Multisample aCGH Data Analysis via Total Variation and Spectral Regularization

Xiaowei Zhou, Can Yang, Xiang Wan, Hongyu Zhao, and Weichuan Yu

Abstract—DNA copy number variation (CNV) accounts for a large proportion of genetic variation. One commonly used approach to detecting CNVs is array-based comparative genomic hybridization (aCGH). Although many methods have been proposed to analyze aCGH data, it is not clear how to combine information from multiple samples to improve CNV detection. In this paper, we propose to use a matrix to approximate the multisample aCGH data and minimize the total variation of each sample as well as the nuclear norm of the whole matrix. In this way, we can make use of the smoothness property of each sample and the correlation among multiple samples simultaneously in a convex optimization framework. We also developed an efficient and scalable algorithm to handle large-scale data. Experiments demonstrate that the proposed method outperforms the state-of-the-art techniques under a wide range of scenarios and it is capable of processing large data sets with millions of probes.

Index Terms—CNV, aCGH, total variation, spectral regularization, convex optimization

1 Introduction

Genetic diseases are caused by a variety of possible alterations in DNA sequences. Traditionally, it was believed that DNA sequences between any two unrelated human individuals are about 99.9 percent identical and the small difference is mainly attributed to single nucleotides polymorphism (SNP). However, recent studies revealed another type of genetic alternation named copy number variation (CNV), which covers more than 12 percent of the human genome [1]. A CNV is defined as a gain or loss in copies of a DNA segment [2]. CNVs can alter gene expression in cells and potentially cause genetic diseases. For instance, it was reported that individuals who carried a lower copy number of gene CCL3L1 than population average were significantly more vulnerable to HIV/acquired immunodeficiency syndrome (AIDS) [3].

One major approach to detecting CNVs is to use array-based comparative genomic hybridization (aCGH) [4]. In a typical aCGH experiment, DNA segments are extracted from test and reference samples and labeled with two different dyes. The labeled DNA segments are hybridized to a microarray spotted with DNA probes. The ratio of fluorescence intensity between the test DNA and the reference DNA ideally represents the relative copy number of the test genome compared to the reference genome. The aCGH data is generally in the form of $\log_2$-ratio. A value greater than zero indicates a gain in the copy number while a value less than zero indicates a loss.

The main goal of analyzing aCGH data is to recover true CNV signals from noisy measurements. Due to measurement noise in aCGH experiments, it is difficult to identify CNVs by simply thresholding the raw $\log_2$-ratios [4]. Traditional methods include break point detection [5], [6], [7], signal smoothing [8], [9], [10], hidden Markov models [11], [12], and variational models [13], [14], among others. Please refer to [15], [16] for a review and comprehensive comparison.

All above-mentioned methods process aCGH profiles from individuals separately. Recently, more efforts are focused on analyzing aCGH data from multiple samples simultaneously. The additional information from a group of samples proved to be useful in analysis. For example, some researchers proposed to use multisample information to normalize the data and remove the reference bias in aCGH profiles [17], [18], [19]. Some aimed to detect recurrent CNVs within multiple samples [20], [21], [22], [23], but these methods rarely considered the CNVs shared by subgroups. Zhang et al. [24], [25] tried to find simultaneous change-points using chi-square statistics and correlation analysis across samples, respectively. Picard et al. [19] extended the dynamic programming method for single aCGH profile segmentation to the multisample case. Recently, Nowak et al. [26] proposed a matrix factorization-based model to explore common CNV patterns among multiple samples. We will discuss this method in detail and compare it to our method in Section 3.

In this paper, we aim to address the problem of identifying CNVs from multisample aCGH data. The main contributions are summarized as follows:

1. We propose a novel framework to denoise multisample aCGH data, which uses both the smoothness property along each sample and the correlation among multiple samples.
2. The problem is formulated as convex optimization and an efficient algorithm is developed to solve the problem exactly.
3. The model naturally handles missing values that usually exist in raw aCGH data.
4. The relationship between our model and other closely related models is discussed.

The MATLAB code of our algorithm is publicly available at http://bioinformatics.ust.hk/tvsp/tvsp.html.

2 Method

2.1 Problem Statement

Let $D \in \mathbb{R}^{m\times n}$ represent an aCGH data set obtained from multiple samples, where m is the number of probes/genes and n is the number of samples. Each entry (i, j) records the $\log_2$-ratio at probe i of sample j and the value of (i, j) is denoted by D_{ij}. The jth column D_j corresponds to an aCGH profile from sample j. We propose to use the following model to describe a given data set

$$D = B + \epsilon. \tag{1}$$

$B \in \mathbb{R}^{m\times n}$ denotes true CNV signals and $\epsilon \in \mathbb{R}^{m\times n}$ is measurement noise. Our goal is to recover B from D.

2.2 Formulation

To make the decomposition in (1) possible, we need some knowledge about the properties of B. Our analysis is based on two assumptions:

- For each sample, the copy numbers at contiguous chromosome positions should be identical except for abrupt changes between different segments, i.e., the signal should be piecewise constant.
- For a set of related samples, the CNV signals are likely to share similar patterns or linearly correlated with each other.

The first assumption is generally required in most methods for aCGH data analysis [15], while the second assumption is the basic

- *X. Zhou and W. Yu are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China. E-mail: eexwzhou@ust.hk.*
- *C. Yang and H. Zhao are with the Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06520.*
- *X. Wan is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China.*

Manuscript received 29 May 2012; revised 21 Nov. 2012; accepted 23 Nov. 2012; published online 14 Dec. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-05-0132. Digital Object Identifier no. 10.1109/TCBB.2012.166.

