

Example 1. This example uses the Example 1: This example uses the same set-up as in Huang et al. (2005). All β_{0i}'s are given in Table 1. The standard deviations of ϵ in

Table 1: The true log2 ratios corresponding to three altered regions along the chromosome in Example 1 and 2.

i	1-100	101-150	151-450	451-600	601-900	901-1000
β_{0i}	0	1	0	0.59	0	-1

these three models are the same.

$$\epsilon_i = e_{i0}, \qquad e_{i0} \sim N(0, 0.15^2),\ i = 1, \cdots, 1000.$$
$$\epsilon_i = 0.60\epsilon_{i-1} + e_{i1}, \qquad e_{i1} \sim N(0, 0.12^2),\ i = 1, \cdots, 1000.$$
$$\epsilon_i = 0.60\epsilon_{i-1} + 0.20\epsilon_{i-2} + e_{i2}, \quad e_{i2} \sim N(0, 0.10^2),\ i = 1, \cdots, 1000.$$

Example 2: To evaluate the robustness property of the LAD-FL estimator, we simulate e_{ij}'s from double exponential (dbexp) distributions in three models. And the standard deviations of ϵ in these three models are also the same. Example 2 shares the same β_{0i}'s with Example 1.

$$\epsilon_i = e_{i0}, \qquad e_{i0}/18 \sim \text{ standard dbexp},\ i = 1, \cdots, 1000.$$
$$\epsilon_i = 0.77\epsilon_{i-1} + e_{i1}, \qquad e_{i1}/17 \sim \text{ standard dbexp},\ i = 1, \cdots, 1000.$$
$$\epsilon_i = 0.60\epsilon_{i-1} + .20\epsilon_{i-2} + e_{i2}, \quad e_{i2}/18 \sim \text{ standard dbexp},\ i = 1, \cdots, 1000.$$

Example 3. In order to demonstrate the performance of the LAD-FL method under both sparsity and smoothness conditions, for all three models in Example 2, we set the true log2 ratios β_{0i}'s to be extremely sparse as given in Table 2.

Table 2: The the true log2 ratios corresponding to three altered regions along the chromosome in Example 3.

i	1-100	101-110	111-450	451-460	461-980	981-1000
β_{0i}	0	1	0	0.59	0	-1

To take into account the unevenly distributed density of markers along the chromosome, we draw one marker from every two markers randomly. Thus we create two subsets with 500 non-equally spaced markers for each data set in all three examples. We simulate 20 data sets for every model in each example. To demonstrate the robustness property of the LAD-FL method, we generate 2 outliers in each data set. Both LB and LAD-FL are applied to detect the number of the significant breakpoints. We run stationary bootstrap sampling 1000 times. The simulation results are given in Table 3.

For each example, we calculate the average number and its standard deviation (in the parenthesis) of all detected breakpoints in each of the 20 data sets. They are listed in the first row of each method. We then calculate the average number of detected breakpoints within and beyond 2 markers of the true breakpoints in every data set. They are listed in the second row of each method. We also count the total number of outliers which are falsely identified as breakpoints. They are listed in the third row of each method.

The results in Table 3 show that the LAD-FL method is more robust than the LB method in all three models. The LAD-FL method detects breakpoints more accurately than the LB method. We can also see that the LB method tends to identify outliers as breakpoints, especially for those data sets that are very sparse (e.g. Example 3). Figure 1 plots two simulated sparse data sets generated from AR(1) and independent models in Example 3. There are two outliers in each data set. All detected breakpoints are indicated by vertical lines. The LB method catches falsely the two outliers as breakpoints in this data set. The LB method catches falsely two outliers as breakpoints in these data sets. In addition to those outliers, the LB model also detects more breakpoints falsely than the LAD-FL method. This is not surprising since unlike

Table 3: Simulation results. There are 5 true break points out of 500 non-equally spaced markers for each data set. In total, there are 40 outliers for all 20 data sets in each scenario.

