Biostatistics (2007), **0**, 0, *pp.* 1–12
doi:10.1093/biostatistics/kxm013

Spatial smoothing and hot spot detection for CGH data using the fused lasso

ROBERT TIBSHIRANI*

Deptartments of Health, Research & Policy, and Statistics, Stanford University
Stanford, CA 94305, USA
tibs@stat.stanford.edu

PEI WANG

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500,
PO Box 19024, Seattle, WA 98109, USA

SUMMARY

We apply the "fused lasso" regression method of Tibshirani *and others* (2004) to the problem of "hot-spot detection", in particular, detection of regions of gain or loss in comparative genomic hybridization (CGH) data. The fused lasso criterion leads to a convex optimization problem, and we provide a fast algorithm for its solution. Estimates of false-discovery rate are also provided. Our studies show that the new method generally outperforms competing methods for calling gains and losses in CGH data.

Keywords: DNA copy number; Signal detection.

1. INTRODUCTION

In this paper, we apply the fused lasso method (Tibshirani *and others*, 2004) to the "hot-spot" detection problem in comparative genomic hybridization (CGH) data. The CGH signal is approximated by a piecewise function that has relatively sparse areas with nonzero values. Hence, the method is useful for determining which areas of the signal are likely to be nonzero.

CGH is a technique for measuring DNA copy numbers of selected genes on the genome. In cells with cancer, mutations can cause a gene to be either deleted from the chromosome or amplified, that is, there are extra DNA copies of the gene. The CGH array experiments return the $\log_2$ ratio between the number of DNA copies of the gene in the tumor cells and the number of DNA copies in the reference cells. A value greater than zero indicates a possible gain in DNA copies of that gene, while a value less than zero suggests possible losses.

The results of a CGH experiment are often interpreted by a biologist, but this is time consuming and not necessarily very accurate. In recent years, a number of algorithms have been developed for automatic interpretation, including Fridlyand *and others* (2004), Olshen & Venkatraman (2004), Myers *and others* (2004), Wang *and others* (2005), Lipson *and others* (2005), and Picard *and others* (2005); a

*To whom correspondence should be addressed.

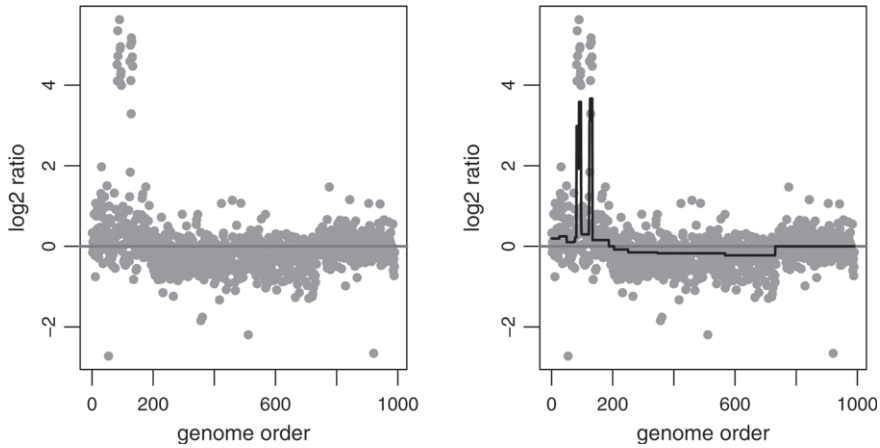

Fig. 1. Fused lasso applied to some GBM data. The data are shown in the left panel, and the solid line in the right panel represents the inferred copy number $\hat{\beta}$ from the fused lasso. The grey line is for $y = 0$.

comprehensive comparison is given by Lai *and others* (2005). Approaches include successive top–down splitting, bottom–up agglomeration along the chromosome, and hidden Markov models.

The left panel of Figure 1 shows an example. The data represent CGH measurements from 2 glioblastoma multiforme (GBM) tumors (see Section 4.2 for details).

In this paper, we apply the fused lasso method of Tibshirani *and others* (2004) to spatial smoothing and the CGH detection problem. The solid line in the right panel of Figure 1 shows the result of the fused lasso method applied to these data. The method has successfully detected the narrow regions of gain and the wide regions of loss.

2. THE FUSED LASSO AND THE PROPOSED METHOD FOR HOT-SPOT DETECTION

The "fused lasso" (Tibshirani *and others*, 2004) is a generalization of the lasso (Tibshirani, 1996) and is defined by

$$\hat{\beta} = \operatorname{argmin} \sum_i \left(y_i - \sum_j x_{ij}\beta_j \right)^2$$
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leqslant s_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leqslant s_2. \tag{2.1}$$

The fused lasso is developed for the situations when $x_{i1}, x_{i2} \ldots, x_{ip}$ have some kind of natural ordering. The additional fused constraint encourages the flatness of the coefficient profiles β_j as a function of j.

