BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 18 2007, pages 2470–2476
doi:10.1093/bioinformatics/btm364

Data and text mining

Analysis of array CGH data for cancer studies using fused quantile regression

Youjuan Li and Ji Zhu*

Department of Statistics, University of Michigan, Michigan, USA

Received on March 25, 2007; revised on June 12, 2007; accepted on July 8, 2007

Advance Access publication July 20, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The identification of DNA copy number changes provides insights that may advance our understanding of initiation and progression of cancer. Array-based comparative genomic hybridization (array-CGH) has emerged as a technique allowing high-throughput genome-wide scanning for chromosomal aberrations. A number of statistical methods have been proposed for the analysis of array-CGH data. In this article, we consider a fused quantile regression model based on three motivations: (1) quantile regression may provide a more comprehensive picture for the ratio profile of copy numbers than the standard mean regression approach; (2) for simplicity, most available methods assume uniform spacing between neighboring clones, while incorporating the information of physical locations of clones may be helpful and (3) most current methods have a set of tuning parameters that must be carefully tuned, which introduces complexity to the implementation.

Results: We formulate the detection of regions of gains and losses in a fused regularized quantile regression framework, incorporating physical locations of clones. We derive an efficient algorithm that computes the entire solution path for the resulting optimization problem, and we propose a simple estimate for the complexity of the fitted model, which leads to convenient selection of the tuning parameter. Three published array-CGH datasets are used to demonstrate our approach.

Availability: R code are available at http://www.stat.lsa.umich.edu/~jizhu/code/cgh/

Contact: jizhu@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Chromosomal aberrations such as deletions, amplifications and structural rearrangements are hallmarks of cancer (Lengauer *et al.*, 1998; Pinkel and Albertson, 2005). Therefore, identifying genomic regions associated with systematic aberrations provides insights into the initiation and progression of cancer, and improves the diagnosis, prognosis and therapy strategies.

In recent years, array-based comparative genomic hybridization (array-CGH) has emerged as a technique offering genome-wide scanning for chromosomal aberrations. In this approach, genomic DNA sequences from tumor and normal reference samples are labeled by different color fluorochrome, and then hybridized to the array consisting of hundreds to thousands of cloned DNA fragments with known exact chromosomal locations. If both samples contain the same quantity of target sequences, they bind equally to the complementary sequences, resulting fluorescent signal ratio equal to 1. In contrast, when the tumor sample contains DNA with duplication (deletion) or other gain (loss) in a particular genomic region, more (less) tumor DNA bind to the complementary sequence, and the ratio is greater (smaller) than 1. Therefore, the ratio represents the relative DNA aberration at a particular genomic region.

A number of statistical methods have been developed to analyze the ratio profile from array-CGH. For instance, a method based on robust locally weighted regression (lowess) has been previously used in Beheshti *et al.* (2003). A quantile smoothing method using the total variation as the roughness penalty has been shown to give desirable visualization of the CGH profile in Eilers and Menezes (2004). Other smoothing algorithms that denoise the array-CGH data and appear suited for handling abrupt changes in the profile include adaptive weights smoothing (Hupe *et al.*, 2004; Polzehl and Spokoiny, 2000), wavelets (Hsu *et al.*, 2005) and lasso (Huang *et al.*, 2005). Jong *et al.* (2004) use a genetic local search algorithm to examine the CGH profile by maximizing a likelihood with a penalty term containing the number of breakpoints. Olshen *et al.* (2004) propose a circular binary segmentation (CBS) method that detects aberrations by recursively using a likelihood ratio test statistics. Other methods include the hidden Markov model (HMM) which intends to model the possible dependence of a clone with its near neighbors (Fridlyand *et al.*, 2004), and a cluster along chromosomes (CLAC) method which builds hierarchical clustering-style trees along the chromosome and then selects the 'interesting' clusters via certain criterion (Wang *et al.*, 2005).

One piece of information not considered in most of these methods is the physical position of the clone along the genome, and uniform spacing between neighboring clones is often assumed. However, incorporating this information can only help if done correctly (Lai *et al.*, 2005). We also note that most of these methods involve tuning parameters that need to be carefully selected. A key issue in the implementation is how to appropriately select these tuning parameters. If an algorithm is

*To whom correspondence should be addressed.

very sensitive to the values of the tuning parameters or if the complexity of the algorithm does not allow the user to determine the tuning parameters easily, it may be viewed as a weakness of the algorithm (Lai *et al.*, 2005).

In this article, we modify the fused quantile regression model in Eilers and Menezes (2004) by incorporating the physical distance between adjacent clones. To solve the resulting optimization problem, we derive an efficient algorithm that computes the entire solution path for *all* values of the tuning parameter. Furthermore, we propose a convenient measure for the complexity of the fitted model, which facilitates selection of the tuning parameter.