Methods	**LAD-FL**			**LB**		
Models	**AR(2)**	**AR(1)**	**Ind.**	**AR(2)**	**AR(1)**	**Ind.**
Example 1	4.8(0.41)[1]	4.9 (0.72)	4.5 (0.61)	5.9 (0.85)	6.3 (1.18)	7.85 (1.53)
	4.8[2], 0[3]	4.85, 0.1	4.5, 0	5.45, 0.45	5.35, 0.95	5.25, 2.6
	No outlier[4]	1 outlier	No outlier	7 outliers	8 outliers	15 outliers
Example 2	4.85 (0.49)	4.75 (0.64)	4.5 (0.61)	5.35 (0.76)	5.45 (0.76)	6.9 (1.37)
	4.8, 0.05	4.65, 0.1	4.35, 0.15	4.95, 0.4	5.05, 0.4	6.0, 0.9
	No outlier	1 outlier	2 outliers	4 outliers	3 outliers	11 outliers
Example 3	5.05 (0.51)	5.05 (0.82)	5.45 (0.76)	6.15 (2.28)	7.6 (1.39)	9.35 (1.42)
	4.9, 0.15	4.65, 0.4	5.0, 0.45	3.9, 2.25	4.95, 2.65	5.75, 3.6
	No outlier	No outlier	3 outlier	9 outliers	17 outliers	31 outliers

[1]The average number (with standard deviation) of all detected breakpoints; [2] The detected breakpoints within 2 markers of the true breakpoints on average; [3]The detected breakpoints beyond 2 markers of the true breakpoints on average; [4] The number of outliers which are falsely identified as breakpoints out of total 40 outliers .

LAD-FL, LB does not take into account the sparsity property of the data set.

5 Two real data sets

To illustrate the LAD-FL method, we analyze two datasets. Following Huang et al. (2005), we apply two empirical conditions for a marker to be a possible breakpoint.

(i) The difference of the means of the two subsegments is greater than .35.

(ii) At least one of the subsegments has a mean greater than .35.

We apply the LAD-FL method to detect breakpoints as follows,

S1. First we use the cross-validation method as given in Section 2 to choose the tuning parameters. Once the tuning parameters Once the tuning parameters are chosen, we can compute $\widehat{\beta}_i$'s. Let $\widehat{\mu}_i = \widehat{\beta}_i - \widehat{\beta}_{i-1}$, for $i > 1$ and $\widehat{\mu}_1 = \widehat{\beta}_1$. Those nonzero $\widehat{\mu}_i$'s satisfying the empirical conditions (i) and (ii) are considered to be the candidates of breakpoints.

S2. Using the stationary bootstrap sampling method to get $N = 500$ bootstrap data sets. From these new data sets, we obtain N estimates of the jump value at each marker using the chosen tuning parameters in S1. Then we use the method in Section 3 to estimate the p-value and choose all significant breakpoints.

5.1 Bacterial Artificial Chromosome (BAC) array

The BAC data set is generated by Snijders et al. (2001). It consists of single experiments on 15 fibroblast cell lines. Each array contains measurements for 2276 mapped BACs spotted in triplicates. There were either one or two alterations in each cell line as identified by spectral karyotyping. There were 15 chromosomes with partial alterations and 8 whole chromosomal alterations. The variable used for analysis is the normalized average of the log2 ratio of test sample over reference sample, as processed by the authors.

Snijders et al. (2001) used spectral karyotyping to confirm that there are 15 chromosomes of BAC array data set which have partial alterations. The LAD-FL method identifies 14 partial chromosomal alterations of them except the one on chromosome 15 in GM07081. See (f) in Figure 3. By using the LB method in Huang

(2005), five single points are detected as breakpoints. But they are not confirmed in Snijders et al. (2001). The LAD-FL method does not identify them as breakpoints either. These five single points are from chromosome 8 of GM13031, chromosome 8 of GM01535, choromsome 8 of GM05296, choromsome 22 of GM13330 and choromsome 23 of GM07081. See (a)–(f) in Figure 3. Figure 4 shows that, like LB, LAD-FL also identifies breakpoints from chromosome 23 of GM01535 and GM05296, which are not confirmed by spectral karotyping.

5.2 Human chromosome 22q11 data

Urban et al. (2006) applied high-resolution CGH (HR-CGH) technology to the analysis of CNV on chromosome 22q11. The DNA samples are collected from patients who have Cat-Eye syndrome, 22q11 deletion syndrome (also called velocardiofacial syndrome or DiGeorge syndrome) and some other symptoms. A large proportion of 22q11DS patients develop learning disabilities and attention-deficit hyperactivity disorder and there are large variations in the symptoms of these patients. For example, patients 03-154 and 97-237 have the typical LCR $A \rightarrow D$ deletion, but they exhibit considerable variation in their symptoms which may be associated with the deletion size. Therefore, it is very important to apply a method to accurately detect the sizes of deletion regions.