1545-5963/13/$31.00 © 2013 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

motivation for researchers to analyze aCGH data from multiple samples simultaneously.

Based on the above two assumptions, we propose to estimate B by minimizing the following energy[1]

$$\min_B \frac{1}{2}\|D-B\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|B_j\|_{TV}. \tag{2}$$

Here, we regularize B with the nuclear norm to encode the message that CNV signals from multiple samples should be correlated with each other as much as possible. Recently, the nuclear norm minimization has proven to be an effective method to reconstruct a low-rank matrix [28]. As mentioned in [29], the nuclear norm is not only a convex surrogate of the rank operator but also a good regularization method, which usually achieves better prediction accuracy for model building. The last term in (2) minimizes the total variation of each sample, which encourages each column of B to be piecewise constant.

There usually exist missing values in real aCGH data sets. Suppose Ω is the set of observed entries. To handle missing values, we modify the formulation in (2) to be

$$\min_B \frac{1}{2}\|\mathcal{P}_\Omega(D-B)\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|B_j\|_{TV}, \tag{3}$$

where $\mathcal{P}_\Omega(\cdot)$ represents the projection to the linear space of matrices whose nonzero entries are restricted in Ω

$$\mathcal{P}_\Omega(X)(i,j)=\begin{cases}X_{ij}, & \text{if } (i,j)\in\Omega,\\ 0, & \text{if } (i,j)\notin\Omega.\end{cases} \tag{4}$$

The first term in (3) means that only the observed entries inside Ω is used for model fitting. By this formulation, we can recover the signal and complete the missing values at the same time.

Both (2) and (3) are convex programming. In Section 2.3, we provide efficient algorithms to solve them with guaranteed optimal solutions.

2.3 Algorithms

To solve (3), we first try to solve (2), which is a special case of (3), with Ω being all entries in the matrix.

The optimization in (2) is convex [30], which can be solved using modern convex optimization software like CVX and SeDumi if the problem size is small. However, these generic methods are not scalable to solve large problems. Here, we provide an efficient algorithm to solve (2) exactly based on singular value thresholding (SVT) [31] and the Alternating Direction Method of Multipliers (ADMMs) [32].

First, we introduce a variable Z with the same size of B to separate the nonsmooth terms in (2)

$$\begin{aligned}&\min_{B,Z}\frac{1}{2}\|D-B\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|Z_j\|_{TV},\\ &\text{s.t. } B=Z.\end{aligned} \tag{5}$$

Obviously, the problems in (5) and (2) have the same solution.

We follow the standard procedure of ADMM to solve (5). The augmented Lagrangian of (5) reads

$$\begin{aligned}\mathcal{L}(B,Z,Y)=&\frac{1}{2}\|D-B\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|Z_j\|_{TV}\\ &+\langle Y,B-Z\rangle+\frac{\rho}{2}\|B-Z\|_F^2,\end{aligned} \tag{6}$$

where Y is the dual variable, $\langle\cdot,\cdot\rangle$ means the inner product and ρ is a positive number controlling the step length of updating variables. Then, the following steps are iterated until convergence

$$B^{k+1}=\arg\min_B \mathcal{L}(B,Z^k,Y^k), \tag{7}$$

$$Z^{k+1}=\arg\min_Z \mathcal{L}(B^{k+1},Z,Y^k), \tag{8}$$

$$Y^{k+1}=Y^k+\rho(B^{k+1}-Z^{k+1}). \tag{9}$$

It can be proved that the sequence of B^k generated by the above iterations will converge to the global optimal solution of (5) [32].

Next, we give the solution to each step in ADMM iterations. The B-step in (7) can be written as

$$\begin{aligned}&\min_B \frac{1}{2}\|D-B\|_F^2+\alpha\|B\|_*+\langle Y^k,B-Z^k\rangle+\frac{\rho}{2}\|B-Z^k\|_F^2\\ &=\min_B \frac{1+\rho}{2}\Big\|\frac{D+\rho\left(Z^k-\frac{1}{\rho}Y^k\right)}{1+\rho}-B\Big\|_F^2+\alpha\|B\|_*,\end{aligned} \tag{10}$$

which has the following closed-form solution [31]

$$B^{k+1}=\mathbf{S}_{\frac{\alpha}{1+\rho}}\left(\frac{D+\rho(Z^k-\frac{1}{\rho}Y^k)}{1+\rho}\right), \tag{11}$$

where $\mathbf{S}_\lambda(\cdot)$ means the SVT operator

$$\mathbf{S}_\lambda(X)=U\Sigma_\lambda V^T, \tag{12}$$

$\Sigma_\lambda=\mathrm{diag}[(d_1-\lambda)_+,\ldots,(d_r-\lambda)_+]$, $U\Sigma V^T$ is the SVD of X, $\Sigma=\mathrm{diag}[d_1,\ldots,d_r]$ and $t_+=\max(t,0)$.

The Z-step in (8) can be written as

$$\begin{aligned}&\min_Z\ \gamma\sum_{j=1}^{n}\|\nabla Z_j\|_1+\langle Y^k,B^{k+1}-Z\rangle+\frac{\rho}{2}\|B^{k+1}-Z\|_F^2\\ &=\min_Z\frac{\rho}{2}\|B^{k+1}+\frac{1}{\rho}Y^k-Z\|_F^2+\gamma\sum_{j=1}^{n}\|Z_j\|_{TV}.\end{aligned} \tag{13}$$

Obviously, this minimization can be operated for each column separately

$$Z_j^{k+1}=\arg\min_{\mathbf{x}}\frac{\rho}{2}\|B_j^{k+1}+\frac{1}{\rho}Y_j^k-\mathbf{x}\|_2^2+\gamma\|\mathbf{x}\|_{TV}. \tag{14}$$

The minimization in (14) is the fused lasso signal approximation (FLSA) problem, which can be solved efficiently [33] using existing algorithms.