Let $y_i \in \mathbb{R}$ denote the $\log_2$-ratio of the normalized signal between the tumor and the reference samples at clone i, and $x_i \in R$ be the physical position of the corresponding clone. If the entire hybridization and measurement process are well behaved, the signal ratio will reflect the relative copy numbers, i.e.

$$y_i = \mu_i + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where μ_i is the true $\log_2$-ratio of DNA copy numbers of tumor over reference samples at position i, and ϵ_i is a random noise.

Similar (but not identical) to Eilers and Menezes (2004), we formulate the detection of DNA copy number changes as a fused quantile regression problem (fused-QR). Let $\tau \in (0, 1)$ be the quantile of interest, and f_i be the smooth series that approximate y_i, specifically, the $100\tau\%$ quantile of the $\log_2$-ratio at clone x_i. We then estimate f_i via:

$$\min_{f_i}\left(\sum_{i=1}^{n}\rho_\tau(y_i - f_i) + \lambda\sum_{i=2}^{n}\left|\frac{f_i - f_{i-1}}{x_i - x_{i-1}}\right|\right). \tag{2}$$

We explain the three parts of criterion (2) as follows:

- The first term $\rho_\tau(.)$ is the so-called check function that measure the closeness of f_i to y_i (Koenker and Bassett, 1978):

$$\rho_\tau(y_i - f_i) = \begin{cases} \tau \cdot (y_i - f_i) & \text{if } y_i - f_i > 0 \\ -(1-\tau)\cdot(y_i - f_i) & \text{otherwise.} \end{cases} \tag{3}$$

A median curve with $\tau = 0.5$ can show the trend of the ratio profile, and the lower and upper quantile curves can give the spread of the ratios.

- The second term is a so-called penalty that measures the smoothness of f_i. Since copy number changes involve chromosome segments, the ratios of adjacent clones should be similar. Hence, we discourage changes in adjacent clones by penalizing $|f_i - f_{i-1}|$, adjusted by the distance between clones, i.e. $|x_i - x_{i-1}|$. A similar penalty without $|x_i - x_{i-1}|$ is called the fused penalty in Tibshirani *et al.* (2005), and it has also been used in Eilers and Menezes (2004) and Huang *et al.* (2005) for smoothing array-CGH data.

  The fused penalty in (2) offers parsimony: making λ sufficiently large will cause some of the fitted $|f_i - f_{i-1}|$ to be *exactly* 0. When $|f_i - f_{i-1}|$ is not equal to 0, it corresponds to a 'jump' in the ratio profile. So the fitted ratio profile will consist of flat plateaus, flat valleys and sharp jumps. A sharp jump corresponds to the beginning or the end of a copy number aberration.
- $\lambda > 0$ is a tuning parameter that controls the balance between the smoothness of the fitted model and its fidelity to the data: the larger the λ, the smoother the f_i, but at the cost of worse fit to the data, and vice versa. Hence, selecting an appropriate value of the tuning parameter is crucial for the performance of the fitted model.

The remainder of the article is organized as follows. In Section 2, we first present an equivalent version of (2), then derive an algorithm for computing the entire solution path for this model, and also propose a convenient method for selecting the tuning parameter. In Section 3, we apply the proposed approach to three published array-CGH datasets. We summarize the article in Section 4.

2 METHODS

2.1 Statistical model

We denote $\beta_0 = f_1$, $\beta_j = (f_{j+1} - f_j)/(x_{j+1} - x_j)$, $j = 1, \ldots, n-1$, then

$$f_i = \beta_0 + \sum_{j=1}^{i-1}\beta_j(x_{j+1} - x_j). \tag{4}$$

Using this new set of parameters $\{\beta_j\}_{j=0}^{j=n-1}$, (2) can be re-written as:

$$\min_{\beta_0,\beta_j}\sum_{i=1}^{n}\rho_\tau(y_i - f_i) + \lambda\sum_{j=1}^{n-1}|\beta_j|, \tag{5}$$

or equivalently,

$$\begin{aligned} \min_{\beta_0,\beta_j} \quad & \sum_{i=1}^{n}\rho_\tau(y_i - f_i) \\ \text{subject to} \quad & \sum_{j=1}^{n-1}|\beta_j| \le s, \end{aligned} \tag{6}$$

where s is a tuning parameter equivalent to λ. Note that (5) and (6) are in a form of L_1 loss plus L_1 penalty. For a given value of the tuning parameter, the optimization can be transformed into a linear programming problem, hence solved efficiently by most commercial packages (Eilers and Menezes, 2004). However, instead of solving the problem for *one* value of the tuning parameter, here we are interested in solving for the entire solution path for all values of the tuning parameter. This will facilitate the selection of the tuning parameter. In the next section, we show that the solution path $\beta(s)$ is piecewise linear in s, which allows us to derive an efficient algorithm to compute the entire $\beta(s)$.

Before delving into technical details, we illustrate the concept of piecewise linearity of the solution path by a simple example. We simulate a ratio profile of thirty clones:

$$y_i = \mu_i + \epsilon_i, i = 1, \ldots, 30, \tag{7}$$

where ϵ_i are independent Gaussian noise with $\sigma = 0.15$. We let $\mu_i = 1$ for $i = 11, \ldots, 20$, indicating gains at the second 10 clones, and $\mu_i = 0$ for all other clones. We fit the fused-QR model with $\tau = 0.5$. Figure 1 shows the solution path of $\beta(s)$ as a function of s, and the final fitted median curve with an appropriately chosen s.