These Human chromosome 22q11 data sets consist of the measurements on chromosome 22 of 12 patients. There are about 372,000 features in the microarray datasets for each patient. In order to apply the LAD-FL method, we partition the whole chromosome into several segments and then apply the method to each segment. We set the cutoff value p to be 0.001 in our analysis. The LAD-FL method is able to identify all the blocks with break points detected in Urban et al. (2006). It can also detect the accurate breakpoints for DNA block deletion and amplification. For

example, Figure 5 show the results for the data from patients 03-154 and 97-237. This plot suggests that the deletion sizes for these two patients are different. In addition, Patient 03-154 appears to have another deletion region, which was not detected by Urban et al. (2006).

6 Concluding remarks

An appealing feature of the proposed LAD-FL method is its resistance against outliers. This robustness property is inherited from the LAD regression, which is useful in reducing the possibility of false positive findings due to outlying intensity measurements. This property is demonstrated in the generating models used in our simulation studies. The fused Lasso penalty in the LAD-FL method incorporate both sparsity and smoothness properties of copy number data. Computationally, the LAD-FL estimator can be computed using the existing efficient programs for LAD regression, since both the loss and penalty functions use the same L_1 norm. While our simulation studies and real data analysis indicate that the LAD-FL method is a useful and robust approach for CNV analysis, there are some important questions that call for further work. For example, in the proposed LAD-FL method, it is assumed that the intensity data have already been properly normalized. It would be useful to examine how sensitive the method is to different normalization methods, or perhaps consider the possibility of incorporating normalization into an integrated model. Another interesting question is regarding the theoretical properties of LAD-FL. It would be interesting to consider under what conditions on the smoothness and sparsity of the underlying copy number process the LAD-FL is able to correctly detect the breakpoints with high probability.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57**, 289–300.

[2] Efron, B., Hastie, T., Johnstone, I. and Tibshirani R. (2004). Least angle regression. *Annals of Statistics*, **32**, 407–499.

[3] Efron, B. and Tibshirani R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70–86.

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

[5] Hsu, L., Self, S.G., Grove, D., Randolph, T., Wang, K., Delrow, J.J., Loo, L. and Porter, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

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

[7] Jong, K., Marchiori, E., Van der Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. *Applications of Evolutionary Computing. Springer LNCS 2611*, 107–116.

[8] Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

[9] Lai, Y.L. and Zhao, H.Y. (2005). A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. *Computational Biology and Chemistry*, **29**, 47–54.

[10] 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.

[11] Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. and Albertson, D.G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*, **20**, 207–211.

[12] Politis, D.N. and Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, **89**, 1303–1313.

[13] Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D. and Alberston, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, **29**, 263–264.

[14] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, **64**, 479–498.

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, **58**, 267–288.

[16] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, **67**, 91–108.

[17] Urban, A.E., Korbel, J.O., Selzer, R., Richmond, T., Hacker, A., Popescu, G.V., Cubells, J.F., Green, R., Emanuel, B.S., Gerstein, M.B., Weissman, S.M. and Snyder, M. (2006). High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *PNAS*, **103**, 4534–4539.

[18] Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

[19] Zhao, X.J., Li, C., Paez, J.G., Chin, K., Jänne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C., Gray, J.W., Sellers, W.R., and Meyerson, M. (2004). An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*. **64**, 3060–3071.

Figure 1: Compare LAD-FL and LB for AR1 model of a dataset in Example 3.

