Biostatistics (2011), **0**, 0, *pp.* 1–16
doi:10.1093/biostatistics/kxr012

C

A fused lasso latent feature model for analyzing multi-sample aCGH data

GEN NOWAK*
Department of Biostatistics, Harvard University, Boston, MA 02115, USA
gen.nowak@gmail.com

TREVOR HASTIE
Departments of Statistics and Health Research and Policy, Stanford University, Stanford, CA 94305, USA

JONATHAN R. POLLACK
Department of Pathology, Stanford University, Stanford, CA 94305, USA

ROBERT TIBSHIRANI
Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305, USA

SUMMARY

Array-based comparative genomic hybridization (aCGH) enables the measurement of DNA copy number across thousands of locations in a genome. The main goals of analyzing aCGH data are to identify the regions of copy number variation (CNV) and to quantify the amount of CNV. Although there are many methods for analyzing single-sample aCGH data, the analysis of multi-sample aCGH data is a relatively new area of research. Further, many of the current approaches for analyzing multi-sample aCGH data do not appropriately utilize the additional information present in the multiple samples. We propose a procedure called the Fused Lasso Latent Feature Model (FLLat) that provides a statistical framework for modeling multi-sample aCGH data and identifying regions of CNV. The procedure involves modeling each sample of aCGH data as a weighted sum of a fixed number of features. Regions of CNV are then identified through an application of the fused lasso penalty to each feature. Some simulation analyses show that FLLat outperforms single-sample methods when the simulated samples share common information. We also propose a method for estimating the false discovery rate. An analysis of an aCGH data set obtained from human breast tumors, focusing on chromosomes 8 and 17, shows that FLLat and Significance Testing of Aberrant Copy number (an alternative, existing approach) identify similar regions of CNV that are consistent with previous findings. However, through the estimated features and their corresponding weights, FLLat is further able to discern specific relationships between the samples, for example, identifying 3 distinct groups of samples based on their patterns of CNV for chromosome 17.

Keywords: Cancer; DNA copy number; False discovery rate; Mutation.

*To whom correspondence should be addressed.

1. INTRODUCTION

Diploid organisms have 2 homologous copies of each chromosome and as such, any segment of DNA in an autosome normally exists in 2 copies. The number of copies of a segment of DNA in the genome is referred to as the DNA copy number for that segment. Changes in genomic DNA copy number can occur when regions of DNA are either lost or gained, resulting in copy numbers that differ from the normal value of 2. Although DNA copy number variations (CNVs) are observed in normal individuals, some CNVs are known to be associated with certain diseases, including cancer (where if they occur as somatic changes, they are more often referred to as copy number alterations). Identifying the regions of CNV and determining whether the regions have gained or lost copy number can give insight into the genetics of these diseases. For example, a region of CNV may encompass genes that contribute to the development and progression of cancer.

Array-based comparative genomic hybridization (aCGH) is a high-throughput, high-resolution method for measuring changes in DNA copy number at thousands of locations in a genome. Further details regarding aCGH are given in Pinkel *and others* (1998) and Pinkel and Albertson (2005). In a typical aCGH experiment, genomic DNA is extracted from test and reference samples and differentially labeled with 2 dyes. The labeled DNA from the 2 samples are mixed together and hybridized to a microarray spotted with DNA probes. The relative fluorescence intensities of the test DNA to that of the reference DNA at a given probe location ideally represents the relative copy number in the test genome compared to the reference genome. The data from an aCGH experiment is generally in the form of log intensity ratios, ordered according to the physical location of the probes along a chromosome. At a given location, if the amount of test DNA is greater than, equal to, or less than the amount of reference DNA, the observed log intensity ratio would be greater than, equal to, or less than zero, respectively. Thus, CNVs in the test genome are signified by deviations from zero in the data.

There are 2 main goals when analyzing aCGH data. The first goal is to estimate the relative copy number at each probe location along the chromosome. The second goal is to identify the regions within the chromosome that display CNV. Due to the genetic mechanisms that lead to CNV, these regions of CNV tend to occur in contiguous blocks, with probe locations within a block sharing the same relative copy number. In recent years, there has been much work done in developing statistical methods for analyzing aCGH data. These include methods based on change-point detection (Olshen *and others*, 2004; Venkatraman and Olshen, 2007), hidden Markov models (Fridlyand *and others*, 2004; Stjernqvist *and others*, 2007), Gaussian models (Picard *and others*, 2005; Hupé *and others*, 2004), latent variable models (Engler *and others*, 2006; Broët and Richardson, 2006; Lai *and others*, 2008), wavelets (Hsu *and others*, 2005), quantile regression (Eilers and de Menezes, 2005; Li and Zhu, 2007), and the fused lasso (Tibshirani and Wang, 2008). A review comparing some of these methods can be found in Lai *and others* (2005) and Willenbrock and Fridlyand (2005).

All the methods mentioned above focus on the analysis of aCGH data obtained from a single experiment. When we have data from multiple experiments, for example, data obtained from a group of cancer patients, these samples should be analyzed collectively rather than individually. A group of related aCGH samples are likely to display shared regions of CNV, for example, in cancer because shared CNVs correspond to specific cancer genes that when gained (or lost) provide selective growth advantage. By analyzing the group as a whole, we are able to draw strength across the samples. This enables us to identify regions that may not be detected in an individual analysis of any sample and also increases our confidence in the legitimacy of any resulting calls of copy number gain or loss. However, it is important that we still maintain the ability to detect any differences that may be present among the samples, for example, small subgroups of samples that share similar patterns of CNV. Some recent work on multi-sample aCGH data analysis includes methods by Diskin *and others* (2006) (Significance Testing of Aberrant Copy number [STAC]), Guttman *and others* (2007) (Multiple Sample Analysis [MSA]), Beroukhim *and others* (2007)

(Genomic Identication of Signicant Targets in Cancer [GISTIC]), and Witten *and others* (2009) (penalized matrix decomposition [PMD]).

Many current approaches for analyzing multi-sample aCGH data involve looking at the frequency of CNVs over all samples and using a threshold to make calls. The problem with basing an analysis on an overall frequency is the resulting difficulty in discovering any heterogeneity (e.g. subgroup structure) that may be present among the samples. Other approaches involve using single-sample methods as a first step to determine the calls for each sample. This has the potential drawback of not taking full advantage of any shared information among the samples. In this paper, we propose a method called the "Fused Lasso Latent Feature Model" (FLLat) for analyzing multi-sample aCGH data that attempts to address the deficiencies of current methods. Initially, we model the samples of aCGH data using a latent feature model. For the identification and quantification of the regions of CNV, we apply the fused lasso penalty to the latent features. FLLat is motivated and described in detail in Section 2. Estimation of the model parameters is covered in Section 3. Finally, some simulations and data analysis are given in Sections 4 and 5, respectively.