2.2 The path algorithm

In this section, we outline the essential components of the path algorithm and leave all details in the Supplementary Material.

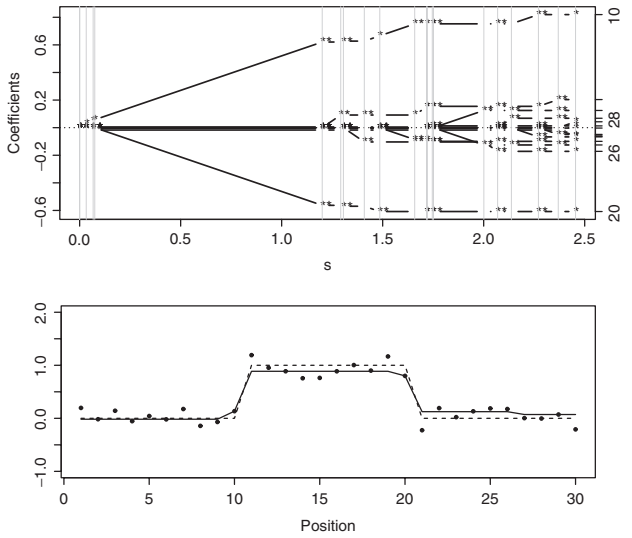

Fig. 1. Illustrative example ($\tau = 0.5$). Upper panel: the solution path $\beta(s)$ as a function of s, which is piecewise linear. Lower panel: the simulation data (dots), the fitted median curve with an appropriately tuned s (solid), and the truth (dashed).

We re-write (6) as:

$$\begin{aligned} \min_{\beta_0,\beta_j} \quad & \tau\sum_{i=1}^{n}\xi_i + (1-\tau)\sum_{i=1}^{n}\zeta_i \\ \text{subject to} \quad & \sum_{j=1}^{n-1}|\beta_j| \le s \\ & -\zeta_i \le y_i - f_i \le \xi_i \\ & \zeta_i, \xi_i \ge 0, i = 1,\ldots,n. \end{aligned} \tag{8}$$

The above gives the Lagrangian primal function:

$$\begin{aligned} L_p \quad : \quad & \tau\sum_{i=1}^{n}\xi_i + (1-\tau)\sum_{i=1}^{n}\zeta_i + \lambda\left(\sum_{j=1}^{n-1}|\beta_j| - s\right) + \\ & \sum_{i=1}^{n}\alpha_i(y_i - f_i - \xi_i + \delta_i^2) + \\ & \sum_{i=1}^{n}\gamma_i(-y_i + f_i - \zeta_i + \upsilon_i^2) + \\ & \sum_{i=1}^{n}\kappa_i(-\xi_i + \varepsilon_i^2) + \sum_{i=1}^{n}\eta_i(-\zeta_i + \iota_i^2), \end{aligned} \tag{9}$$

where λ, α_i, γ_i, κ_i, η_i are non-negative Lagrangian multipliers. Setting the derivatives of L_p to zero, we achieve

$$\frac{\partial}{\partial\beta} \quad : \quad \lambda \cdot \text{sign}(\beta_j) = \sum_{i>j}(\alpha_i - \gamma_i)d_j, \text{ for } \beta_j \ne 0, \tag{10}$$

$$\frac{\partial}{\partial\beta_0} \quad : \quad \sum_{i=1}^{n}(\alpha_i - \gamma_i) = 0, \tag{11}$$

$$\frac{\partial}{\partial\xi_i} \quad : \quad \tau = \alpha_i + \kappa_i, \tag{12}$$

$$\frac{\partial}{\partial\zeta_i} \quad : \quad 1 - \tau = \gamma_i + \eta_i, \tag{13}$$

where $d_j = x_{j+1} - x_j$, and the Karush–Kuhn–Tucker (KKT) conditions are

$$\alpha_i(y_i - f_i - \xi_i) = 0, \tag{14}$$

$$\gamma_i(y_i - f_i + \zeta_i) = 0, \tag{15}$$

$$\kappa_i\xi_i = 0, \tag{16}$$

$$\eta_i\zeta_i = 0. \tag{17}$$

Since the Lagrange multipliers must be non-negative, we conclude from (12) and (13) that both $0 \le \alpha_i \le \tau$ and $0 \le \gamma_i \le 1 - \tau$. We can also see that when $y_i - f_i > 0$ (hence $\xi_i > 0$), we have $\alpha_i = \tau$ and $\gamma_i = 0$; when $y_i - f_i < 0$ ($\zeta_i > 0$), we have $\alpha_i = 0$ and $\gamma_i = 1 - \tau$. These lead to the following relationships:

$$\begin{array}{llllll} y_i - f_i > 0 & \Rightarrow & \xi_i > 0, & \zeta_i = 0, & \alpha_i = \tau, & \gamma_i = 0; \\ y_i - f_i < 0 & \Rightarrow & \xi_i = 0, & \zeta_i > 0, & \alpha_i = 0, & \gamma_i = 1 - \tau; \\ y_i - f_i = 0 & \Rightarrow & \xi_i = 0, & \zeta_i = 0, & \alpha_i \in [0,\tau], & \gamma_i \in [0, 1-\tau]. \end{array}$$