Here, we apply the fused lasso in the special case in which $\{x_{ij}\}$ is the identity matrix. Hence, we are smoothing the sequence $y_1, y_2, \ldots, y_n$ along the 1-dimensional index i.

Denote the $\log_2$ ratio measurement of a chromosome (or chromosome arm) as $Y = \{y_i\}_{i=1}^n$; we are interested in finding coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$ satisfying

$$\hat{\beta} = \operatorname{argmin} \left\{ \sum_i (y_i - \beta_i)^2 \right\} \text{ subject to } \sum_j |\beta_j| \leqslant s_1, \sum_j |\beta_j - \beta_{j+1}| \leqslant s_2, \tag{2.2}$$

and $\hat{\beta}_j$ is the inferred DNA copy number for gene j. Here, s_1 controls the overall DNA copy number alteration amount of the target chromosome (or chromosome arm), while s_2 controls the frequency of the alterations in the target region.

We investigated 2 different methods for estimating s_1, s_2 and calling gains and losses.

2.1 *Method 1*

The first method estimates s_1 and s_2 from pre-smoothed version of the data, applies the fused lasso with these values, and then thresholds the estimate to determine regions of gains or losses.

Since the overall DNA copy number alteration amount is contributed mainly by the large gain/loss regions, we derive the value of s_1 by using a heavily smoothed version of Y:

1) Apply "lowess" to Y with fraction parameter $f = \max(1/50, 50/p)$, where p is the length of Y. (The lowess window will be at least 50.)
2) Denote the lowess result by $\tilde{Y}$. Define $\hat{s}_1 = \sum_i |\tilde{y}_i|$.

On the other hand, s_2 controls the frequency of the alterations in the target region. Thus, we use moderately smoothed Y to infer s_2:

1) Apply lowess to Y with fraction parameter $f = \min(1/20, 10/p)$. (The lowess window will be at most 10.)
2) Denote the result as $\breve{Y}$. Let $d_i = \breve{y}_i - \breve{y}_{i+1}$ and calculate the median absolute deviation of $\{d_i\}$:

$$\delta = \text{median}\left(\{|d_i - \mu_d|\}_i\right),$$

   where $\mu_d = \text{median}(\{d_i\}_i)$.
3) It is reasonable to assume that d_i with absolute values greater than 4δ corresponds to copy number alteration break points on the genome. Thus, we define $\hat{s}_2 = 2\delta + \sum_i |d_i| \cdot I\,(|d_i| > 4\delta)$.

After we obtain $\hat{s}_1$ and $\hat{s}_2$, we apply the fused lasso to estimate the underlying signals $\hat{\beta}_i, i = 1, 2, \ldots, n$. Finally, we threshold $|\hat{\beta}_i|$ by a value θ to obtain the final regions of gain or loss. The choice of θ is discussed in Section 2.3, where it is used to control the false-discovery rate (FDR).

2.2 *Method 2*

Here, we vary both s_1 and s_2 in a systematic way. Define $S_1(X) = \sum_i |x_i|$. We calculate $s_1 = S_1(\text{lowess}(Y, f = w/p))$ and $s_2 = 2s_1/w$, where w represents the average length of alteration regions. The value of w is varied and then the fused lasso is computed for the resulting values of s_1 and s_2. Figure 2 shows an example with some simulated data from a normal and tumor array. As w increases, the estimate does a better job of isolating the gain regions in the tumor sample. However, there is still a large region of low but nonzero values that would require further thresholding before gains and losses can be called. Hence, this approach seems too complicated, and we therefore settle on method 1.

2.3 *Estimation of FDR*

For 1 array, when we are trying to decide whether the ith gene/clone has significant DNA copy number alteration, we are actually doing a hypothesis testing with

H_0: Gene/clone i does not belong to any gain/loss region.