The overall algorithm to solve (2) is summarized in Algorithm 1. The optimality of the solution can be guaranteed [32]. Please refer to [32] for detailed description on selection of coefficient ρ and the criterion of convergence.

Algorithm 1. The algorithm to solve (2)

1. **Input:** D
2. **Initialize:** $\hat{B}=\mathbf{0}$, $Z=\mathbf{0}$ and $Y=\mathbf{0}$
3. **repeat**
4. $\quad\hat{B}\leftarrow\arg\min_{\hat{B}}\frac{1}{2}\|\frac{D+\rho(Z^k-\frac{1}{\rho}Y^k)}{1+\rho}-\hat{B}\|_F^2+\frac{\alpha}{1+\rho}\|\hat{B}\|_*$
5. $\quad$**for** $j=1$ to n **do**
6. $\quad\quad Z_j\leftarrow\arg\min_{\mathbf{x}}\frac{1}{2}\|\hat{B}_j+\frac{1}{\rho}Y_j-\mathbf{x}\|_2^2+\frac{\gamma}{\rho}\|\mathbf{x}\|_{TV}$
7. $\quad$**end for**
8. $\quad Y\leftarrow Y+\rho(\hat{B}-Z)$

1. The following norms are used throughout this paper: For any vector $\mathbf{x}\in\mathbb{R}^n$, the ℓ_2-norm is defined as $\|\mathbf{x}\|_2=\sqrt{\sum_{i=1}^n x_i^2}$, the ℓ_1-norm is defined as $\|\mathbf{x}\|_1=\sum_{i=1}^n|x_i|$, and the total variation [27] is defined as $\|\mathbf{x}\|_{TV}=\sum_{i=2}^n|x_i-x_{i-1}|$, which measures the smoothness of $\mathbf{x}$. For any matrix $X\in\mathbb{R}^{m\times n}$, the Frobenius norm is defined as $\|X\|_F=\sqrt{\sum_{i=1}^m\sum_{j=1}^n X_{ij}^2}$, and the nuclear norm is defined as $\|X\|_*=\sum_{i=1}^r\sigma_i$, where $\sigma_1,\ldots,\sigma_r$ are the singular values of X and r is the rank of X.

9. **until** convergence
10. **Output:** $\hat{B}$.

Next, we give the algorithm to solve the extended model in (3). First, we define $\pi_\alpha^\gamma(D)$ to be the solution of (2) with given D and fixed α and γ. To solve (3), we first rewrite it as

$$\min_B \frac{1}{2}\|\mathcal{P}_\Omega(D)+\mathcal{P}_{\Omega^\perp}(B)-B\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|B_j\|_{TV}, \quad (15)$$

where $\Omega^\perp$ is the complementary set of Ω. Comparing (15) with (2), we propose to solve the extended model by iteratively updating B using

$$B^{k+1}=\pi_\alpha^\gamma(\mathcal{P}_\Omega(D)+\mathcal{P}_{\Omega^\perp}(B^k)). \quad (16)$$

Theorem 1. *The sequence B^k generated by (16) converges to a limit B^∞ that solves the problem in (3).*

The proof of Theorem 1 is given in the supplementary document, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.166, and the algorithm to solve (3) is summarized in Algorithm 2.

Algorithm 2. The algorithm to solve (3)

1. **Input:** D, the set of observed entries Ω
2. **Initialize:** $\hat{B}=\mathbf{0}$
3. **repeat**
4. $\quad\hat{B}\leftarrow\pi_\alpha^\gamma(\mathcal{P}_\Omega(D)+\mathcal{P}_{\Omega^\perp}(\hat{B}))$
5. **until** convergence
6. **Output:** $\hat{B}$.

2.4 Parameter Tuning

We have two parameters in our model: α controls the nuclear norm of $\hat{B}$ and γ controls the total variation of each $\hat{B}_j$. Here, we propose to choose the parameters by formulating the problem as a matrix completion problem [29] and use the prediction error to guide the parameter selection.

Let Ω_0 be the observed entries in matrix D. We further divide Ω_0 into a training set Ω_1 and a testing set Ω_2. $\Omega_1\cup\Omega_2=\Omega_0$ and $|\Omega_1|/|\Omega_0|=50\%$.[2] First, we use entries in Ω_1 to fit the model by solving

$$\min_B \frac{1}{2}\|\mathcal{P}_{\Omega_1}(D-B)\|_F^2+\alpha\|B\|_*+\gamma\sum_{j=1}^{n}\|B_j\|_{TV}, \quad (17)$$

and denote the solution by $\hat{B}(\alpha,\gamma)$. Then, we evaluate the prediction error over the testing set Ω_2

$$Err(\alpha,\gamma)=\frac{1}{2}\|\mathcal{P}_{\Omega_2}(D-\hat{B}(\alpha,\gamma))\|_F. \quad (18)$$

The problem in (17) is solved multiple times for a grid of (α,γ) values. Finally, we choose $(\hat{\alpha},\hat{\gamma})$ that gives the minimum prediction error as the final parameters, and we run Algorithm 1 again with full observation in Ω_0.