Let $\theta_i = \alpha_i - \gamma_i$. Using these relationships, we can define the following four sets which are useful for the path algorithm:

- $\mathcal{E} = \{i : y_i - f_i = 0, -(1-\tau) \le \theta_i \le \tau\}$ (elbow)
- $\mathcal{L} = \{i : y_i - f_i < 0, \theta_i = -(1-\tau)\}$ (left of the elbow)
- $\mathcal{R} = \{i : y_i - f_i > 0, \theta_i = \tau\}$ (right of the elbow)
- $\mathcal{V} = \{j : \beta_j \ne 0\}$ (active set)

Since our goal is to compute the solution path $\beta(s)$, we are interested in how the KKT conditions change when the parameter s increases. As s increases, we define an *event* to be

(1) a data point hits the elbow, i.e. a residual $y_i - f_i$ changes from non-zero to zero, or

(2) a coefficient β_j changes from non-zero to zero, i.e. an index leaves $\mathcal{V}$.

Notice that these two events correspond to the non-smooth points of $\sum_i \rho_\tau(y_i - f_i)$ and $||\beta||_1$, respectively. Then we can see:

- As s increases, the sets $\mathcal{V}$, $\mathcal{L}$, $\mathcal{R}$ and $\mathcal{E}$ will not change (or equivalently, the KKT conditions will not change), unless an event happens. When the KKT conditions do not change, for uniqueness of the solution, we have the number of non-zero coefficients equal to the number of observations on the elbow, i.e. $|\mathcal{E}| = |\mathcal{V}|$ according to (10) and (11).
- As s increases, points in $\mathcal{E}$ stay at the elbow, unless an event happens. Therefore, β_j satisfy:

$$y_i - (\beta_0 + \sum_{j\in V, j<i}\beta_j d_j) = 0 \text{ for } i \in \mathcal{E}. \tag{18}$$

Since $|\mathcal{V}| = |\mathcal{E}|$, there is one free unknown in this set of equations, which allows β to change linearly when s increases, unless an event happens. Thus, overall the entire solution path $\beta(s)$ is piecewise linear.

The basic idea of our algorithm is as follows: we start with $s = 0$ and increase it, keeping track of the location of all data points relative to the elbow and also of the magnitude of the fitted coefficients along the way. As s increases, by continuity, points in $\mathcal{E}$ must linger on the elbow. Since all points at the elbow have $y_i - f_i = 0$, we can establish a path for β. The elbow set will stay stable until either some other point comes to the elbow or one non-zero fitted coefficient becomes zero. The path terminates when we reach the interpolating solution.

2.3 Selection of the tuning parameter

Our path algorithm facilitates selection of the tuning parameter among all possible values. In this section, we propose a convenient approach for this selection. We choose to use the Schwarz information criterion (SIC) (Schwarz, 1978), which is common in the quantile regression literature (Koenker *et al.* 1994):

$$\text{SIC}(s) = \ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - f_i)\right) + \frac{\ln n}{2n}df, \tag{19}$$