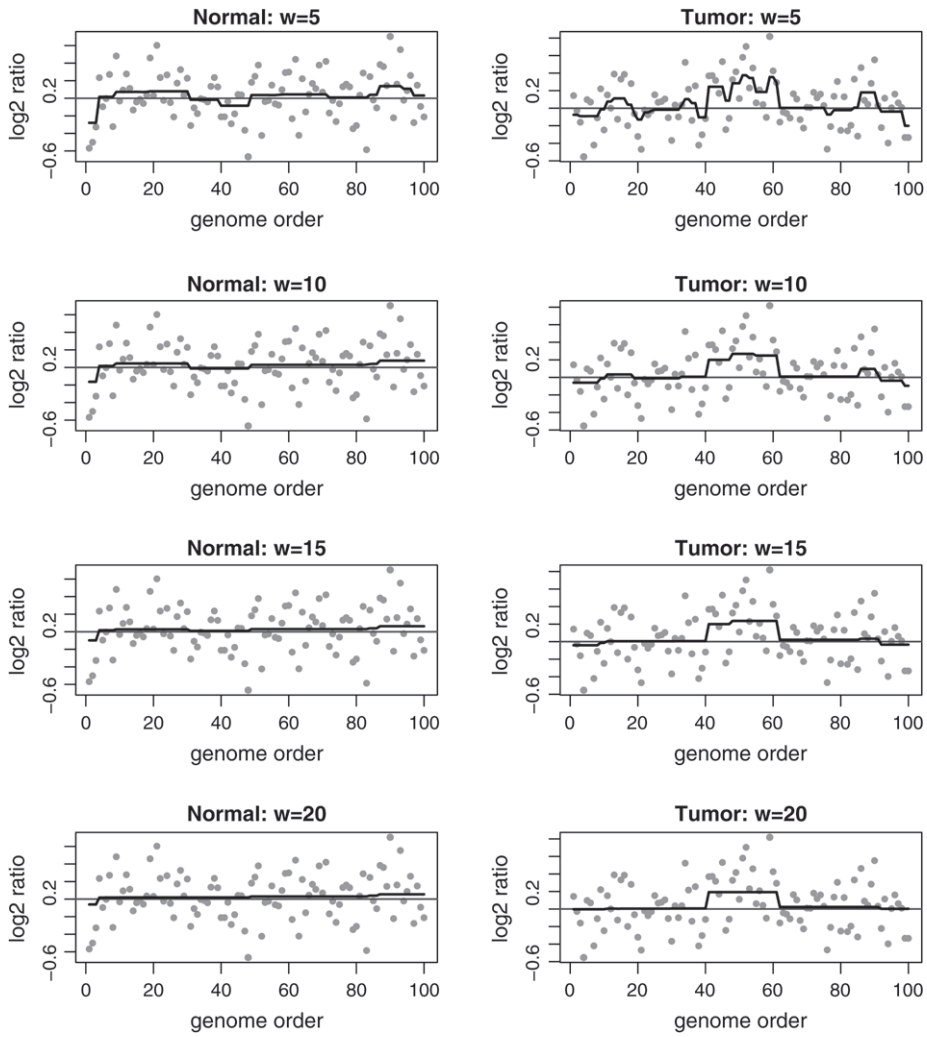

Fig. 2. Simulated data: the effect of parameter ω (the average length of alterations) on the fused lasso solution. The panels of the left column represent a chromosome with no copy number alteration; the panels of the right column represent a chromosome with an amplification region of genes 40–60. The solid lines illustrate the estimated copy numbers of fused lasso; a horizontal line is drawn at $y = 0$.

Here, rises the issue of multiple hypothesis testing, for, in each array, tens of thousands of genes/clones need to be considered simultaneously.

Although we do not have independent H_i for each gene in our problem, we can still use

$$\widehat{\text{FDR}} = \frac{\text{number of genes picked under the null distribution}}{\text{number of genes picked in the observed data}} \tag{2.3}$$

as a rough estimator for (FDR) (Benjamini & Hochberg, 1995; Chu *and others*, 2002; Storey, 2002; Efron & Tibshirani, 2002). We assume that the denominator is greater than 0 with probability 1 (i.e. the event that no significant genes are selected is rare).

We estimate the FDR in 2 different ways according to the availability of normal reference arrays.

Estimation of FDR when normal reference arrays are available. Suppose there are M normal reference arrays $\{R^m\}_{m=1}^M$, where $R^m = \{r_i^m\}_{i=1}^n$. We first scale the normal reference arrays according to the target tumor array and then use the copy number inferred from the reference arrays to approximate the null distribution of $\{\hat{\beta}_i\}_i$.

1) The contiguous genes/clones with the same $\hat{\beta}_i$ are defined as 1 segment. Suppose there are K segments resulting from $\{\hat{\beta}_i\}$ and denote them as $S_k = \{i_{k-1}+1, \ldots, i_k\}$. We then calculate the within-segment standard deviation of the target array

$$\hat{\sigma} = \frac{1}{n} \sum_k \sum_{i \in S_k} (y_i - \mu_k)^2,$$

where $\mu_k = \text{mean}(y_i : i \in S_k)$.
2) Normalize each reference array according to $\hat{\sigma}$:

$$\widetilde{R^m} = \hat{\sigma} \cdot R^m / sd(R^m).$$

3) Estimate s_1 and s_2 for each $\widetilde{R^m}$ and then apply fused lasso on it. Denote the resulting coefficients by $\{\widetilde{\beta_i^m}\}_i$.
4) For a given threshold θ, FDR of the procedure is estimated as a function of θ:

$$\widehat{\text{FDR}}(\theta) = \frac{\frac{1}{M} \sum_m \sum_i I(\widetilde{\beta_i^m} > \theta)}{\sum_i I(\hat{\beta}_i > \theta)}. \tag{2.4}$$

Usually, the data analyst pre-chooses a target FDR and vary θ over a range to seek the solution with $\widehat{\text{FDR}}(\theta)$ closest to the target.