In implementation, since the 2D searching of parameters is computationally expensive, we first search for $\hat{\gamma}$ by fixing $\alpha=0$ and then search for $\hat{\alpha}$ by fixing $\gamma=\hat{\gamma}$. We let $\alpha=c_\alpha\alpha_{max}$ and $\gamma=c_\gamma\gamma_{max}$, where α_{max} and γ_{max} are fixed upper bounds for α and γ, respectively. c_α and c_γ are coefficients selected from $\{0.1,0.2,\ldots,1\}$. In all experiments, we choose $\alpha_{max}=0.2\sqrt{m}\hat{\sigma}$ and $\gamma_{max}=2\hat{\sigma}$ empirically. $\hat{\sigma}$ is the standard deviation of noise in the data set, which can be estimated robustly by the median absolute deviation [34].

2. For a set Ω, $|\Omega|$ means the number of elements in Ω.

Please refer to the online supplemental document for the experiments on stability and effectiveness of our parameter selection method.

2.5 Estimation of FDR

After processing the aCGH data, we use a threshold T to determine whether (i,j) is an abberation or not by comparing $|\hat{B}_{ij}|$ with T. The false discovery rate (FDR) [35] is usually used for statistical assessment of detection accuracy, which is defined as

$$\mathrm{FDR}(T)=\frac{|\mathcal{N}_T|}{|\mathcal{A}_T|}, \quad (19)$$

where $\mathcal{A}_T=\{(i,j):|B_{ij}|>T\}$ is the set of declared abberations and $\mathcal{N}_T$ is the set of false abberations. To estimate the FDR for a given T, $|\mathcal{N}_T|$ needs to be calculated, which is unknown in real experiments. However, it can be roughly approximated by the number of abberations picked from the null data. Since there is no reference data in practice, the null data is usually generated by permutation [10], [26]. More specifically, during the kth permutation, we randomly permute the probe locations for each sample and form a null data set $\overline{D}^{(k)}$. Then, we apply our algorithm on $\overline{D}^{(k)}$ and produce an approximated matrix $\overline{B}^{(k)}$. Hence, $\overline{\mathcal{N}}_T^{(k)}=\{(i,j):|\overline{B}_{ij}^{(k)}|>T\}$ gives an estimate to the number of false detections. To reduce bias, we run K times of permutation, and the FDR for threshold T is estimated by

$$\widehat{\mathrm{FDR}}(T)=\frac{\frac{1}{K}\sum_{k=1}^{K}|\overline{\mathcal{N}}_T^{(k)}|}{|\mathcal{A}_T|}. \quad (20)$$

3 Relation to Other Methods

In this section, we discuss the relationship between our method and two closely related methods for aCGH data analysis.

The first method is fused lasso for signal approximation (FLSA) [10]. Briefly, they process each sample separately. If D_j represents the aCGH profile of sample j, FLSA tries to find a sparse and piecewise-constant vector B_j to approximate D_j by solving

$$\min_{B_j}\|D_j-B_j\|_2^2+\lambda_1\|B_j\|_1+\lambda_2\|B_j\|_{TV}. \quad (21)$$

Comparing (21) and (2), we can find that our model differs from FLSA by replacing the ℓ_1-norm of individual columns with the nuclear norm of the whole matrix. The nuclear norm regularization prefers that the detected CNVs are shared by as many samples as possible. The utilization of information among multiple samples can improve the accuracy of detection.

Another closely related method is named the Fused Lasso Latent Feature Model (FLLat) proposed by Nowak et al. [26]. In this model, each aCGH profile is modeled as a linear combination of some latent features $D=UV+\epsilon$, where $D\in\mathbb{R}^{m\times n}$ is the input data set, $U\in\mathbb{R}^{m\times J}$ is the feature matrix with each column representing a latent feature, $V\in\mathbb{R}^{J\times n}$ is the weight matrix, ϵ denotes noise and J is the predefined number of features. U and V are estimated by solving

$$\begin{aligned}&\min_{U,V}\|D-UV\|_F^2+\lambda_1\sum_{j=1}^{J}\|U_j\|_1+\lambda_2\sum_{j=1}^{J}\|U_j\|_{TV},\\&\text{s.t. }\|V_j\|_2\le 1.\end{aligned} \quad (22)$$

Essentially, the underlying assumptions of our method and FLLat are identical. On the one hand, we can notice that $\mathrm{rank}(UV)=J$, which means that FLLat will output a low-rank matrix if J is relatively small. On the other hand, the smoothness constraint on latent features is equivalent to the smoothness constraint on each profile, which is just a linear combination of

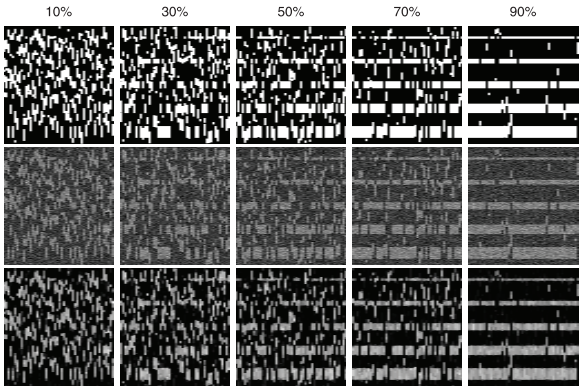

Fig. 1. Examples of synthesized data. Top: data without noise; Middle: data with noise ($\mathrm{SNR} = 2$); Bottom: signals recovered by Algorithm 1. The columns from left to right correspond to various shared percentages.

the features. However, there are differences between our method and FLLat:

1. *Convex* versus *nonconvex*. The formulation in (2) is convex. Hence, a global optimal solution can be guaranteed. While the formulation of FLLat can be solved efficiently, its solution depends on initialization and may get stuck at local optimum.
2. *Nuclear norm* versus *rank operator*. Comparison between the nuclear norm used in our model and the rank operator used in FLLat is analogous to comparison of the ℓ_1-norm versus the ℓ_0-norm used in regression problems [28]. To see this, let $\hat{B}$ be the matrix to be estimated and $\mathbf{w} = [\sigma_1, \ldots, \sigma_r]^T$ be the vector of singular values of $\hat{B}$. Then, in this paper, $\|\hat{B}\|_*$ or $\|\mathbf{w}\|_1$ is minimized, while in FLLat $\mathrm{rank}(\hat{B})$ or $\|\mathbf{w}\|_0$ is fixed to be J. For regression problems, it has been stated that the ℓ_1-norm achieves better consistency of feature selection [36]. Similarly, for matrix learning, the nuclear norm regularization usually performs more stably [29], while the hard constraint on the matrix rank used in FLLat may be too aggressive in selecting the singular vectors (i.e., latent features in FLLat).

