BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 14 2014, pages 1943–1949
doi:10.1093/bioinformatics/btu131

Genome analysis

Advance Access publication March 17, 2014

Piecewise-constant and low-rank approximation for identification of recurrent copy number variations

Xiaowei Zhou[1], Jiming Liu[2], Xiang Wan[2,*] and Weichuan Yu[1,*]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon and [2]Department of Computer Science and Institute of Theoretical and Computational Study, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

Associate Editor: John Hancock

ABSTRACT

Motivation: The post-genome era sees urgent need for more novel approaches to extracting useful information from the huge amount of genetic data. The identification of recurrent copy number variations (CNVs) from array-based comparative genomic hybridization (aCGH) data can help understand complex diseases, such as cancer. Most of the previous computational methods focused on single-sample analysis or statistical testing based on the results of single-sample analysis. Finding recurrent CNVs from multi-sample data remains a challenging topic worth further study.

Results: We present a general and robust method to identify recurrent CNVs from multi-sample aCGH profiles. We express the raw dataset as a matrix and demonstrate that recurrent CNVs will form a low-rank matrix. Hence, we formulate the problem as a matrix recovering problem, where we aim to find a piecewise-constant and low-rank approximation (PLA) to the input matrix. We propose a convex formulation for matrix recovery and an efficient algorithm to globally solve the problem. We demonstrate the advantages of PLA compared with alternative methods using synthesized datasets and two breast cancer datasets. The experimental results show that PLA can successfully reconstruct the recurrent CNV patterns from raw data and achieve better performance compared with alternative methods under a wide range of scenarios.

Availability and implementation: The MATLAB code is available at http://bioinformatics.ust.hk/pla.zip.

Contact: xwan@comp.hkbu.edu.hk or eeyu@ust.hk

Received on December 10, 2013; revised on February 24, 2014; accepted on March 2, 2014

1 INTRODUCTION

Copy number variations (CNVs) are genomic alterations characterized as abnormal number of copies in one or more segments of DNA. The size of segments varies from 1 kb to 3 Mb (Hastings *et al*., 2009). Studies have indicated that CNVs are associated with human diseases. For instance, Gonzalez *et al*. (2005) found that individuals who carried fewer copies of gene CCL3L1 than average were significantly more susceptible to HIV. Each missing copy increases the susceptibility by 4.5–10.5%. Several other types of human diseases, such as cancers and autoimmune diseases, have also been linked with CNVs (Lee and Lupski, 2006; Lee *et al*., 2007; Lupski, 2007). Therefore, it is important to investigate the contribution of CNV to complex diseases.

Array-based comparative genomic hybridization (aCGH) is a high-throughput and high-resolution approach for measuring changes of copy numbers in thousands of DNA regions (Pinkel and Albertson, 2005). In a typical aCGH experiment, genomic DNAs are extracted from test samples and reference samples and differentially labeled with two dyes. The labeled DNAs are mixed together and hybridized to a microarray spotted with DNA probes. The ratio between the fluorescence intensity of the test DNA and that of the reference DNA at a given probe measures the ratio between the copy number in the test genome and that in the reference genome. Data extracted from an aCGH experiment are generally in the form of $\log_2$ ratios. A value greater than zero indicates a possible gain in the copy number, while a value less than zero indicates a possible loss.

Detection of CNVs from aCGH data is to locate change-points in $\log_2$-ratio profiles that partition each chromosome into discrete segments. Because of the high noise level in the intensity values of aCGH data, it is difficult to identify the boundaries of CNVs using the typical thresholding approach (Pinkel and Albertson, 2005). Many methods have been developed to analyze single-sample aCGH data, including breakpoint detection (Olshen *et al*., 2004; Picard *et al*., 2005; Rancoita *et al*., 2009), signal smoothing (Ben-Yaacov and Eldar, 2008; Hupé *et al*., 2004; Tibshirani and Wang, 2008) and Hidden Markov Models (Marioni *et al*., 2006; Stjernqvist *et al*., 2007). Reviews and comparisons can be found in Lai *et al*. (2005) and Willenbrock and Fridlyand (2005). However, finding CNVs from single samples is only the initial step in the search of disease-associated genes. It has been pointed out that the recurrent CNVs (CNVs at the same genomic locations appearing frequently over multiple individuals) are more likely to encompass disease-critical genes (Beroukhim *et al*., 2007; Rueda and Diaz-Uriarte, 2010).