Estimation of FDR when normal reference arrays are not available. In real experiments, there are often situations where appropriate normal reference arrays are not available due to sample/lab limitations. Thus, there is a need to control the FDR in the absence of reference arrays.

We approach this by considering segments as the hypothesis unit first. For 1 segment S_k, the hypothesis of interest can be stated as

$$H_0^k: \mu_k, \text{ the mean of } \{y_i\}_{i \in S_k}, \text{ is equal to } 0.$$

This makes it natural to examine statistics $z_k = \frac{\sum_{i \in S_k} y_i}{\sqrt{n_k}\hat{\sigma}}$, where n_k is the size of S_k. To take advantage of the inferred copy number $\{\widehat{\beta_i}\}_i$, we further define $\widehat{z_k} = \frac{\sum_{i \in S_k} \widehat{\beta_i}}{\sqrt{n_k}\hat{\sigma}}$.

We approximate the null distribution of $\widehat{z_k}$ with standard normal and define $p_k = P(Z > |\widehat{z_k}|)$, where $Z \sim N(0, 1)$.

If we assume $\widehat{z_k}$ are independent from each other, then, for a given $q \in (0, 1)$, a conservative estimator of the genome-wide FDR in (2.3) is

$$\widehat{\text{FDR}}(q) = \frac{Kq \cdot \frac{1}{K} \sum_k n_k}{\sum_k n_k \cdot I(p_k \leqslant q)} = \frac{n \cdot q}{\sum_k n_k \cdot I(p_k \leqslant q)}. \tag{2.5}$$

Here, genes/clones in the segments with $p_k \leqslant q$ are considered to have copy number alterations.

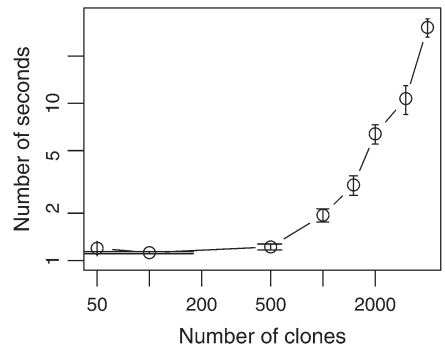

Fig. 3. Number of CPU seconds required for the fused lasso algorithm as a function of the number of clones.

Again, data analyst can vary q over a range and choose the solution with $\widehat{\mathrm{FDR}}(q)$ closest to the preselected value.

2.4 *Computational considerations*

The optimization problem for the fused lasso is a quadratic programming problem. Criterion (2.1) leads to a quadratic programming problem. For large n, the problem is difficult to solve and special care must be taken to avoid the use of p^2 storage elements. We use the approach for the general fused lasso problem outlined in Tibshirani *and others* (2004): the 2-phase active set algorithm "sqopt" of Gill *and others* (1999), which is designed for quadratic programming problems with sparse linear constraints.

Figure 3 shows a log-log plot of the number of seconds required for the computation as a function of the number of clones (genes), on a Linux computer. We simulated 10 data sets of each size, and the figure shows the mean plus or minus standard error; the computation took on average about 30 s for 4000 clones. Beyond $n = 1000$ clones, the plot is roughly linear with a slope of about 2. This suggests that the computation is of order n^2. The estimated computation time for a data set with 100 000 clones is about 135 min.

While this may be sufficiently fast for practical use, clearly it would be preferable to reduce the computation time for large data sets. We are currently working with Jerome Friedman and Trevor Hastie on specialized algorithms for the fused lasso, and the initial results show considerable promise.

3. COMPARISON TO OTHER SMOOTHING METHODS

General smoothing methods are not typically used for analyzing CGH data because their results can be difficult to interpret (Lai *and others*, 2005). This is illustrated in Figure 5, where 2 popular smoothing methods—lowess (Becker *and others*, 1988) and penalized smoothing splines (Ruppert *and others*, 2003)—are applied to the example data (see Section 4.2 for details). We used R function lowess (smooth window $= 10$) and "spm" (default parameters) to compute the results. As we can see from the figure, the 2 smoothing methods do not provide direct calls for copy number gains/losses, and thus require additional thresholding for identifying regions with significant alterations. In addition, the smoothing curves do not catch the piecewise constant shape of copy number changes, which raises additional challenges for controlling the FDR. Moreover, copy number alterations can be both large chromosome segmentation gain/loss and also abrupt local amplification/deletion. Therefore, different degrees of smoothness are needed for different chromosome regions, which adds another layer of complexity to the kernel- and spline-based approaches.

In the fused lasso, the fused term in the loss function can be viewed as the first derivative of the coefficient profiles, and thus the method can also be deemed as a "smoothing" approach. However, the use of L_1-norm on the fused penalty term enables the method to capture both the piecewise flatness patterns and the abrupt local jumps at the same time (see Figure 5). In addition, the control on the overall sparsity of the coefficient solution helps to screen away the "cold"-spot regions. These make the fused lasso a more attractive approach for analyzing CGH data.