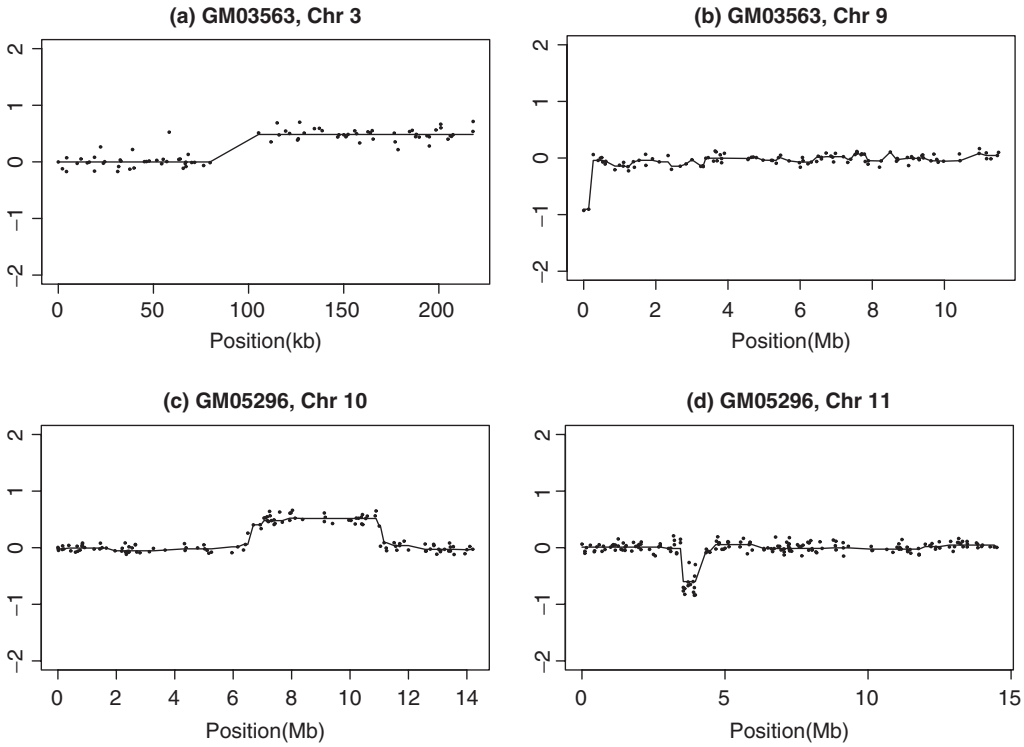

Fig. 2. Examples of the application of the fused-QR to the fibroblast cell lines ($\tau = 0.5$): (**a**) a copy gain at one side of the chromosome; (**b**) two altered points at one end; (**c**) a copy gain in the middle of the chromosome and (**d**) a copy deletion in the middle of the chromosome.

where *df* is a measure of the complexity of the fitted model. Stein (Stein, 1981), Efron (Efron, 1986) and several other researchers argue that *df* can be estimated by $\sum_{i=1}^{n} \partial f_i/\partial y_i$. We have been able to prove that in the fused-QR setting, $\sum_{i=1}^{n} \partial f_i/\partial y_i$ is equal to the number of non-zero β_j's in the fitted model (details of the proof are in the Supplementary Material). Furthermore, according to (4), a non-zero β_j corresponds to a jump between adjacent clones in the fitted model. Therefore, the number of fitted jumps gives a convenient estimate for *df*.

In practice, we can first use the path algorithm to compute the entire solution path, then select s that minimizes the SIC.

3 APPLICATIONS

In this section, we apply our method to three real datasets.

3.1 Fibroblast cell line data

This is a widely analyzed BAC array dataset provided by Snijders *et al*. (2001). The arrays consist of approximately 2400 BAC clones and provide precise measurement with an SD of ~0.05–0.1 $\log_2$-ratio. Single experiments were conducted on 15 fibroblast cell lines and the true copy number alterations were previously characterized by cytogenetics. One attractive feature of using this dataset is to prove the validation of our method since the true copy number changes are known.

We applied our fused-QR algorithm to four chromosomes: chromosome 3 and 9 on GM03563, and chromosome 10 and 11 on GM05296. Figure 2 shows the fitted median curves for the four datasets. As we can see, all four chromosomes show partial chromosomal alterations. These alterations detected by the fused-QR median curves, compared with the results from Snijders *et al*. (2001), agree with the cytogenetic analysis very well.

3.2 Colorectal cancer data

The development and progression of colorectal cancer is a multi-step process leading to genomic alterations in tumors. Nakao *et al*. (2004) report the results of array-CGH in a set of 125 primary colorectal tumors. In the study, DNA was extracted from 125 frozen colorectal cancer samples obtained from the University of Barcelona Hospital Clinic in Barcelona, Spain. The arrays used in this study consisted of 2463 BAC clones that covered the human genome at a 1.5 mb resolution. After applying certain spot exclusion criteria, there were 2120 clones in the final dataset.

Following Eilers and Menezes (2004), we applied our method to samples X59, X524, X186 and X204 with a focus on chromosome 1. For a given quantile τ, we first computed the entire fused-QR solution path using our algorithm, then selected the tuning parameter using the SIC criterion. Figure 3 shows the ratio profiles for the four samples: the solid lines are the fitted median curves, and the dashed lines are upper/lower 15% quantile curves.

The interesting thing is that, when using our path algorithm to select the tuning parameter s and when using the actual physical locations of the clones, the results do not completely agree with Eilers and Menezes (2004). In Eilers and