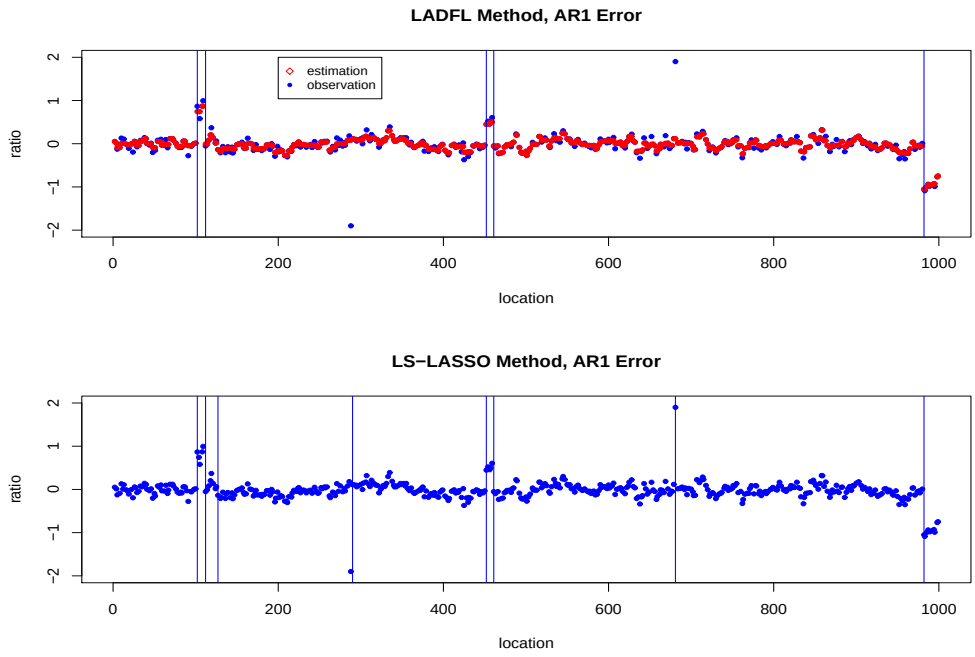

Figure 2: Compare LAD-FL and LB for independent model of a dataset in Example 3.

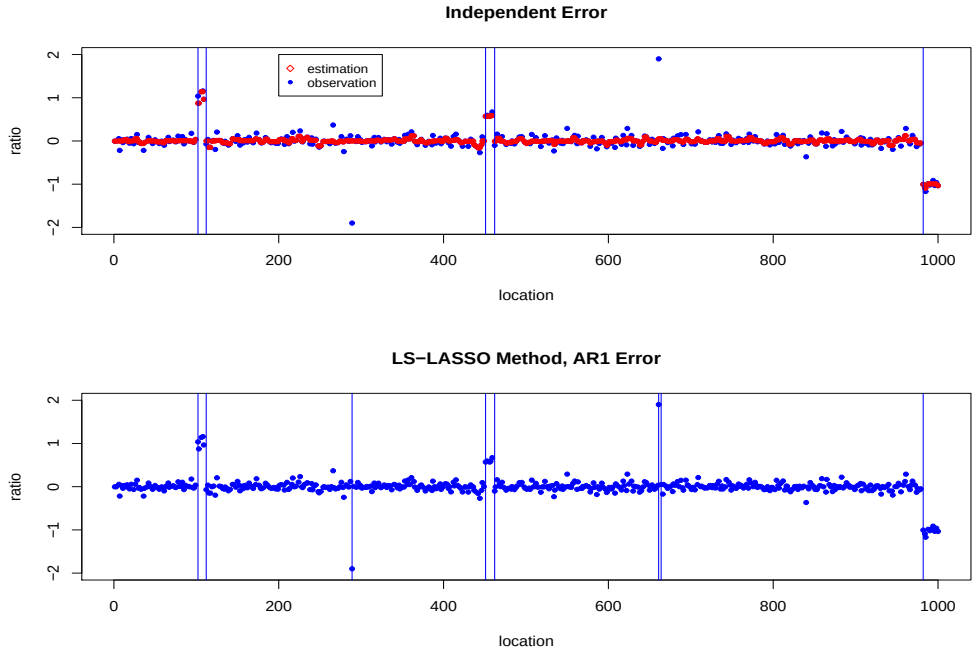

Figure 3: In (a)−(e), LAD-FL is consistent with spectral karyotyping, no breakpoint is detected. However LB detected 5 single points as breakpoints in these choromsomes. Not like spectral karyotyping, neither LAD-FL nor LB detects any breakpoint in (f). Blue dots are the observation. Red dots are the estimates from LAD-FL.