2. Fused lasso latent feature model

2.1 *Latent feature model for multi-sample aCGH data*

As mentioned in Section 1, we believe that through inheritance and mutations (combined with selective pressures), there are regions of CNV that are shared among a group of related samples. We would like our procedure to draw strength from these similarities among the samples while at the same time identifying any heterogeneity that may exist. With this in mind, we propose the following latent feature model to model multi-sample aCGH data:

$$y_{ls} = \sum_{j=1}^{J} \beta_{lj}\theta_{js} + \epsilon_{ls}, \tag{2.1}$$

where y_{ls} denotes the observed log intensity ratio at probe location l ($l = 1, \ldots, L$) for sample s ($s = 1, \ldots, S$). Equivalently, we can express the model using matrix notation:

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\Theta} + \boldsymbol{E}, \tag{2.2}$$

where $\boldsymbol{Y}$ is an $L \times S$ matrix with $\boldsymbol{Y}_{l,s} = y_{ls}$, $\boldsymbol{B}$ is an $L \times J$ matrix with $\boldsymbol{B}_{l,j} = \beta_{lj}$, $\boldsymbol{\Theta}$ is a $J \times S$ matrix with $\boldsymbol{\Theta}_{j,s} = \theta_{js}$, and $\boldsymbol{E}$ is an $L \times S$ matrix with $\boldsymbol{E}_{l,s} = \epsilon_{ls}$.

The model states that each sample $\boldsymbol{y}_{\cdot s} = (y_{1s}, \ldots, y_{Ls})^T$ can be expressed as a weighted linear combination of J latent features plus some noise, with the $\boldsymbol{\beta}_{\cdot j} = (\beta_{1j}, \ldots, \beta_{Lj})^T$, for $j = 1, \ldots, J$, representing the latent features and $\boldsymbol{\theta}_{\cdot s} = (\theta_{1s}, \ldots, \theta_{Js})^T$ representing the weights. Typically $J < S$, where S is the number of samples. Also, we fit the model separately for each chromosome, so that $\boldsymbol{y}_{\cdot s} = (y_{1s}, \ldots, y_{Ls})^T$ corresponds to the observed log intensity ratios for a given sample along a single chromosome. Figure 1 describes how a sample of aCGH data, ignoring noise, is derived from this model for a simple example with $J = 3$ latent features.

The motivation behind the model is that the J features collectively summarize the important characteristics, with respect to CNV, of the group of samples. Specifically, each feature represents a particular pattern of CNV. The weights for a given sample then determine how much each feature contributes to that sample. In other words, the features can be thought of as common ingredients that are shared by each sample. The weights are then the recipe that determines the composition of each sample. The weights can also give insight into the distribution of the features among the samples. For example, looking at all the weights that are applied to a particular feature can tell us how often and how prominently that feature

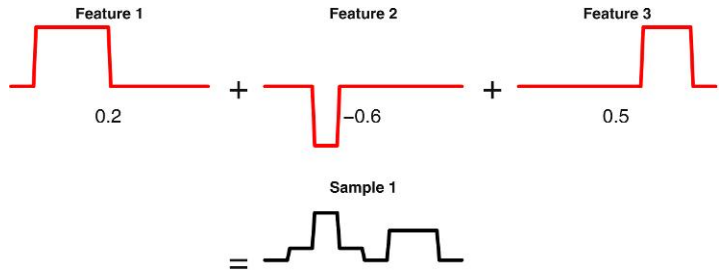

Fig. 1. An example of a sample of aCGH data, ignoring noise, generated using the latent feature model with 3 latent features. The latent features ($\boldsymbol{\beta}_{\cdot 1}$, $\boldsymbol{\beta}_{\cdot 2}$, and $\boldsymbol{\beta}_{\cdot 3}$) are displayed in the top line. The weights corresponding to Sample 1 are $\theta_{11} = 0.2$, $\theta_{21} = -0.6$, and $\theta_{31} = 0.5$.

appears among the samples. Finally, given estimates of the features and weights, $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{\Theta}}$, the estimated relative copy number at each probe location for each sample is given by the fitted values, $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{B}}\hat{\boldsymbol{\Theta}}$.

2.2 *Applying the fused lasso*

As described previously, regions of CNV tend to occur in contiguous blocks throughout the chromosome, with probe locations within a block having the same relative copy number. For the rest of the chromosome not displaying CNV, the expected log intensity ratio should be zero. Therefore, if we treat the aCGH data as a 1D signal along the chromosome, the majority of the signal is zero, with the nonzero regions occurring in smooth blocks. This combination of sparsity and smoothness for a 1D signal leads us intuitively toward the fused lasso signal approximator (FLSA, Tibshirani *and others*, 2005; Friedman *and others*, 2007). When we have ordered outcomes y_i, for $i = 1, \ldots, n$, the FLSA solves the following optimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=2}^{n} |\beta_i - \beta_{i-1}|. \tag{2.3}$$

Here, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of parameters that estimates the ordered outcomes. The first penalty term, which penalizes the size of each parameter, encourages the solution to be sparse, and the second penalty term, which penalizes the absolute difference between adjacent parameters, encourages the solution to be smooth. There are 2 corresponding tuning parameters, λ_1 and λ_2, that control the amount of sparsity and smoothness, respectively.

The application of the FLSA to aCGH data in the single-sample case is described in Tibshirani and Wang (2008). Briefly, they set $(y_1, \ldots, y_n)^T$ to be the observed log intensity ratios and used the FLSA to estimate the inferred relative copy number ratios, represented by the single parameter vector $\boldsymbol{\beta}$. In the multi-sample case, the situation is slightly different since we now want to estimate multiple parameter vectors, specifically, the J features $\boldsymbol{\beta}_{\cdot 1}, \ldots, \boldsymbol{\beta}_{\cdot J}$ in model (2.1). Recall that each feature describes a particular pattern of CNV. Therefore, we will apply the fused lasso penalty to each feature in order to encourage both smoothness and sparsity in each estimated feature. Specifically, to fit model (2.1), we estimate the β_{lj} and the θ_{js} by minimizing the following criterion:

$$F(\boldsymbol{B}, \boldsymbol{\Theta}) = \sum_{s=1}^{S} \sum_{l=1}^{L} \left(y_{ls} - \sum_{j=1}^{J} \beta_{lj}\theta_{js} \right)^2 + \sum_{j=1}^{J} P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}_{\cdot j}) \tag{2.4}$$

or in matrix notation:

$$F(\boldsymbol{B}, \boldsymbol{\Theta}) = \|\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\Theta}\|_F^2 + \sum_{j=1}^{J} P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}_{\cdot j}), \tag{2.5}$$

where $P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}_{\cdot j}) = \lambda_1 \sum_{l=1}^{L} |\beta_{lj}| + \lambda_2 \sum_{l=2}^{L} |\beta_{lj} - \beta_{l-1,j}|$.

The first term in (2.4) is the usual sum of squared errors, and the second term is the fused lasso penalty, applied to each feature. We can solve this minimization problem using an alternating least squares–type algorithm where we alternate between fixing $\boldsymbol{B}$ and solving for $\boldsymbol{\Theta}$ and vice versa, until the solutions converge. Regarding the issue of convergence, this is a biconvex optimization problem and as such there potentially exist many local minima. Therefore, we are not guaranteed to converge to the global minimum. However, provided the criterion (2.4) is decreased at each step of the algorithm, we will converge to a local minimum. One possible way to increase the chance of attaining the global minimum is to run the alternating algorithm with a range of different initial values and choose the solution that achieves the smallest criterion. The initial values can be chosen by setting $\boldsymbol{B}$ to be a random selection of J columns of $\boldsymbol{Y}$. Alternatively, we can initialize the algorithm by setting $\boldsymbol{B}$ to be the first J principal components of $\boldsymbol{Y}$.