4 Results

4.1 Synthesized Data

Synthesized data is generated to test the proposed method. For each data set, 50 samples of aCGH profiles with a length of 500 probes are generated. The intensity at probe i of sample j is given by $D_{ij} = B_{ij} + \epsilon_{ij}$, where B_{ij} is true signal intensity and ϵ_{ij} is noise. We set $B_{ij} = 1$ if (i, j) is located in an abberation segment and $B_{ij} = 0$ otherwise. $\epsilon_{ij} \in \mathcal{N}(0, \sigma)$ and the signal-to-noise ratio (SNR) is defined as $1/\sigma$. Two types of abberation segments are added. The first type is unshared segment that is independently added to each sample at random locations. The second type is shared segment that presents at the same location for multiple samples. The *shared percentage* is defined as the ratio between the number of shared abberation segments and the total number of abberation segments. The length of segments L is selected from $\{10, 20, 30, 40, 50\}$. For each L, 50 segments are added to 50 samples. If the shared percentage is p, we randomly choose $p \times n$ samples to add a segment with length L to each of them at the same probe location. Then, we add a segment with the same length to each of the other $(1 - p) \times n$ samples at random probe locations. Fig. 1 gives illustrative examples. With the shared percentage increasing,

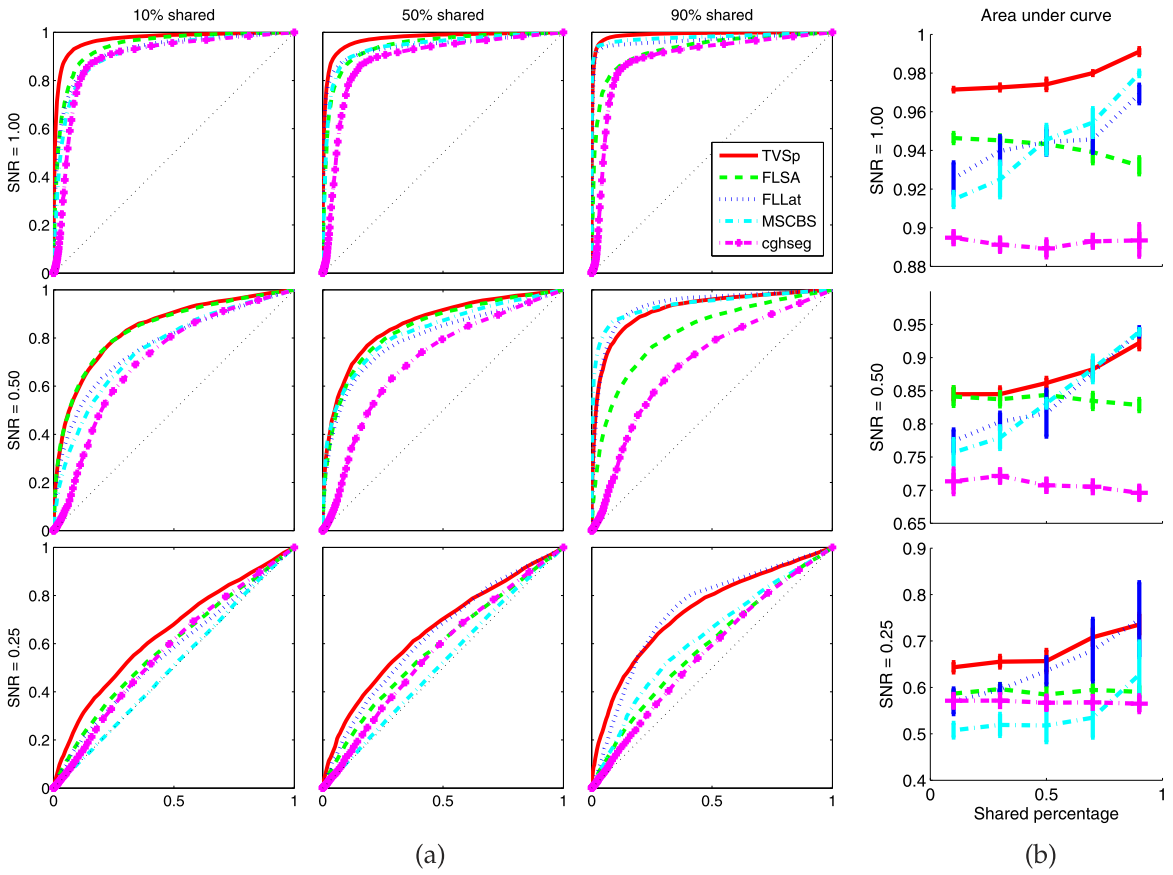

Fig. 2. Performance comparison of our method (*abbr.* TV-Sp), FLSA [10], FLLat [26], MSCBS [24], and cghseg [19] (a) The ROC curves of different methods on synthesized data sets with various shared percentages and SNRs. The y-axis and x-axis of each plot represent the true positive rate and the false positive rate, respectively. (b) The AUC versus the shared percentage. The bar length means the standard deviation.