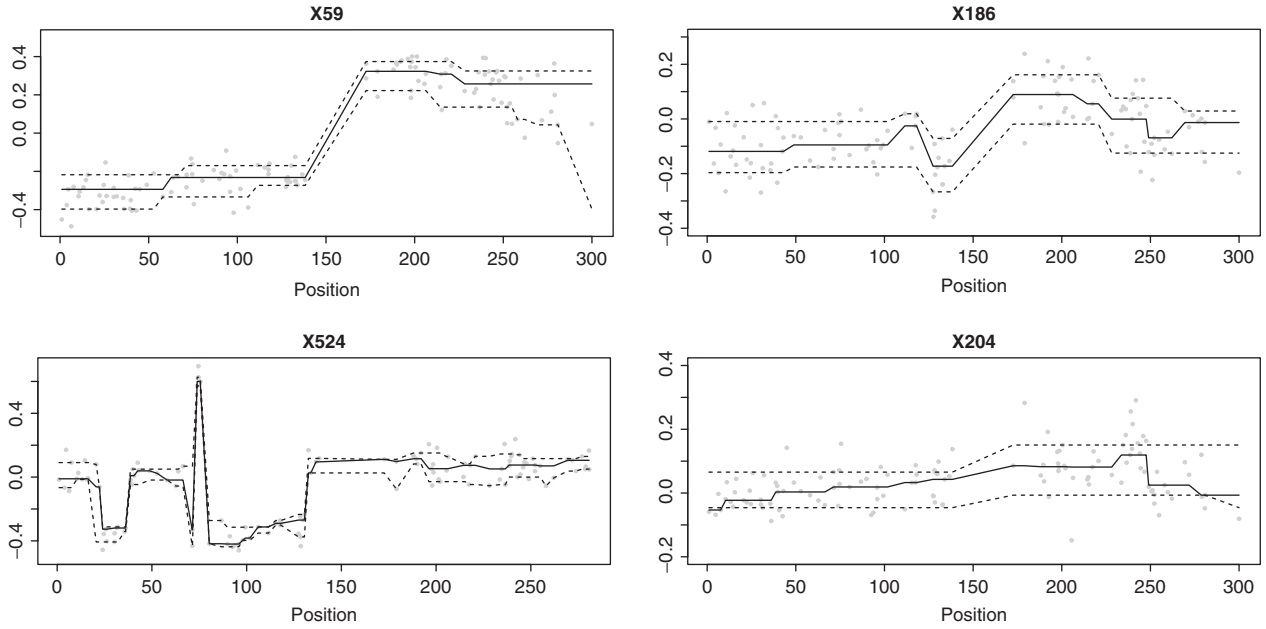

Fig. 3. Chromosome 1 from samples X59, X524, X186 and X204. Horizontal axis: physical position (mb) of the clones. Vertical axis: $\log_2$-ratio of intensities from the tumor versus reference samples. Solid lines represent the fitted median curves, and the dashed lines are fitted upper and lower 15% quantile curves.

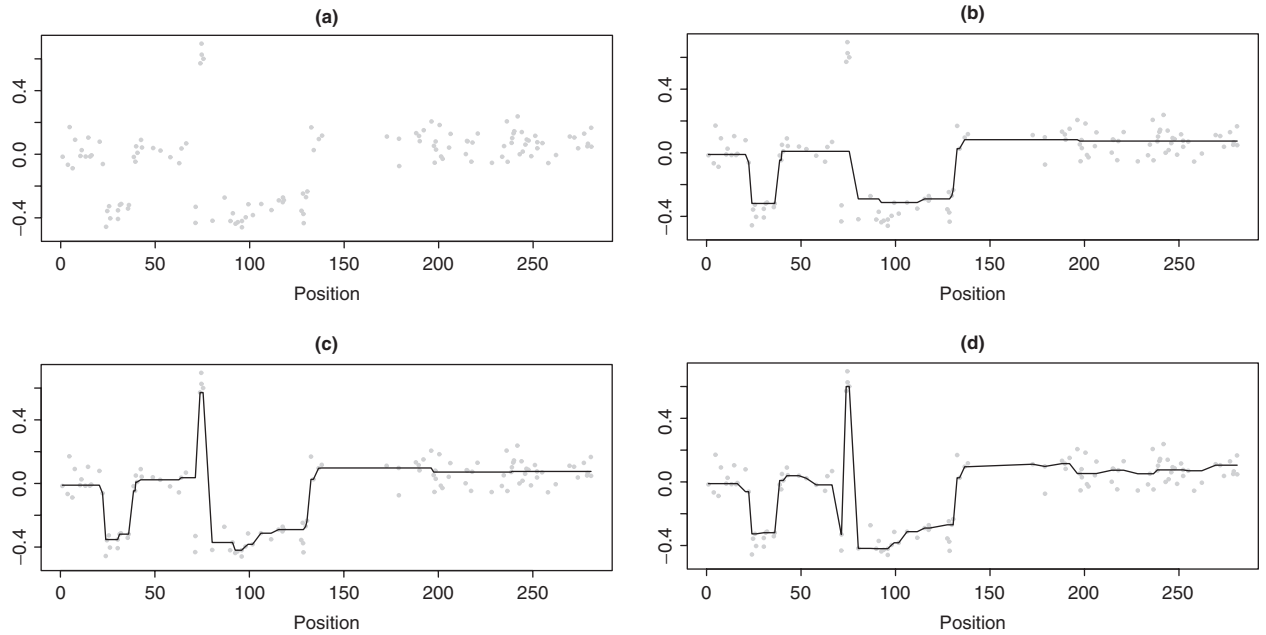

Fig. 4. Chromosome 1 from sample X524. (**a**) The scatterplot of observed ratios. (**b**) Results from Eilers and Menezes (2004). (**c**) Results from fused-QR using uniform spacing. (**d**) Results from fused-QR using the clones' positions. The solid lines are the fitted median curves.