2.3 *Constraining the weights*

For the estimates of the features to be useful, it is necessary to appropriately constrain the weights. Leaving the weights unconstrained can lead to problems with the fused lasso penalty and also with model identifiability. For example, multiplying a particular feature $\boldsymbol{\beta}_{\cdot j}$ by a constant $0 < c < 1$, and dividing the corresponding weights by the same constant leaves the fit unchanged, but reduces the penalty. With these issues in mind, we placed the following L_2 constraint on the weights:

$$\sum_{s=1}^{S} \theta_{js}^2 \leqslant 1 \text{ for each } j. \tag{2.6}$$

On a technical note, we do not constrain the different weight vectors to be orthogonal to each other. Although this would simplify their estimation somewhat, it would interfere with the interpretation of the penalties. Constraint (2.6) places a restriction on the size of each row of $\boldsymbol{\Theta}$, that is, the weights corresponding to a given feature. We felt that this was the most suitable way of constraining the size of the weights. First, it makes direct comparisons among the estimated features more meaningful. For example, a larger feature would indicate that it appeared more prominently among the samples. Second, it prevents the majority of the weight being distributed onto only a few features, which would potentially allow these features to circumvent the fused lasso penalty.

2.4 *The number of features J*

Inherent in model (2.1) is that a choice needs to be made for the number of features J. Theoretically, J can take any value in $\{1, \ldots, S\}$, where S is the number of samples. The "best" choice of J for any given data set is difficult to determine and is likely to depend on a number of factors, for example, the level of noise, the value of the tuning parameters λ_1 and λ_2, and S. Therefore, the value of J is generally left to the user to specify, with the default set as $\min\{15, S/2\}$. Alternatively, we provide a semiautomatic process for choosing J that is based on the percentage of variation explained (PVE). For a given value of J, the PVE is defined to be

$$\text{PVE}_J = 1 - \frac{\sum_{s=1}^{S} \sum_{l=1}^{L} \left(y_{ls} - \sum_{j=1}^{J} \hat{\beta}_{lj}\hat{\theta}_{js}\right)^2}{\sum_{s=1}^{S} \sum_{l=1}^{L} (y_{ls} - \bar{y}_s)^2}, \tag{2.7}$$

where $\hat{\beta}_{lj}$ and $\hat{\theta}_{js}$ are the estimates produced by FLLat and $\bar{y}_s = \sum_{l=1}^{L} y_{ls}/L$. As more features are added to the model, the estimated fit is improved and the PVE increases. However, after a certain point, additional features will not significantly improve the estimated fit and will essentially be superfluous. Consequently, the PVE will tend to plateau beyond this point. Thus, by plotting the PVE against the number of features, users can choose a value of J at which the PVE begins to plateau.

3. PARAMETER ESTIMATION FOR THE FUSED LASSO LATENT FEATURE MODEL

3.1 *Estimating* $\boldsymbol{B}$ *and* $\boldsymbol{\Theta}$

When $\boldsymbol{\Theta}$ is held fixed, we use a block coordinate descent approach (see Tseng, 1988, 2001) to estimate $\boldsymbol{B}$. That is, we estimate each feature $\boldsymbol{\beta}_{\cdot j}$ then cycle through $j = 1, \ldots, J$ until the estimates converge. Specifically, for fixed $\boldsymbol{\Theta}$ and $\{\boldsymbol{\beta}_{\cdot k}\}_{k \neq j}$, the solution for $\boldsymbol{\beta}_{\cdot j}$ is given by

$$\boldsymbol{\beta}_{\cdot j} = \underset{\boldsymbol{\beta}_{\cdot j}}{\operatorname{argmin}} \sum_{l=1}^{L} (\check{y}_{lj} - \beta_{lj})^2 + P_{\check{\lambda}_1, \check{\lambda}_2}(\boldsymbol{\beta}_{\cdot j}), \tag{3.1}$$

where $\check{y}_{lj} = \sum_{s=1}^{S} \check{y}_{lsj}\theta_{js} / \sum_{s=1}^{S} \theta_{js}^2$, $\check{y}_{lsj} = y_{ls} - \sum_{k \neq j} \beta_{lk}\theta_{ks}$, and $\check{\lambda} = \lambda / \sum_{s=1}^{S} \theta_{js}^2$. We can solve (3.1) by applying the FLSA to $\left(\check{y}_{1j}, \ldots, \check{y}_{Lj}\right)^T$. Since this solution depends on $\{\boldsymbol{\beta}_{\cdot k}\}_{k \neq j}$, we cycle through each j until the solutions converge

When $\boldsymbol{B}$ is held fixed, we again use a block coordinate descent approach to estimate $\boldsymbol{\Theta}$. We estimate the weights $\boldsymbol{\theta}_{j\cdot} = (\theta_{j1}, \ldots, \theta_{jS})^T$ then cycle through $j = 1, \ldots, J$ until the estimates converge. For fixed $\boldsymbol{B}$ and $\{\boldsymbol{\theta}_{k\cdot}\}_{k \neq j}$, the solution for $\boldsymbol{\theta}_{j\cdot}$ is given by

$$\boldsymbol{\theta}_{j\cdot} = (\tilde{y}_{j1}, \ldots, \tilde{y}_{jS}) \Big/ \max\left\{ \sum_{l=1}^{L} \beta_{lj}^2, \left(\sum_{s=1}^{S} \tilde{y}_{js}^2 \right)^{1/2} \right\}, \tag{3.2}$$

where $\tilde{y}_{js} = \sum_{l=1}^{L} \check{y}_{lsj}\beta_{lj}$ and $\check{y}_{lsj} = y_{ls} - \sum_{k \neq j} \beta_{lk}\theta_{ks}$, for $s = 1, \ldots, S$. Since this solution depends on $\{\boldsymbol{\theta}_{k\cdot}\}_{k \neq j}$, we cycle through each j until the solutions converge.

Detailed derivations of (3.1) and (3.2), and an algorithm for estimating $\boldsymbol{B}$ and $\boldsymbol{\Theta}$, are given in Section S.1 of the supplementary material available at *Biostatistics* online.

3.2 *Selecting the fused lasso tuning parameters* λ_1 *and* λ_2

In general, the selection of optimal tuning parameters for a given model can be a difficult task, which is further complicated as the number of tuning parameters increases. To simplify the search for the optimal tuning parameters, we reparameterize λ_1 and λ_2 in terms of λ_0 and $\alpha \in (0, 1)$ such that $\lambda_1 = \alpha\lambda_0$ and $\lambda_2 = (1-\alpha)\lambda_0$. We can think of λ_0 as an overall tuning parameter with α determining how much emphasis is placed on sparsity versus smoothness. By fixing the possible values that α can take, we effectively reduce the search over 2 parameters, λ_1 and λ_2, to a search over one parameter, λ_0.