4. Some results

4.1 *Simulated data*

We apply the proposed method on artificial chromosomes simulated by Lai *and others* (2005) (downloaded from http://www.chip.org/∼ppark/Supplements/Bioinformatics05b.html). We consider the most challenging situation where the signal-to-noise ratio is equal to 1. For each of the 4 different aberration widths (5, 10, 20, and 40 probes), there are 100 independently simulated chromosomes with 100 probes in total.

We first compare the inferred copy number $\{\hat{\beta}_i\}_i$ by fused lasso with the estimated copy numbers from 3 other popular programs "CGHseg" (Picard *and others*, 2005), "CBS" (Olshen & Venkatraman, 2004), and "CLAC" (Wang *and others*, 2005). The receiver operating curves of different methods shown in Figure 4 suggest that fused lasso better captures the true DNA copy number alterations than the other 3 methods, especially when the aberration width is small.

We then investigate the 2 FDR estimators proposed in Section 2.3. $\widehat{\mathrm{FDR}}(\theta)$ is computed based on 10 simulated reference arrays with no copy number alterations (i.i.d. from $N(0, 1)$). We then control $\widehat{\mathrm{FDR}}(\theta)$ and $\widehat{\mathrm{FDR}}(q)$ at the level of 0.01 and calculate the true FDR for the selected solutions. The result is summarized in Table 1. Overall, when we control $\widehat{\mathrm{FDR}}(\theta) = 0.01$, 80.75% simulated cases have true FDR less than 0.01; when we control $\widehat{\mathrm{FDR}}(q) = 0.01$, 84.5% simulated cases have true FDR less than 0.01. Generally, $\widehat{\mathrm{FDR}}(q)$ is more conservative than $\widehat{\mathrm{FDR}}(\theta)$, for the former is derived on the assumption that segments are independent from each other.

4.2 *GBM data*

The glioma data from Bredel *and others* (2005) contain samples representing primary GBMs, a particular malignant type of brain tumor. We investigate the performance of various methods on the array CGH profiles of the GBM samples examined in Lai *and others* (2005). To generate a more challenging situation where both local amplification and large region loss exit in the same chromosome, we paste together the following 2 array regions: (1) chromosome 7 in GBM29 from 40 to 65 Mb and (2) chromosome 13 in GBM31. The performance of different methods on this pseudo chromosome is illustrated in Figure 5. We can see that the proposed method using fused lasso successfully identified both the local amplification and the big chunk of copy number loss.

4.3 *Breast tumor data*

In the study conducted by Pollack *and others* (2002), cDNA microarray CGH was profiled across 6691 mapped human genes in 44 breast tumor samples and 10 breast cancer cell lines. The scanned raw data were downloaded from the Stanford Microarray Database (http://smd.stanford.edu). We pick the breast cancer cell line (MDA157) as an example, which has a large degree of copy number alterations, and apply our proposed method as well as the other 3 methods on it to estimate the underlying copy number changes. From the results shown in Figure 6, we can see that fused lasso successfully recognized varies copy number alterations.