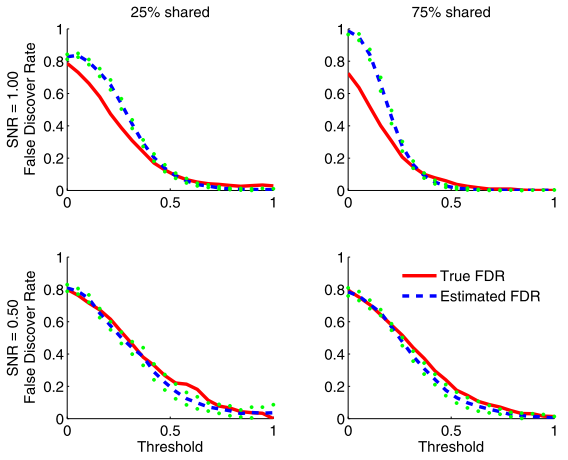

Fig. 3. The comparison between the estimated false discovery rate and the true FDR on synthesized data sets.

the common patterns becomes more and more visible. The results produced by Algorithm 1 are given in the last row. Compared to the raw input in the middle row, random noise is suppressed while the CNV signal is maintained.

The pattern of shared abberations could be more complex than what has been synthesized in our simulation. For example, six possible scenarios of recurrent regions are described in the work by Rueda and Diaz-Uriarte [37]. Our simulation only covers the first two. In practice, our method can be applied to all scenarios since the patterns in these scenarios all admit our assumptions: The abberation region covers a set of contiguous probes and affects a group or a subgroup of samples, regardless of the type of the abberation and whether a region overlaps another or not.

4.2 Performance Comparison

We compare our method with FLSA [10], FLLat [26], MSCBS [24], and cghseg [19] on synthesized data sets. The R packages of all methods are downloaded and the default parameter settings are applied. Fig. 2a plots the receiver operating characteristic (ROC) curves under different cases. A ROC curve deviating more from the diagonal line generally indicates better performance. To better display, we also plot the area under curve (AUC), as shown in Fig. 2b.

Our model consistently outperforms FLSA, especially for large shared percentages. This demonstrates that utilization of multisample information via the nuclear norm regularization can increase the power of detecting common abberations among multiple samples. FLLat performs extremely well when the shared percentage is high, since the data structure almost meets its underlying model. However, when the shared percentage gets lower, the performance of FLLat drops dramatically. This is due to the fact that the hard constraint on the matrix rank used in FLLat is not flexible enough when the low-rank assumption is not rigorously satisfied, which was discussed in Section 3. Also, the variance of AUC for FLLat is relatively large due to its nonconvex formulation. Instead, the nuclear norm regularization and the convex formulation of our method performs consistently well under various cases. Compared with other two alternative methods MSCBS and cghseg, our method also shows better performance.

4.3 FDR Estimation

To verify the reliability of FDR estimation introduced in Section 2.5, we compare the estimated FDR with its true value on synthesized data sets. As shown in Fig. 3, the FDR is overestimated when the value is large, while it is approximated well when the value is small. Generally, the FDR should be controlled under 0.2, where our estimation is fairly accurate.

4.4 Application on Breast Cancer Data

We apply our method on two independent breast cancer data sets. The first one is from Pollack et al. [38], which consists of aCGH profiles over 6,691 mapped human genes for 44 locally advanced primary breast tumors. The other one is from Chin et al. [39], which includes aCGH profiles over 2,149 DNA clones for 141 primary breast tumors.

The results for chromosome 17 are presented in Fig. 4. The recovered matrices are displayed using the heat map and the bar plot at the bottom shows the number of gains summed over all samples for each probe with a threshold equal to 1. The high-amplification regions discovered from the two data sets, i.e., probes 178-184 from Pollack et al. and probes 38-39 from Chin et al., coincide with each other regarding their locations on the chromosome. Several genes that have been verified to be functionally important to breast cancer are located within this region [39], such as the transcription regulation protein PPARBP, the receptor tyrosine kinase ERBB2 and the adaptor protein GRB7. Fig. 4c shows the log_2-ratio of a selected sample from Pollack et al. before and after processing. Compared with FLLat, our method gives a more smooth profile while keeping more candidate signals unsuppressed, e.g., the amplifications around probe 321, which were also reported by [38] with high-elevated mRNA levels.

Further downstream analysis can be carried out on our results, e.g., identifying disease-related CNVs by existing tools such as CanPredict [40] or using gene expression data [41].

4.5 Computational Time

The time cost of our algorithm is mainly from the singular value decomposition (SVD) and FLSA computed in each iteration. For a matrix $X \in \mathbb{R}^{m \times n}$, the complexity of SVD is $\mathcal{O}(mn^2)$ if $n \ll m$. For a vector $x \in \mathbb{R}^m$, the complexity of current FLSA algorithms is $\mathcal{O}(m)$ empirically [33]. Overall, the complexity of our algorithm is $\mathcal{O}(mn^2)$ for limited iterations. Fig. 5 shows the real CPU time of our