Specifically, we initially fixed the possible values of α (e.g. {0.1, 0.3, 0.5, 0.7, 0.9}). For each value of α, we determined the value of λ_0 that resulted in each estimated feature being constant, denoting this value by $\lambda_{0,\alpha}^{\max}$. We then chose a fixed number of candidate values (e.g. 5) for λ_0 from the interval $(0, \lambda_{0,\alpha}^{\max})$. The optimal values of α and λ_0 were selected by searching over this 2D grid for the values that minimized the following criterion:

$$(SL) \cdot \log\left(\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{B}}\hat{\boldsymbol{\Theta}}\|_F^2}{SL} \right) + k_{\alpha,\lambda_0} \log(SL). \tag{3.3}$$

Here, we define $k_{\alpha,\lambda_0} = \sum_{j=1}^{J} k_{\alpha,\lambda_0}(j)$, where $k_{\alpha,\lambda_0}(j)$ is the number of unique nonzero elements in the jth feature, $\boldsymbol{\beta}_{\cdot j}$. The term k_{α,λ_0} represents the complexity of the model, with larger values indicating greater complexity. S and L are the number of samples and probe locations, respectively. Criterion (3.3) is similar to the Bayesian information criterion when we assume that the model errors in (2.1) are normally distributed. The rationale behind this criterion is that by minimizing (3.3), we are attempting to find an appropriate model without overfitting the data, as the first term will tend to be smaller for complex models, whereas the second term will tend to be smaller for simple models. For computational reasons, we prefer this approach for selecting the optimal tuning parameters as opposed to an approach based on cross-validation.

4. Simulation studies

4.1 *Comparing FLLat to single-sample methods*

Here, we present some simulations to demonstrate the advantages that a true multi-sample approach has over single-sample approaches. We simulated 3 different data sets, each consisting of $S = 20$ samples and $L = 1000$ probes. For the first 2 data sets, samples were generated in a manner similar to that employed by Olshen *and others* (2004). We used the model $y_{ls} = \mu_{ls} + \epsilon_{ls}$, $l = 1, \ldots, L$, $s = 1, \ldots, S$, where μ_{ls} is the mean and $\epsilon_{ls} \sim N(0, \sigma^2)$ with σ determined by the signal-to-noise ratio (SNR), as described below. The mean is given by $\mu_{ls} = \sum_{m=1}^{M_s} c_m I_{\{l_m \leqslant l \leqslant l_m + k_m\}}$, where M_s is the number of segments generated for sample s and c_m, l_m, and k_m are the height, starting position, and length, respectively, of each segment. For the first data set, the samples were designed to share no segments (regions of CNV). Therefore, separately for each sample, we chose the value of M_s from $\{1, 2, 3, 4, 5\}$, then chose c_m from $\{\pm 1, \pm 2, \pm 3, \pm 4, \pm 5\}$, l_m from $\{1, \ldots, L - 100\}$, and k_m from $\{5, 10, 20, 50, 100\}$. For the second data set, the samples were designed to share segments. We set the number of shared segments to 5 and generated starting positions and lengths for each shared segment, as above. In order to determine the number of samples in which each shared segment appeared, we selected a proportion from $(0.25, 0.75)$ for each shared segment and randomly selected the corresponding number of samples that would then share the segment. For each sample containing shared segments, heights for each shared segment were chosen as above. Finally, each sample was also populated with unshared segments, chosen as above, with the additional requirement that no sample contained more than 5 total segments. The third data set was generated according to model (2.1) with $J = 5$ and $\epsilon_{ls} \sim N(0, \sigma^2)$ with σ again determined by the SNR. The features were generated using the model $\beta_{lj} = \sum_{m=1}^{M_j} c_m I_{\{l_m \leqslant l \leqslant l_m + k_m\}}$, $l = 1, \ldots, L$, $j = 1, \ldots, J$. The value of M_j was chosen from $\{1, 2, 3\}$, whereas c_m, l_m, and k_m were chosen as above. The weights θ_{js} were generated by creating a matrix of $N(0, 1)$ variables and normalizing the rows to satisfy (2.6).

We applied FLLat, cghFLasso (Tibshirani and Wang, 2008), quantsmooth (Eilers and de Menezes, 2005), CBS (Venkatraman and Olshen, 2007) and PMD (L_1, FL) (Witten *and others*, 2009, a multi-sample method) to each of the 3 data sets for SNRs of 0.1, 0.5, and 1. The SNR was defined to be the mean magnitude of the aberrations divided by σ. Here, an aberration is any probe with nonzero signal, where the signal is given by the μ_{ls} for data sets 1 and 2 and $\boldsymbol{B\Theta}$ for data set 3. For each method, we generally used the default settings but also followed any recommendations and attempted to apply any tuning procedures, as outlined in the software documentation. In particular, for FLLat, we used the PVE plots in Figure 3 to choose J and used criterion (3.3) to select the optimal tuning parameters; for PMD (L_1, FL), we set the number of factors to J (to be comparable with FLLat) and used the provided cross-validation function when estimating each factor; for quantsmooth, as the provided cross-validation function produced an error on the simulated data, we used the default value for the smoothing parameter. A comparison of the computation times for each method can be found in Section S.2 of the supplementary material available at *Biostatistics* online.

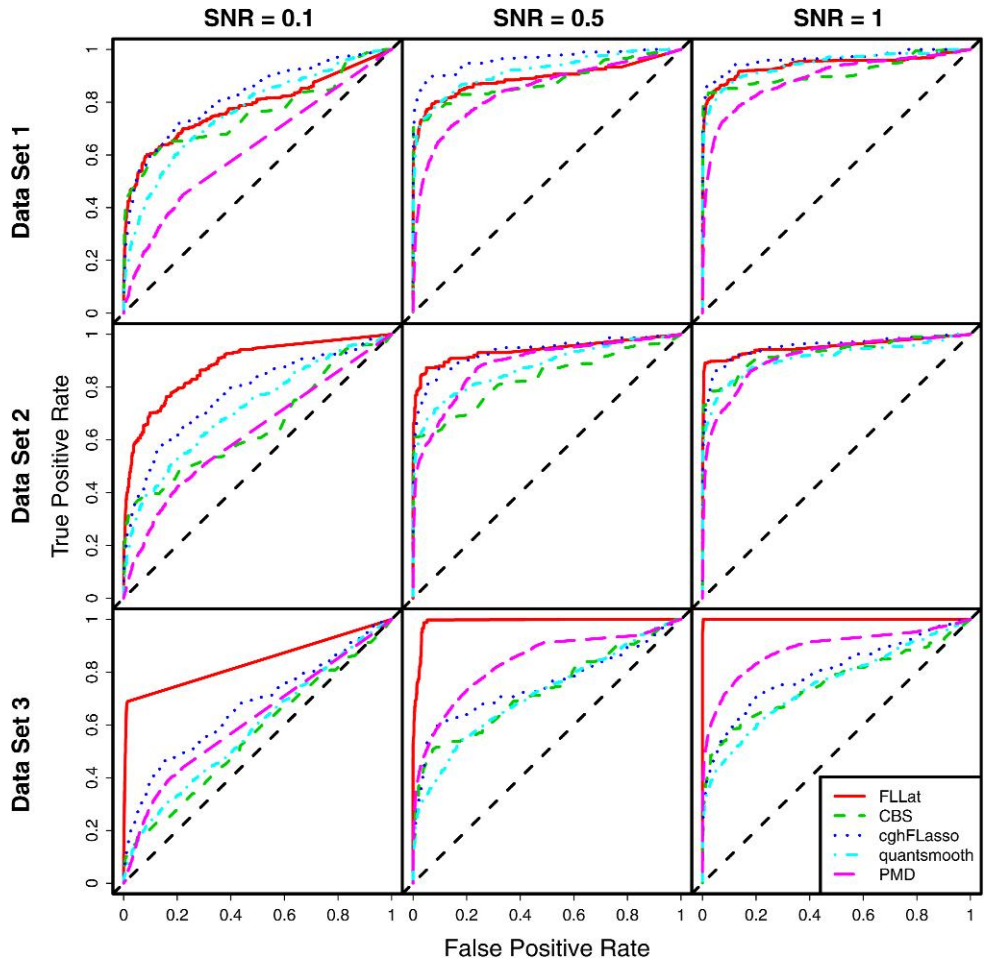

Fig. 2. ROC curves comparing FLLat to single-sample methods. FLLat is represented by the solid line. The 3 simulated data sets correspond to samples generated with no shared segments (data set 1), samples generated with shared segments (data set 2) and samples generated from model (2.1) (data set 3).