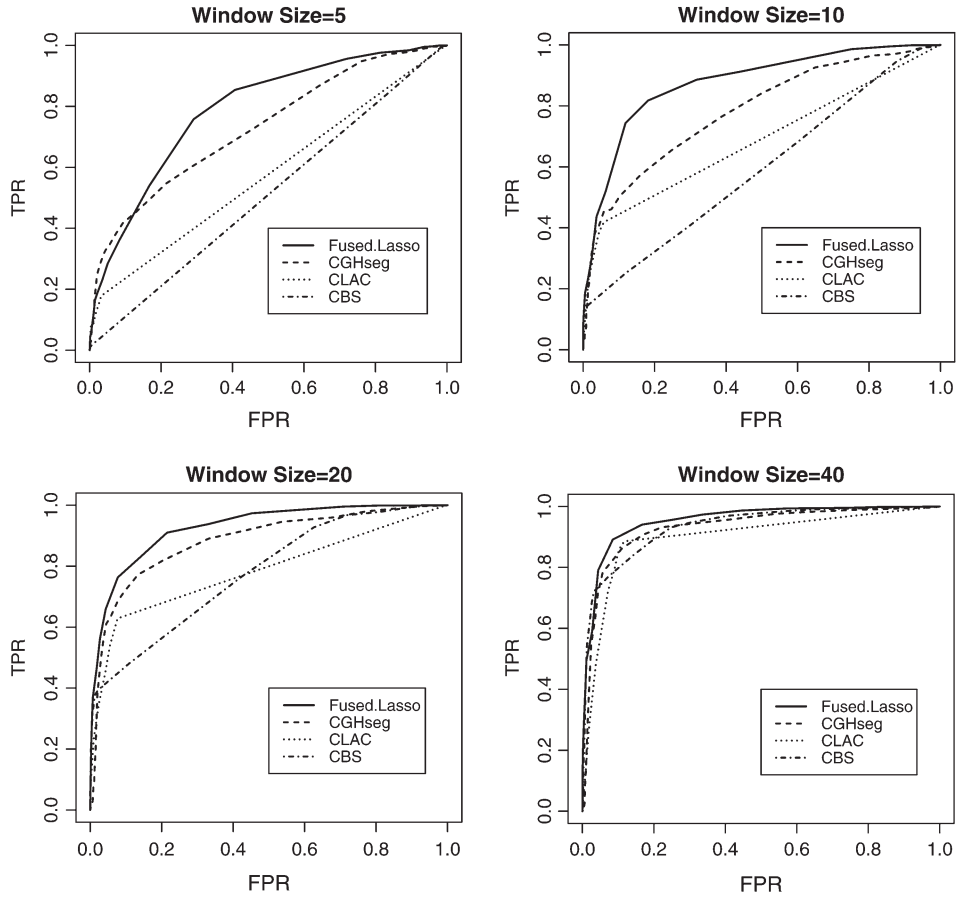

Fig. 4. TPR = (the number of probes within the aberration width that is above a threshold)/(the total number of probes within the aberration width); FPR = (the number of probes outside the aberration width that is above a threshold)/(the total number of probes outside the aberration width).

Table 1. *The median,* 75% *quantile, and* 90% *quantile of the true FDR values*

Window size	$\widehat{\text{FDR}}(\theta) = 0.01$			$\widehat{\text{FDR}}(q) = 0.01$		
	Median	75%	90%	Median	75%	90%
5	0	0	0.499	0	0	0
10	0	0	0	0	0	0
20	0	0	0.108	0	0	0.09
40	0	0.037	0.133	0	0.03	0.29

CGHseg appears to be very sensitive to outlier measurements and thus will be more suitable for detecting single-gene copy number mutations of high-quality arrays. CLAC is conservative in handling outliers with opposite signs in 1 alteration region and therefore tends to break large alteration segments into small blocks. CBS provides clean solutions for segmentations but has the limitation to detect break points whose alteration signals are weak (e.g. chromosome 7 and 15 of the selected cell line).

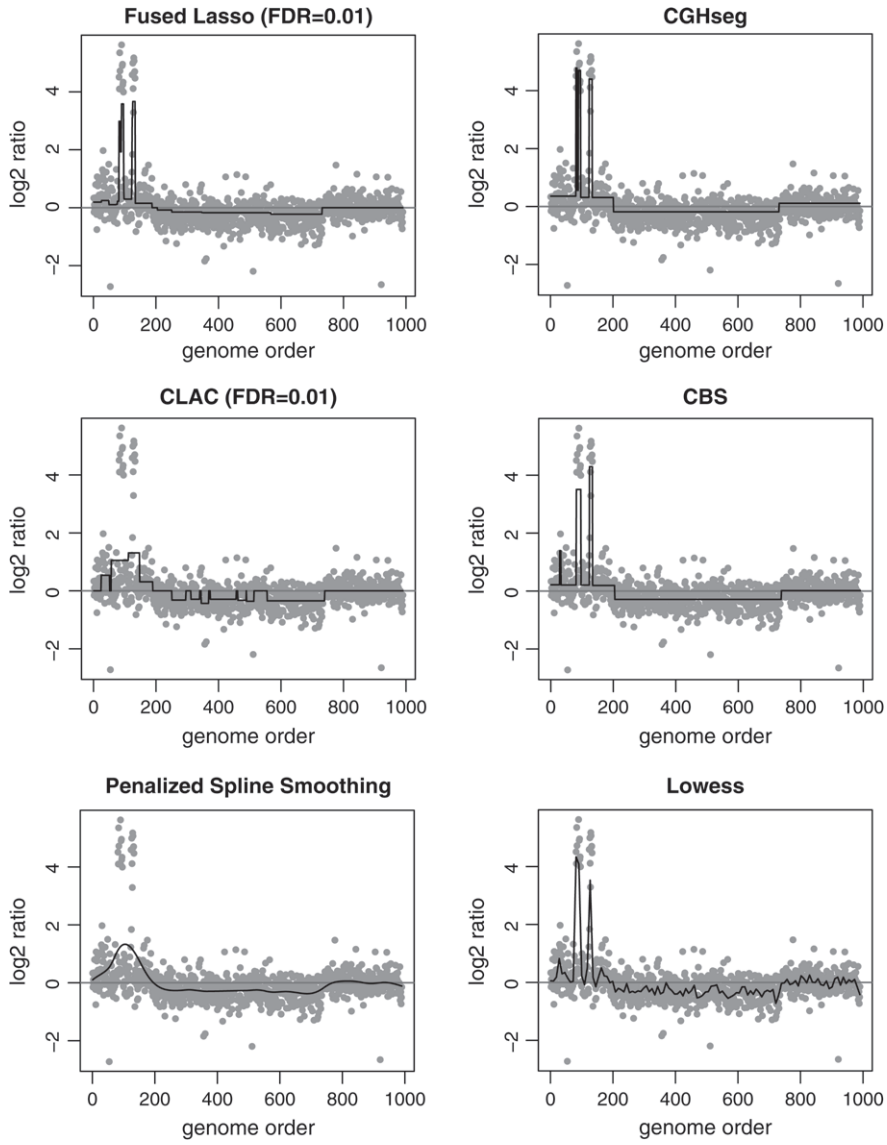

Fig. 5. Chromosome 7 and chromosome 13 from 2 GBM tumors.