Menezes (2004), uniform spacing between adjacent clones was assumed, and cross-validation was used to select the tuning parameter from a pre-specified grid of values. On the other hand, we utilized the physical locations of the clones in our model, and selected the tuning parameter from all possible values for s, as opposed to from a grid.

In particular, Figure 4 compares the results for sample X524 from three methods: (1) Eilers and Menezes (2004);

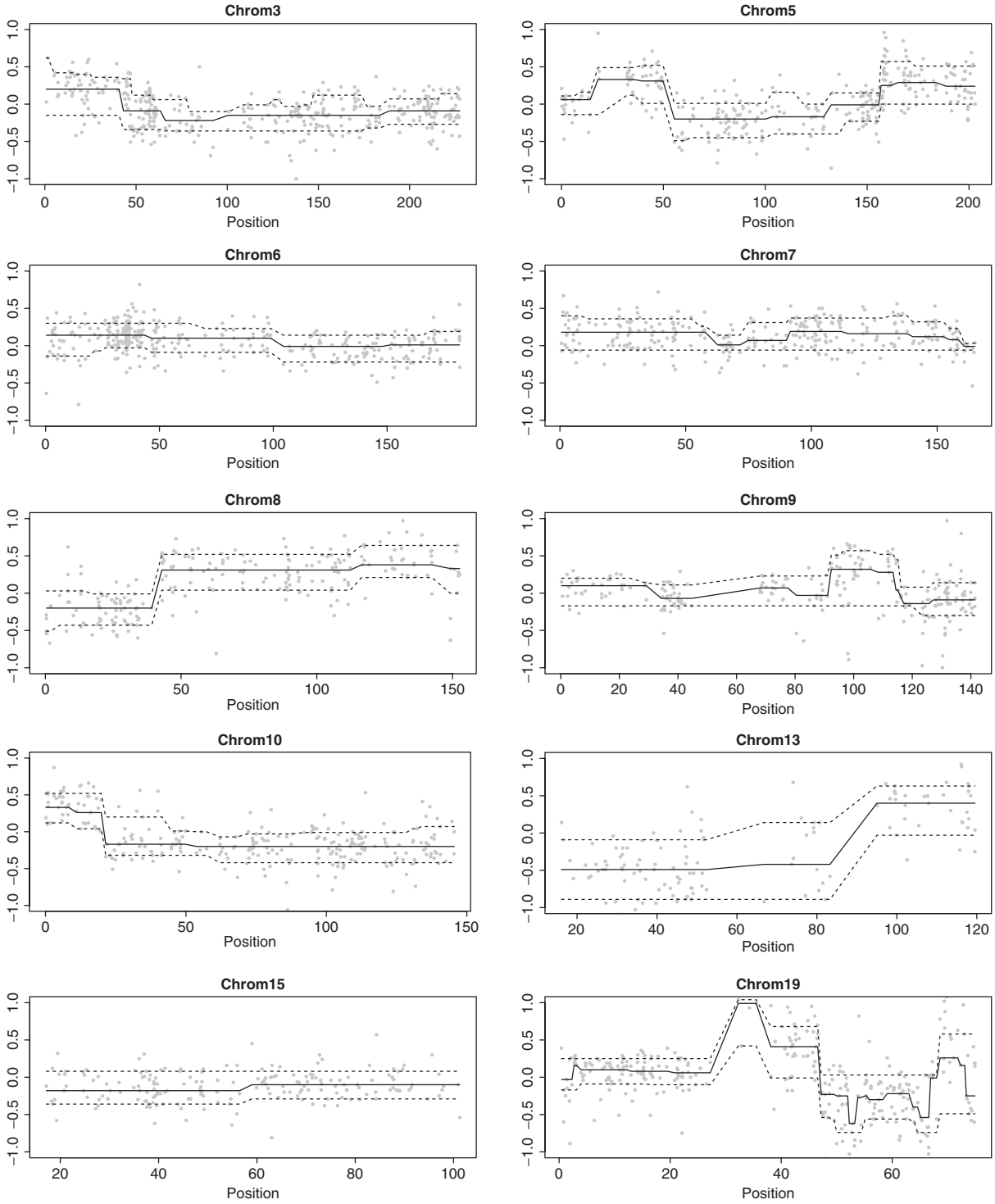

Fig. 5. Array-CGH profiles of 10 chromosomes of breast cancer cell line MDA157. The solid lines are fitted median curves and the dashed lines are fitted lower/upper 15% quantile curves.

(2) fused-QR using uniform spacing but the path algorithm for selecting s and (3) fused-QR using both the actual clone positions and the path algorithm. As we can see from the median curves, Eilers and Menezes (2004) missed a region of gain and a region of loss around the position of 70 Mb. Fused-QR using uniform spacing detected the region of gain but failed to detect the region of loss at this position, while fused-QR using actual physical locations captured both the gain and the loss.