For each method, we generated receiver operating characteristic (ROC) curves, shown in Figure 2, by comparing the true signal to the estimated signal. For each sample, any probe that had an estimated signal greater in magnitude than a fixed threshold was declared an aberration. The ROC curves were produced by varying the threshold and calculating the true positive and false positive rates for each value of the threshold. The true positive rate was defined to be the proportion of true aberrations that were declared to be aberrations by a given method. Similarly, the false positive rate was defined to be the proportion of true nonaberrations that were declared to be aberrations.

When comparing ROC curves, curves that lie further above the 45° line indicate better performance. For the first data set, where the samples share no segments, although FLLat is not likely to have much advantage over single-sample methods, we see that it still performs comparably to the other methods. For the second data set, where samples now shared segments, FLLat performed very well, and this is especially evident at the lowest SNR. Finally, for the third data set, where the samples are generated from

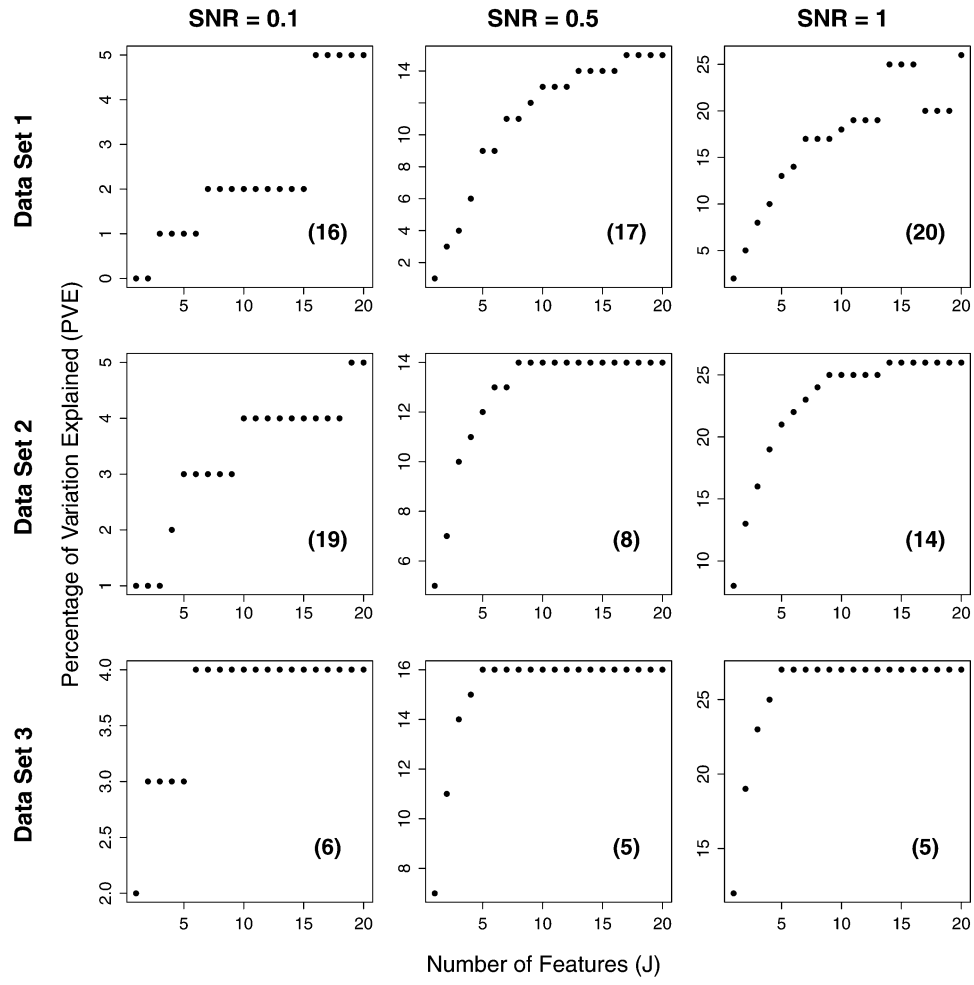

Fig. 3. The percentage of variation explained by FLLat (from (2.7)) as we vary the number of features for the 3 data sets of Section 4.1 at SNRs of 0.1, 0.5, and 1. The numbers in brackets indicate the value of J chosen from the particular PVE plot.

model (2.1), FLLat significantly outperformed all the other methods. Overall, these simulations show that FLLat performs very favorably compared to single-sample methods when samples truly share similarities. Further, FLLat performs particularly well in situations with low SNRs, which is often the case with aCGH data. With regard to the PVE plots in Figure 3, we see that for the third data set, where we know the true number of features ($J = 5$), the PVE accurately determines the correct number of features.

4.2 *Estimating the FDR*

For a given threshold T, letting $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{B}}\hat{\boldsymbol{\Theta}}$ denote the fitted values produced by FLLat, we can declare probe location l for sample s to be an aberration if $|\hat{y}_{ls}| \geqslant T$. Given these declared aberrations, we would like to have a measure of the proportion that are falsely called. Using a similar approach to