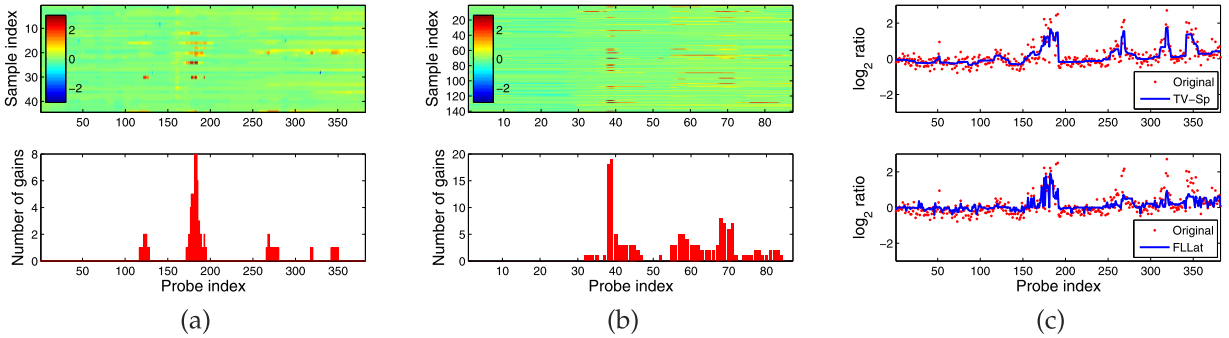

Fig. 4. (a) The results of applying our method to the data set from Pollack et al. [38]. From top to bottom are the 2D images of recovered signals and the number of detected gains along the chromosome. (b) The results on the data set from Chin et al. [39]. (c) The 1D profile of a selected sample of the data set from Pollack et al. Our result is shown in the top panel. The FLLat result is shown in the bottom panel.

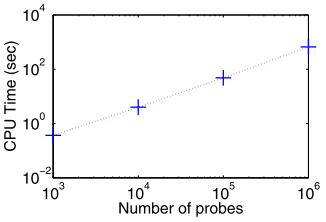

Fig. 5. Time cost versus number of probes on synthesized data sets. The number of samples is 50.

algorithm to solve synthesized problems with different numbers of probes. Here, parameter α and γ are fixed. The algorithm is run on a desktop PC with a 3.4-GHz Intel i7 CPU and 8-GB RAM. As we can see, the computational time increases almost linearly over the number of probes. In aCGH experiments, the number of probes is always much larger than the number of samples, and our algorithm shows great scalability to process large data sets, e.g., the data from the next generation of microarrays. Specifically, the computational time is 516 s for $(m, n) = (10^6, 50)$, while those of FLLat and FLSA are 1,932 and 39 s, respectively.

5 Conclusion

In this paper, we propose a convex formulation for analyzing multisample aCGH data. There are two major advantages to formulate the problem as convex optimization. First, the global optimal solution is guaranteed, which makes the method perform stable in various problems. Second, a very efficient and scalable algorithm can be developed based on modern convex optimization techniques. Moreover, we explain the relationship between our model and two closely related models. The experiments demonstrate that our method is competitive to the state-of-the-art approaches and more robust under various cases.

Acknowledgments

This work was partially supported by the National Institutes of Health grant R01 GM 59507.