3.3 Breast tumor data

Pollack *et al*. (2002) used cDNA array-based CGH to profile DNA copy number alterations in a series of breast cancer cell lines and primary tumors. Specifically, they performed CGH on 44 breast tumor samples and 10 breast cancer cell lines, using cDNA microarrays containing 6691 different mapped human genes. To demonstrate our fused-QR algorithm for identifying copy number aberrations, we focused on 10 chromosomes of the breast cancer cell line MDA157, which is known to have a large number of alterations. The fitted profiles are shown in Figure 5. When compared with the results from the CBS method (Olshen *et al*., 2004) and the CLAC method (Wang *et al*., 2005), as summarized in (Tibshirani and Wang 2007), our results seem promising, in terms of both resisting outlier measurements (chromosome 3, 5 and 9) and detecting weak alterations (chromosome 7 and 15).

4 CONCLUSION

We have proposed a fused quantile regression algorithm for smoothing array-CGH data. It is well known that quantile regression enjoys several attractive features. For example, quantile curves tend to present a more complete picture of the data than a single mean curve: the median curve shows the overall trend, while the upper and lower quantile curves show the spread. Quantile regression is also known to be not subject to the distributional assumption of the noise term in (8), while several other methods (Beheshti *et al*., 2003; Hsu *et al*., 2005; Olshen *et al*., 2004) rely on the constant variance assumption.

In our work, we have made three additional contributions:

- We developed an efficient algorithm that computes the entire solution path for the fused-QR model. This facilitates selection of the tuning parameter. For example, Eilers and Menezes (2004) pointed out that the choice of the tuning parameter is 'unlikely to be unique' in real applications, and it is worth to try different values of the tuning parameter to 'see whether strong patterns present themselves'. Our path algorithm allows biologists to play with the tuning parameter more easily and see how pattern changes with the value of the tuning parameter.
- We proposed a very convenient estimate for the complexity of the fitted model, i.e. the number of jumps in the fitted curve. This further allows convenient selection of the tuning parameter.
- We have incorporated the physical location of clones into the modeling procedure. Clone locations contain important information for understanding the aberrations. For example, if two regions indicating the same direction of aberrations are far apart, the chance that they refer to the same aberration should be lower than when they were closer. As we have seen in Section 3.2, using the physical location information can result in different fitted models in identifying copy number changes. However, we note that further complementary experiments are needed to test whether these differences are biologically meaningful.

However, we note that similar to most current algorithms, we did not address the issue of how to set a threshold to call for detection, i.e. we return only a fitted profile of the copy number changes, without calling the detected regions as significant or not. In practice, we may use the false discovery rate (FDR) to help use decide an appropriate threshold for the detection. Tibshirani and Wang (2007) proposed two ways to estimate the FDR for a given threshold, one when normal reference samples are available, and the other when normal reference samples are not available. We plan to explore along the same line in our future studies.

ACKNOWLEDGEMENTS

We would like to thank the Editor, the Associate Editor and three reviewers for their thoughtful and useful comments. Y.L. and J.Z. are partially supported by grant DMS-0505432 from the National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Beheshti,B. *et al*. (2003) Chromosomal localization of DNA amplifications in neuroblastoma tumors using cDNA microarray comparative genomic hybridization. *Neoplasia*, **5**, 53–62.

Eilers,P.H.C. and Menezes,R.X. (2004) Quantile smoothing of array CGH data. *Bioinformatics*, **21**, 1146–1153.

Efron,B. (1986) How biased is the apparent error rate of a prediction rule? *J. Am. Stat. Assoc*., **81**, 461–470.

Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal*., **90**, 132–153.

Hsu,L. *et al*. (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

Huang,T. *et al*. (2005) Detection of DNA copy number alterations using penalized least squares regression. *Bioinformatics*, **21**, 3811–3817.

Hupe,P. *et al*. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Biostatistics*, **20**, 3413–3422.

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Koenker,R. and Bassett,G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.

Koenker,R. *et al*. (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Lai,W.R. *et al*. (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

Lengauer,C. *et al*. (1998) Genetic instabilities in human cancers. *Nature*, **396**, 643–649.

Nakao,K. *et al*. (2004) High-resolution analysis of DNA copy number alterations in colorectal cancer by array-based comparative genomic hybridization. *Carcinogenesis*, **25**, 1345–1357.

Olshen,A.B. *et al*. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Pinkel,D. and Albertson,D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet*., **37**, s11–s17.

Pollack,J.R. *et al*. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci*., **99**, 12963–12968.

Polzehl,J. and Spokoiny,S. (2000) Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B*, **62**, 335–354.

Schwarz,G. (1978) Estimating the dimension of a model. *Annu. Stat*., **3**, 98–108.

Snijders,A.M. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet*., **29**, 263–264.

Stein,C. (1981) Estimation of the mean of a multivariate normal distribution. *Annu. Stat*., **9**, 1135–1151.

Tibshirani,R. and Wang,P. (2007) Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 1–12.

Tibshirani,R. *et al*. (2005) Sparsity and smoothness via the fused Lasso. *J. R. Stat. Soc. Ser. B*, **67**, 91–108.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.