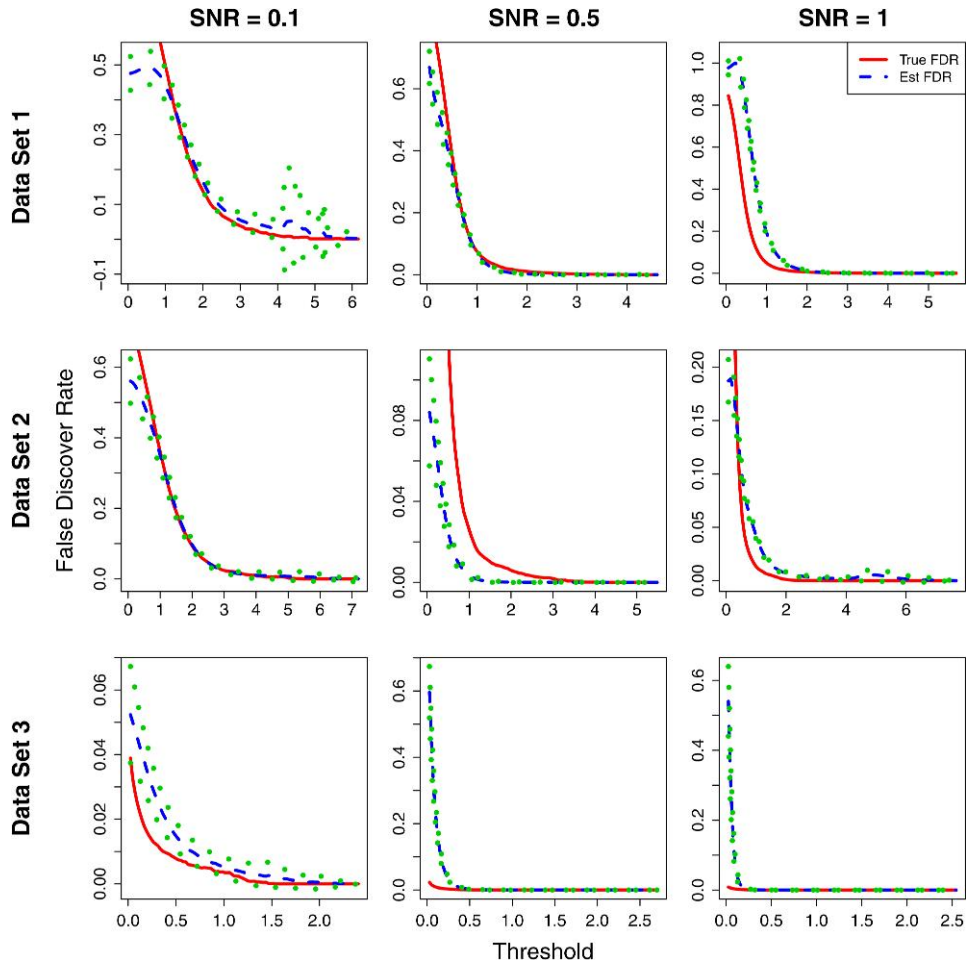

Fig. 4. The true FDR (solid line) and the average estimated FDR (dashed line) ± one standard deviation (dotted line) for the 3 data sets of Section 4.1 at SNRs of 0.1, 0.5, and 1. The average and standard deviation of the estimated FDR were calculated from 100 realizations of each simulated data set.

Tibshirani and Wang (2008), identifying aberrations can be thought of as a multiple-testing problem, where we are testing the following hypothesis for each probe location within each sample:

$$H_0(l, s) = \text{no aberration at probe location } l \text{ for sample } s.$$

The false discovery rate (FDR, Benjamini and Hochberg, 1995), for threshold T, is then defined to be

$$\text{FDR}(T) = \mathbf{E}\left(\frac{\sum_{s=1}^{S} \sum_{l=1}^{L} I(|\hat{y}_{ls}| \geqslant T) I(H_0(l, s) \text{ is true})}{\sum_{s=1}^{S} \sum_{l=1}^{L} I(|\hat{y}_{ls}| \geqslant T)} \right). \tag{4.1}$$

The numerator within the expectation is the number of declared aberrations that are not true aberrations and the denominator is the total number of declared aberrations. We propose a permutation-based method for estimating the FDR.

To estimate (4.1), we need some information regarding the null distribution of the data. If a reference data set were available, we could apply FLLat to this reference data set and use the resulting declared aberrations to estimate the numerator of (4.1). When an appropriate reference data set is not available, which is often the case in real experiments, an alternative is to approximate the null distribution of the data. Under the null hypothesis, the log intensity ratio for probe location l in sample s should be distributed as random noise, with no correlation between neighboring locations. One way to approximate this null distribution is to permute the probe locations within each sample. This has the effect of destroying any linear structure along the chromosome. Suppose the data were permuted in this fashion K times, and let $\hat{\boldsymbol{Y}}^k$ denote the fitted values produced by applying FLLat to the kth permuted data set. Then the FDR can be estimated by

$$\widehat{\text{FDR}(T)} = \frac{\pi_0 \sum_{k=1}^{K} \left(\sum_{s=1}^{S} \sum_{l=1}^{L} I\left(|\hat{y}_{ls}^k| \geqslant T\right) \right) \Big/ K}{\sum_{s=1}^{S} \sum_{l=1}^{L} I\left(|\hat{y}_{ls}| \geqslant T\right)}, \tag{4.2}$$

where π_0 is the proportion of true null hypotheses.

We calculated both the true FDR and also the permutation-based estimate for each of the 3 data sets described in Section 4.1 at the same 3 SNRs of 0.1, 0.5, and 1. Plots of the true FDR and the average estimated FDR (calculated from 100 realizations of each simulated data set) for varying threshold values are displayed in Figure 4. The estimated FDR was based on $K = 20$ permuted data sets and the true value of $\pi_0 = \sum_{s=1}^{S} \sum_{l=1}^{L} I\left(\sum_{j=1}^{J} \beta_{lj} \theta_{js} = 0 \right) \Big/ S \times L$ (i.e. the proportion of total probe locations having zero true signal) was used in (4.2). The true value of π_0 is generally unknown and can either be estimated from the data, which can be a difficult task or set to the upper bound of 1, which results in conservative estimates of the FDR. We see from Figure 4 that in each plot, the estimated FDR approximates the true FDR fairly well for smaller values of the FDR, although there are some deviations at the larger values.

5. Analysis of breast cancer data

To explore the performance of FLLat on real data, we analyzed some data from Pollack *and others* (2002). The data set consisted of aCGH data across 6691 mapped human genes for 44 locally advanced primary breast tumors. We focused our analyses on chromosomes 8 and 17, as the study by Pollack *and others* (2002) found these chromosomes to exhibit extensive CNV. Of the 6691 genes, 241 were from chromosome 8 and 382 from chromosome 17. We applied FLLat separately for each chromosome. The number of features J was chosen by examining the percentage of variation explained over a range of values of J and the tuning parameters were selected using criterion (3.3). The results and figures for chromosome 8 are displayed in Section S.3 of the supplementary material available at *Biostatistics* online.

Displayed in Figure 5 are the features produced by FLLat for chromosome 17. The features are plotted in order of decreasing total magnitude, $\sum_{l=1}^{L} \hat{\beta}_{jl}^2$. The orientation of each feature was determined by sign $\left(\sum_{s=1}^{S} \hat{\theta}_{js} / S \right)$, which essentially corresponds to the most common orientation of the feature among the samples. Five features were found for chromosome 17, displaying gains in the 17q arm. For chromosome 8, six features were discovered, exhibiting some gains in the 8q arm and some losses in the 8p arm. Also included in Figure 5 is a heatmap of the estimated weights which indicates how much each feature contributes to a given sample. The samples have also been clustered according to their weights, with the potential to reveal sample group structure.

We also applied STAC (Diskin *and others*, 2006) to the same data. Results for chromosome 8 can again be found in Section S.3 of the supplementary material available at *Biostatistics* online. STAC takes as input binary data where, for a particular location, a 1 signifies an aberration and a 0 signifies no aberration. STAC also requires gains and losses to be analyzed separately. Therefore, in order to generate the appropriate input data, we first applied cghFLasso to each individual sample. This produced an estimated signal for