5. Discussion and future work

The fused lasso can be generalized to other analysis besides the calling of gains and losses in CGH data. For example, biologists have great interest in understanding the interactions between copy number alterations and mRNA expression levels. To study this, a commonly used method is to identify gene pairs, of which one gene's copy number and the other gene's expression level are significantly correlated. However, genome-wide screening for such pairs is quite challenging: (1) the large dimensionality of the problem raises the requirement of controlling the sparsity in the solutions; (2) the strong spacial correlations of gene copy number changes need to be taken into account. The fused lasso can be used again to tackle these challenges.

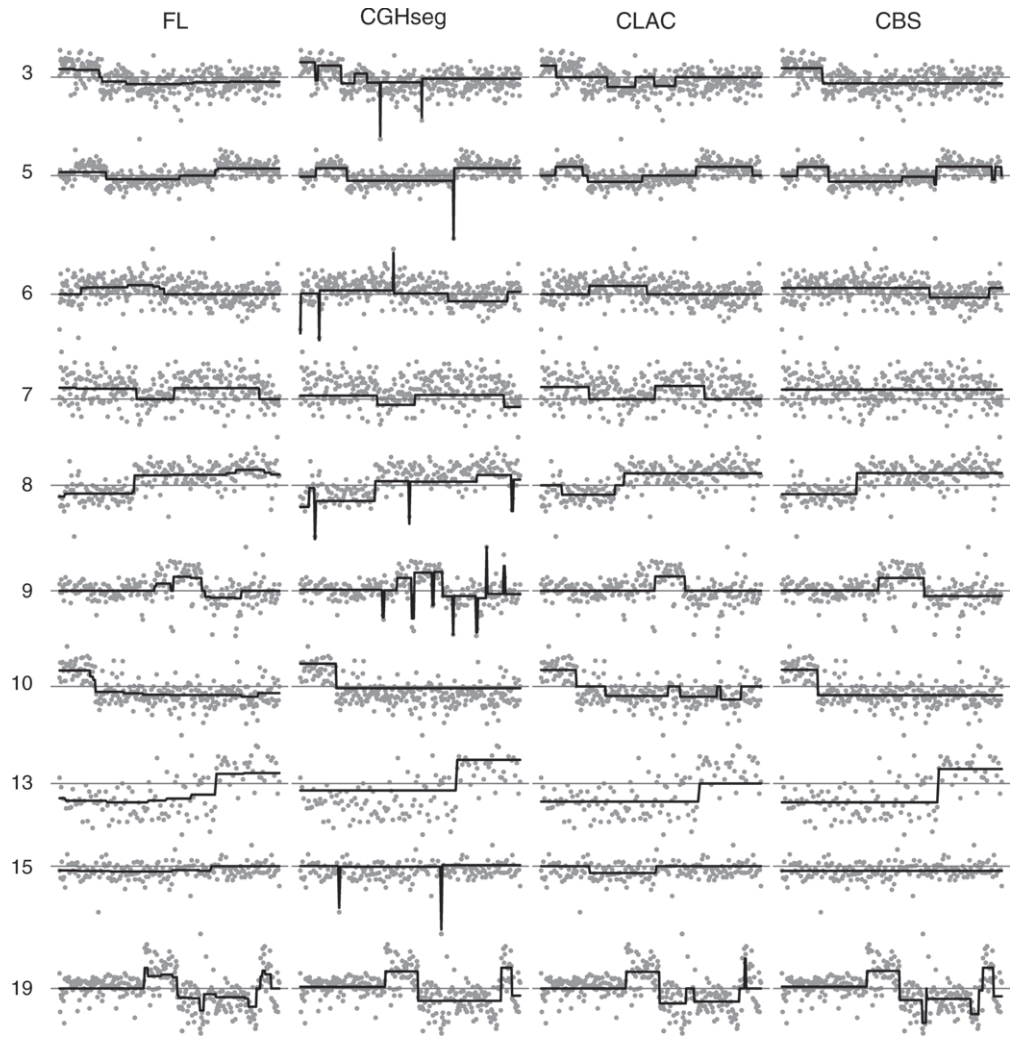

Fig. 6. Array CGH profile of 10 chromosomes of breast cancer cell line MDA157. Panels in the same row are for the same chromosome. The integers at the beginning of each row are chromosome indexes. The dark black line in each panel represents the estimated copy number of a particular method, whose name is shown on the top of each column. The black horizontal line in each panel represents $y = 0$.