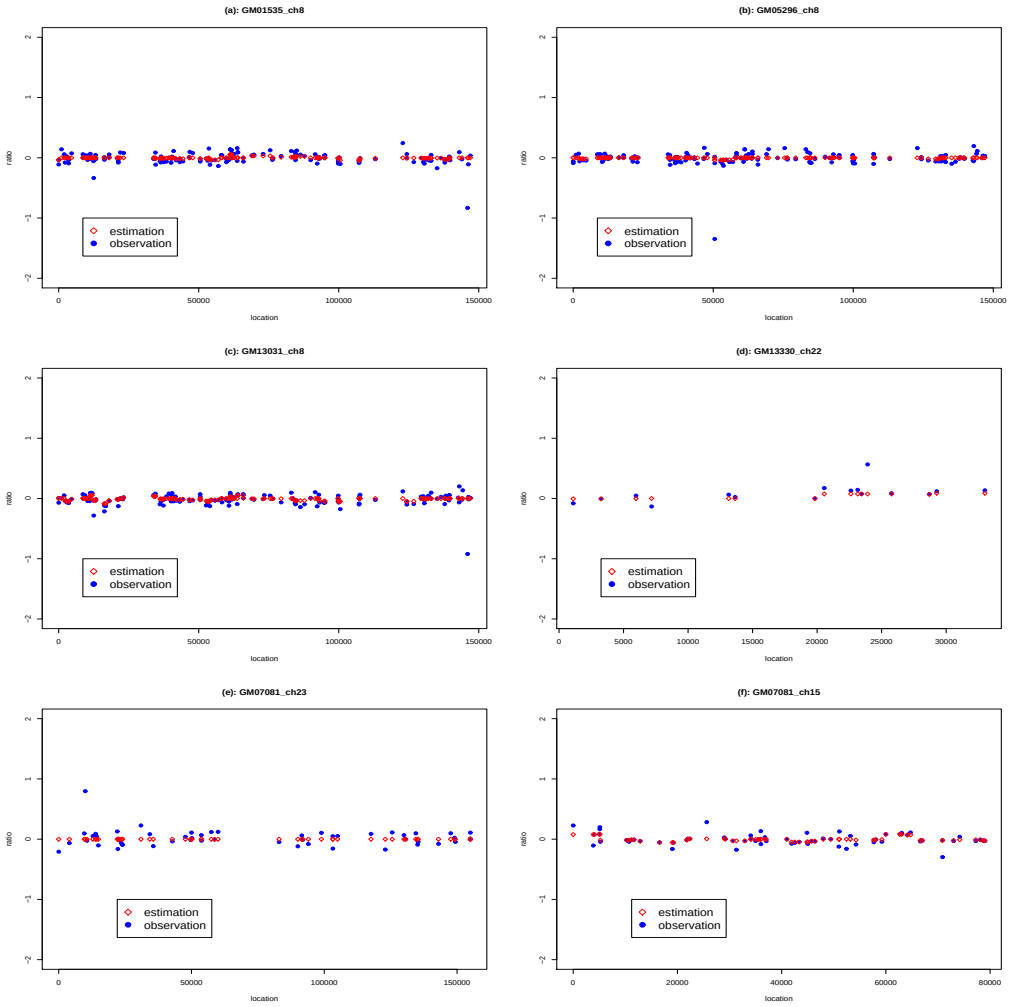

Figure 4: Both LAD-FL and LB detect alterations in these 2 chromosomes. Blue dots are the observation. Red dots are the estimates by using the LAD-FL method. The vertical line marked the location for detected breakpoints by from LAD-FL.

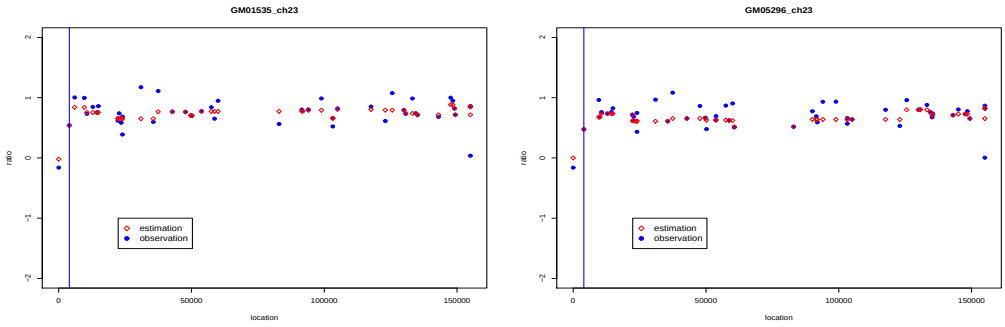

Figure 5: Human Chormosom 22q11 datasets of patient 03-154 and patient 97-237. The first two plots are the observations of 180,000 markers. The last two plots are the observations (gray) and estimates (blue) from LAD-FL.