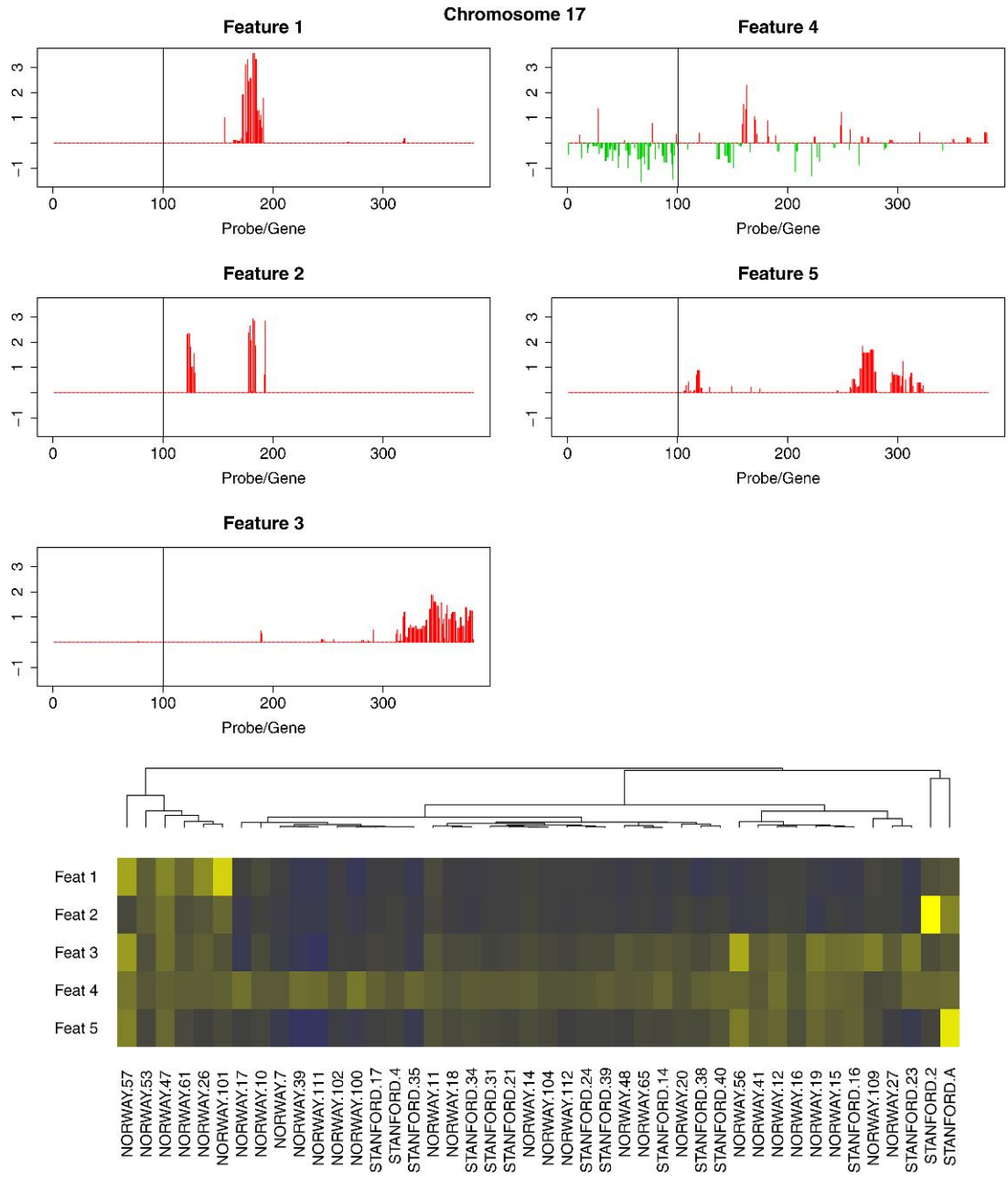

Fig. 5. Top: the 5 estimated features produced by FLLat for chromosome 17. The vertical black line indicates the approximate position of the centromere. Bottom: a heatmap of the estimated weights. Blue and yellow in the color online version indicate negative and positive weights, respectively. The samples have also been clustered based on their weights.

each sample. For the input data for gains, we then set any probe location with an estimated signal greater than 0 to 1 and to 0 otherwise. Similarly, for the input data for losses, we set any probe location with an estimated signal less than 0 to 1 and to 0 otherwise. We note that the STAC analysis clearly depends

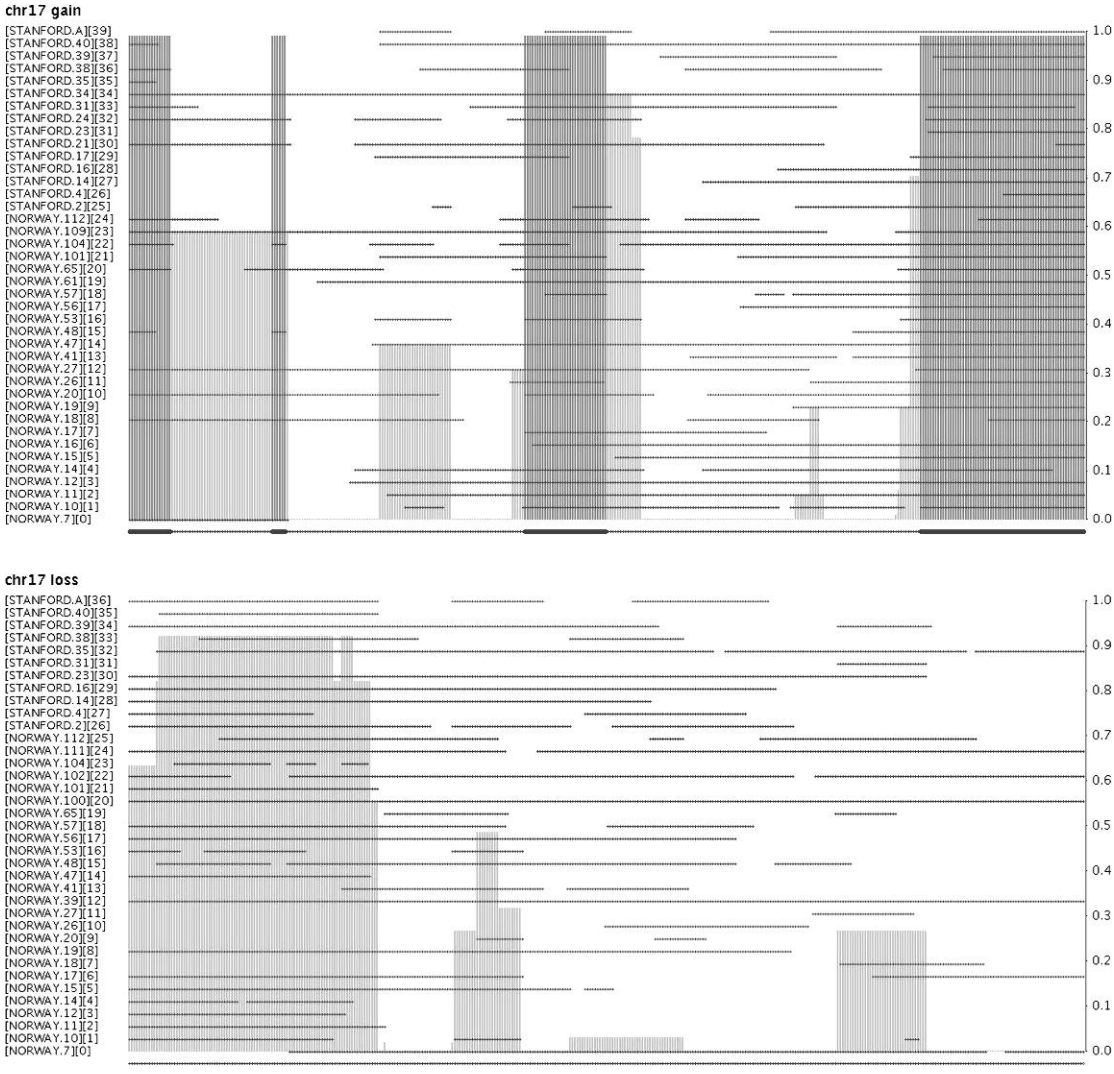

Fig. 6. The STAC analysis for gains (top) and losses (bottom) on chromosome 17. The gray bars represent the STAC confidence ($1-p$, where p is the footprint-based p-value) for each location, which are ordered along the chromosome. Darker bars indicate a confidence greater than 0.95. Significant gains were identified in both the 17p and 17q arms.

on the single-sample method that was used to make the initial calls, thus, the results of the analysis may be different if another single-sample method was used. Diskin *and others* (2006) recommend analyzing each arm of a chromosome separately. However, since FLLat was applied to the whole chromosome, for a more direct comparison, we did likewise with STAC. STAC calculates a p-value for each probe location based on a "footprint" statistic. These p-values for each probe location on chromosome 17 are displayed in Figures 6. The height of the bars represents the quantity $1 - p$, with the darker bars corresponding to $1 - p$ greater than 0.95. Based on these p-values, STAC has identified some gains in both the 17p and 17q arms and a region of loss in the 17p arm (and also some gains towards the end of the 8q arm and perhaps a region of loss in the 8p arm).