Denote the expression measurements of a target gene as $Y = (y_1, y_2, \ldots, y_J)$, where J represents the total number of samples. Denote the CGH measurement of the ith gene across the J samples as $X^i = (x_1^i, \ldots, x_J^i)$. Suppose Y and $\{X^i\}$ have all been normalized to mean 0 and variance 1. $\operatorname{cor}(Y, X^i)$ is then exactly the least-squares coefficient of the linear model $Y \sim \beta X^i$. Thus, we can infer the correlation coefficients by solving

$$\{\hat{\beta}\} = \operatorname{argmin}\left\{\sum_i \|Y - \beta_i X^i\|^2\right\} \text{ subject to } \sum_j |\beta_j| \leqslant s_1, \sum_j |\beta_j - \beta_{j+1}| \leqslant s_2, \tag{5.1}$$

where a nonzero β_j suggests that Y and X^i are significantly correlated.

In addition, for CGH data, our paper focuses on the calling of gains and losses from a single array. Typically, the researcher collects data on a few dozen arrays, often divided into normal and diseased groups. Ideally, one would like to carry out a joint analysis of the data set, finding regions of gains and losses that occur commonly among the patients and also ones that occur differentially in the normal and diseased groups. One interesting proposal along these lines is the STAC procedure of Diskin *and others* (2006). This method projects the calls of gains and losses for individual arrays along chromosome, measuring the frequency and "footprint" of calls in each local region. The STAC procedure could be applied to the gains and losses called by the fused lasso. However, a more integrated approach might be possible. For example, given data on arrays $j = 1, 2, \ldots, J$, one could first apply the fused lasso to the data from each array. Suppose that we parse the resulting coefficient vectors into a collection of nonzero basis functions $\hat{\beta}_{ik}$, $k = 1, 2, \ldots, K$. Each function is nonzero over a single interval of the real line. Then, we can do some kind of joint regression of the data from all arrays on this collection, allowing the arrays to share the basis functions as efficiently as possible. This would reveal basis functions (areas of the chromosome) that are common among many or all arrays and also those that are unique to only a few. This is a topic for further study.

An R language package for the fused lasso will be freely available at http://www.stat.stanford.edu/~tibs/cghFLasso.

Acknowledgments

We would like to thank Michael Saunders for his expert advice on computational issues and would like to thank the referees and editors for helpful comments that led to improvements in this manuscript. Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183. *Conflict of Interest:* None declared.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language*. Pacific Grove, CA: Wadsworth Brooks Cole.

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* **85**, 289–300.

Bredel, M., Bredel, C., Juric, D., Harsh, G. R., Vogel, H., Recht, L. D. and Sikic, B. I. (2005). High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Research* **65**, 4088–4096.

Chu, G., Narasimhan, B., Tibshirani, R. and Tusher, V. (2002). *Significance Analysis of Microarrays (SAM) Software.* http://www-stat.stanford.edu/~tibs/SAM/. Accessed July 16, 2003.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, Jr, C. J., Weber, B. L., Maris, J. M. and Grant, G. R. (2006). Stac: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Research* **16**, 1149–1158.

Efron, B. and Tibshirani, R. (2002). Microarrays, empirical Bayes methods, and false discovery rates. *Genetic Epidemiology* **1**, 70–86.

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

Gill, P., Murray, W. and Saunders, M. (1999). Users guide for SQOPT 5.3: a Fortran package for large-scale linear and quadratic programming. *Technical Report*. Palo Alto, CA: Stanford University.

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. *Biostatistics* **21**, 3763–3770.

LIPSON, D., AUMANN, Y., BEN-DOR, A., LINIAL, N. AND YAKHINI, Z. (2005). Efficient calculation of interval scores for DNA copy number data analysis. *Journal of Computational Biology* **13**, 215–228.

MYERS, C. L., DUNHAM, M. J., KUNG, S. Y. AND TROYANSKAYA, O. G. (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics* **20**, 3533–3543.

OLSHEN, A. AND VENKATRAMAN, E. (2004). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Biostatistics* **5**, 557–572.

PICARD, F., ROBIN, S., LAVIELLE, M., VAISSE, C. AND DAUDIN, J. (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics* **11**, 6–27.

POLLACK, J. R., SÓRLIE, T., PEROU, C. M., REES, C. A., JEFFREY, S. S., LONNING, P. E., TIBSHIRANI, R., BOTSTEIN, D., BÓRRESEN-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 USA* **99**, 12963–12968.

RUPPERT, D., WAND, M. P. AND CARROLL, R. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

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

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

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.

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.

[*Received November 28, 2006; revision March 15, 2007; accepted for publication March 21, 2007*]