References

[1] R. Redon et al., "Global Variation in Copy Number in the Human Genome," *Nature*, vol. 444, no. 7118, pp. 444-454, 2006.
[2] L. Feuk, A. Carson, and S. Scherer, "Structural Variation in the Human Genome," *Nature Rev. Genetics*, vol. 7, no. 2, pp. 85-97, 2006.
[3] E. Gonzalez et al., "The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility," *Science*, vol. 307, no. 5714, pp. 1434-1440, 2005.
[4] D. Pinkel and D. Albertson, "Array Comparative Genomic Hybridization and Its Applications in Cancer," *Nature Genetics*, vol. 37, pp. S11-S17, 2005.
[5] A. Olshen, E. Venkatraman, R. Lucito, and M. Wigler, "Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data," *Biostatistics*, vol. 5, no. 4, pp. 557-572, 2004.
[6] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. Daudin, "A Statistical Approach for Array CGH Data Analysis," *BMC Bioinformatics*, vol. 6, no. 1, article 27, 2005.
[7] P. Rancoita, M. Hutter, F. Bertoni, and I. Kwee, "Bayesian DNA Copy Number Analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 10, 2009.
[8] P. Hupé, N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot, "Analysis of Array CGH Data: From Signal Ratio to Gain and Loss of DNA Regions," *Bioinformatics*, vol. 20, no. 18, pp. 3413-3422, 2004.
[9] E. Ben-Yaacov and Y. Eldar, "A Fast and Flexible Method for the Segmentation of aCGH Data," *Bioinformatics*, vol. 24, no. 16, pp. i139-i145, 2008.
[10] R. Tibshirani and P. Wang, "Spatial Smoothing and Hot Spot Detection for CGH Data Using the Fused Lasso," *Biostatistics*, vol. 9, no. 1, pp. 18-29, 2008.
[11] J. Marioni, N. Thorne, and S. Tavaré, "BioHMM: A Heterogeneous Hidden Markov Model for Segmenting Array CGH Data," *Bioinformatics*, vol. 22, no. 9, pp. 1144-1146, 2006.
[12] S. Stjernqvist, T. Rydén, M. Sköld, and J. Staaf, "Continuous-Index Hidden Markov Modelling of Array CGH Copy Number Data," *Bioinformatics*, vol. 23, no. 8, pp. 1006-1014, 2007.
[13] B. Nilsson, M. Johansson, F. Al-Shahrour, A. Carpenter, and B. Ebert, "Ultrasome: Efficient Aberration Caller for Copy Number Studies of Ultra-High Resolution," *Bioinformatics*, vol. 25, no. 8, pp. 1078-1079, 2009.
[14] S. Morganella, L. Cerulo, G. Viglietto, and M. Ceccarelli, "VEGA: Variational Segmentation for Copy Number Detection," *Bioinformatics*, vol. 26, no. 24, pp. 3020-3027, 2010.
[15] W. Lai, M. Johnson, R. Kucherlapati, and P. Park, "Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data," *Bioinformatics*, vol. 21, no. 19, pp. 3763-3770, 2005.
[16] H. Willenbrock and J. Fridlyand, "A Comparison Study: Applying Segmentation to Array CGH Data for Downstream Analyses," *Bioinformatics*, vol. 21, no. 22, pp. 4084-4091, 2005.
[17] R. Pique-Regi, A. Ortega, and S. Asgharzadeh, "Joint Estimation of Copy Number Variation and Reference Intensities on Multiple DNA Arrays Using GADA," *Bioinformatics*, vol. 25, no. 10, pp. 1223-1230, 2009.
[18] M. Van De Wiel, R. Brosens, P. Eilers, C. Kumps, G. Meijer, B. Menten, E. Sistermans, F. Speleman, M. Timmerman, and B. Ylstra, "Smoothing Waves in Array CGH Tumor Profiles," *Bioinformatics*, vol. 25, no. 9, pp. 1099-1104, 2009.
[19] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin, "Joint Segmentation, Calling, and Normalization of Multiple CGH Profiles," *Biostatistics*, vol. 12, no. 3, pp. 413-428, 2011.
[20] S. Diskin, T. Eck, J. Greshock, Y. Mosse, T. Naylor, C. Stoeckert Jr., B. Weber, J. Maris, and G. Grant, "STAC: A Method for Testing the Significance of DNA Copy Number Aberrations Across Multiple Array-CGH Experiments," *Genome Research*, vol. 16, no. 9, pp. 1149-1158, 2006.
[21] M. Guttman, C. Mies, K. Dudycz-Sulicz, S. Diskin, D. Baldwin, C. Stoeckert, and G. Grant, "Assessing the Significance of Conserved Genomic Aberrations Using High Resolution Genomic Microarrays," *PLoS Genetics*, vol. 3, no. 8, p. e143, 2007.
[22] R. Beroukhim et al., "Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma," *Proc. Nat'l Academy of Sciences of USA*, vol. 104, no. 50, p. 20007, 2007.
[23] S. Shah, W. Lam, R. Ng, and K. Murphy, "Modeling Recurrent DNA Copy Number Alterations in Array CGH Data," *Bioinformatics*, vol. 23, no. 13, pp. i450-i458, 2007.
[24] N. Zhang, D. Siegmund, H. Ji, and J. Li, "Detecting Simultaneous Changepoints in Multiple Sequences," *Biometrika*, vol. 97, no. 3, pp. 631-645, 2010.
[25] Q. Zhang et al., "Cmds: A Population-Based Method for Identifying Recurrent DNA Copy Number Aberrations in Cancer from High-Resolution Data," *Bioinformatics*, vol. 26, no. 4, pp. 464-469, 2010.
[26] G. Nowak, T. Hastie, J. Pollack, and R. Tibshirani, "A Fused Lasso Latent Feature Model for Analyzing Multi-Sample a CGH Data," *Biostatistics*, vol. 12, no. 4, pp. 776-791, 2011.
[27] A. Rinaldo, "Properties and Refinements of the Fused Lasso," *Annals of Statistics*, vol. 37, no. 5B, pp. 2922-2952, 2009.
[28] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471-501, 2010.
[29] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," *J. Machine Learning Research*, vol. 11, pp. 2287-2322, 2010.
[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.
[31] J. Cai, E. Candès, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM J. Optimization*, vol. 20, pp. 1956-1982, 2010.
[32] S. Boyd, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2010.
[33] J. Liu, L. Yuan, and J. Ye, "An Efficient Algorithm for a Class of Fused Lasso Problems," *Proc. ACM SIGKDD 16th Int'l Conf. Knowledge Discovery and Data Mining*, pp. 323-332, 2010.
[34] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision*, vol. 6, no. 1, pp. 59-70, 1991.
[35] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. Royal Statistical Soc., Series B (Methodological)*, vol. 57, pp. 289-300, 1995.
[36] P. Zhao and B. Yu, "On Model Selection Consistency of Lasso," *J. Mach. Learn. Res*, vol. 7, pp. 2541-2563, 2006.
[37] O. Rueda and R. Diaz-Uriarte, "Finding Recurrent Copy Number Alteration Regions: A Review of Methods," *Current Bioinformatics*, vol. 5, no. 1, pp. 1-17, 2010.
[38] J. Pollack, T. Sørlie, C. Perou, C. Rees, S. Jeffrey, P. Lonning, R. Tibshirani, D. Botstein, A. Børresen-Dale, and P. Brown, "Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors," *Proc. Nat'l Academy of Sciences USA*, vol. 99, no. 20, pp. 12963-12968, 2002.
[39] K. Chin et al., "Genomic and Transcriptional Aberrations Linked to Breast Cancer Pathophysiologies," *Cancer Cell*, vol. 10, no. 6, pp. 529-541, 2006.
[40] J. Kaminker, Y. Zhang, C. Watanabe, and Z. Zhang, "CanPredict: A Computational Tool for Predicting Cancer-Associated Missense Mutations," *Nucleic Acids Research*, vol. 35, no. suppl 2, pp. W595-W598, 2007.
[41] L. Tran, B. Zhang, Z. Zhang, C. Zhang, T. Xie, J. Lamb, H. Dai, E. Schadt, and J. Zhu, "Inferring Causal Genomic Alterations in Breast Cancer Using Gene Expression Data," *BMC Systems Biology*, vol. 5, no. 1, article 121, 2011.