The gains and losses identified for chromosomes 8 and 17 by FLLat are consistent with the findings of Pollack *and others* (2002). Some of these regions contain known oncogenes, for example, FGFR1 (8p11), MYC (8q24), and ERBB2 (17q12). Further, it has been shown by Pollack *and others* (2002) that a significant proportion of highly amplified genes on chromosome 17 were also highly expressed. This indicates that CNVs strongly influence the transcriptional program of human breast tumors.

The findings of the STAC analysis generally agree with the regions of gains and losses identified by FLLat. However, FLLat has a number of practical advantages. Since FLLat produces multiple features for each chromosome, we feel that FLLat is able to capture more detailed structure regarding gains and losses than STAC. Also, the weights reveal some interesting relationships among the samples. For example, based on the clustering of the weights for chromosome 17, displayed in Figure 5, we see that there appear to be 3 distinct groups of samples. From the heatmap of the weights, we also see that feature 4, which represents some losses in the p arm, appears in almost all the samples, whereas feature 1, which represents some gains in the q arm, is localized to the group of samples at the far left of the dendrogram. Further, for this group of samples, feature 1 seems to consistently co-occur with feature 2.

6. DISCUSSION

Although much work has been done on identifying regions of CNV in single-sample aCGH data, the analysis of multi-sample aCGH data is a relatively new area. When dealing with multi-sample aCGH data, it is important to take advantage of the similarities among samples while also maintaining the ability to identify any heterogeneity that may be present. Some of the current methods for analyzing multi-sample data use frequency thresholds for calling gains and losses, which can obscure important differences (like subgrouping) between samples or rely on single-sample methods for making initial calls of gain or loss, which does not take full advantage of the similarities among samples and can lead to false negatives.

We proposed a method called FLLat. This method involves modeling the aCGH data with a latent feature model, where each sample is modeled by a weighted combination of a fixed number of features. These features represent the key regions of CNV for the group of samples and combined with the weights describe the regions of CNV for each individual sample. We used the fused lasso penalty in the estimation of the features, which encourages both smoothness and sparsity in the estimates. This is a desirable property given that regions of CNV tend to occur in infrequent contiguous blocks along the chromosome.

Our simulation studies showed that FLLat outperformed single-sample methods when the simulated samples shared common information. Further, we found that FLLat is able to effectively draw strength from the similarities among samples and performs quite well in situations of low SNR. When applied to some aCGH data from human breast tumors, FLLat identified regions of gain in 8q and 17q and regions of loss in 8p. These were consistent with previous findings by Pollack *and others* (2002). STAC identified similar regions when applied to the same data. However, FLLat was able to find more detailed patterns of CNV and also discovered interesting relationships among the samples.

FLLat is a fast, flexible procedure for analyzing multi-sample aCGH data and produces interesting and interpretable results. With regard to computational time and the complexity of the method, the following example demonstrates how the number of samples (S) and probes (L) can affect the run times. We applied FLLat to 4 data sets, consisting of $S = \{50, 500\}$ and $L = \{1000, 10\,000\}$. The data sets were generated according to the model used to simulate the second data set in Section 4.1, with the SNR set to 1. We set the number of features (J) to the default value of 15, and the values of λ_1 and λ_2 were chosen using criterion (3.3) on the smallest data set ($S = 50$ and $L = 1000$) and were then also used for the other data sets. For $S = 50$, the run times were 1.97 s ($L = 1000$) and 52.36 s ($L = 10\,000$) and for $S = 500$, they were 83.22 s ($L = 1000$) and 393.54 s ($L = 10\,000$). The run times will also depend on factors such as the values of λ_1 and λ_2, the SNR and how much information is shared among samples. FLLat scales relatively

well to large data sets mainly due to the fast and efficient algorithms available for solving the fused lasso problem. An R package, FLLat, will be available from the Comprehensive R Archive Network.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank Nicholas Johnson and Holger Hoefling for developing fast and efficient algorithms for solving the fused lasso. The authors also thank the editors and reviewers for their helpful comments which led to improvements in the paper. *Conflict of Interest:* None declared.

Funding

National Science Foundation (DMS-9971405 to R.T.); National Institutes of Health (N01-HV-28183 to R.T., 5R01 EB 001988-13 to T.H.).

References

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 (Methodological)* **57**, 289–300.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S. *and others* (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007–20012.

Broët, P. and Richardson, S. (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics* **22**, 911–918.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert, 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.

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

Engler, D. A., Mohapatra, G., Louis, D. N. and Betensky, R. A. (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics* **7**, 399–421.

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.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332.

Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S. J., Baldwin, D. A., Stoeckert, Jr, C. J. and Grant, G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* **3**, 1464–1486.

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.

Hupé, P., Stransky, N., Thiery, J. P., Radvanyi, F. and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422.

LAI, T. L., XING, H. AND ZHANG, N. (2008). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290–307.

LAI, W. R., JOHNSON, M. D., KUCHERLAPATI, R. AND PARK, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770.

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

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. AND WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572.

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

PINKEL, D. AND ALBERTSON, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature Genetics* **37** (Suppl), 11–17.

PINKEL, D., SEGRAVES, R., SUDAR, D., CLARK, S., POOLE, I., KOWBEL, D., COLLINS, C., KUO, W.-L., CHEN, C., ZHAI, Y. *and others* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics* **20**, 207–211.

POLLACK, J., SORLIE, T., PEROU, C., REES, C., JEFFREY, S., LONNING, P., TIBSHIRANI, R., BOTSTEIN, D., BORRESEN-DALE, A. AND BROWN, P. (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 of the United States of America* **99**, 12963–12968.

STJERNQVIST, S., RYDÉN, T., SKÖLD, M. AND STAAF, J. (2007). Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics* **23**, 1006–1014.

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 (Statistical Methodology)* **67**, 91–108.

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

TSENG, P. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. *Technical Report LIDS-P-1840*. Cambridge, MA: Laboratory for Information and Decision Systems, Massachusetts Institute of Technology.

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* **109**, 475–494.

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

WILLENBROCK, H. AND FRIDLYAND, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics* **21**, 4084–4091.

WITTEN, D. M., TIBSHIRANI, R. AND HASTIE T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

[*Received October 22, 2009; revised April 15, 2011; accepted for publication April 22, 2011*]