Recently, an increasing number of methods have become available for finding recurrent CNV regions from aCGH data of multiple samples. A recurrent CNV region is often defined as a set of consecutive probes that are altered in a group of samples (Rueda and Diaz-Uriarte, 2010). Some methods identify recurrent CNV regions via hypothesis testing by comparing the frequency of alternations at each probe with a null distribution. Examples include Beroukhim *et al*. (2007), Diskin *et al*. (2006), Guttman *et al*. (2007). These methods rely on the preprocessing

*To whom correspondence should be addressed.

of single-sample profiles, including smoothing, segmentation and CNV calling, which may miss some weak but common variations (Shah *et al.*, 2007). Some other methods attempt to detect recurrent CNV regions directly from raw profiles through multiple-chain Hidden Markov Models (Shah *et al.*, 2007), joint segmentation (Picard *et al.*, 2011; Zhang *et al.*, 2010a, b) and matrix factorization methods (Nowak *et al.*, 2011; Zhou *et al.*, 2013). Most of these methods assume a generative model that describes the characteristics of recurrent variations and fit the model under the assumption of Gaussian noise. However, these methods rarely consider the influence of individual-specific variations, which typically have large intensity values and cannot be modeled as Gaussian noise. As a result, individual-specific variations will be mixed with recurrent variations and eventually corrupt the model fitting.

In this article, we develop a robust and efficient method to identify recurrent CNVs from raw aCGH profiles. The main contributions are summarized as follows:

- We formulate the problem as a matrix decomposition problem, where the raw data matrix is decomposed into a low-rank component, a sparse component and a noise component. These three components correspond to recurrent CNVs, individual-specific variations and random noises, respectively. The recurrent CNVs can be easily identified from the low-rank component by thresholding or more sophisticated statistical analysis.
- The proposed formulation is convex. An efficient and scalable algorithm is developed to find the globally optimal solution based on the state-of-art convex optimization techniques.
- The relationship between our model and other related models is explained via simulation of six popular scenarios.

The rest of this article is organized as follows. In Section 2, we introduce our model, formulation and algorithm. In Section 3, we report experimental results. Finally, we conclude our article with some discussions in Section 4.

2 METHODS

2.1 Formulation

In Rueda and Diaz-Uriarte (2010), six popular recurrent CNV scenarios are defined in detail. Our method for detecting recurrent CNVs is motivated by the low-rank property of these scenarios as illustrated in Figure 1. Basically, the number of recurrent CNV regions determines the rank of the matrix composed of multi-sample profiles. In Scenarios (1), (3) and (5), for example, the rank of the matrix equals to 1 if we assume the matrix values are $+1$, -1 and 0 for the probes marked as gain, loss and no alteration, respectively. In Scenarios (2) and (6), two recurrent regions exist and the rank is 2. In Scenario (4), the pattern is more complex and the rank is 3. Therefore, the problem of identifying recurrent CNVs can be treated as a problem of recovering a low-rank matrix from input data.

Mathematically, we express each dataset by a matrix $\mathbf{D} \in R^{n\times p}$, where each row $\mathbf{d}_i$ is a $\log_2$-ratio profile of one sample, and n and p are the numbers of samples and probes, respectively. Our task is to recover a low-rank component $\mathbf{X}$ from $\mathbf{D}$, such that rank($\mathbf{X}$) is small. The typical approach for low-rank approximation is to compute the singular value decomposition of the input matrix and form a new low-rank matrix using

Fig. 1. Six popular scenarios of recurrent CNV regions defined in Rueda and Diaz-Uriarte (2010). The vertical axis indicates the sample index and the horizontal axis indicates the probe index. If the intensity ratio is piecewise constant along the probe axis for each sample, the intensity ratios of the entire dataset can be approximated by a low-rank matrix

the first few singular vectors (Eckart and Young, 1936). This procedure is also named principal component analysis in statistics. The drawback of traditional principal component analysis is that it is only optimal under the assumption of Gaussian noise. However, aCGH data have large intensity values, which cannot be modeled as Gaussian noise. These individual-specific intensity values will severely corrupt the low-rank approximation and make the fitted model to deviate far away from the true model.

Inspired by Candès *et al.* (2011), we propose to use the following decomposition model to detect the recurrent CNVs from noisy input:

$$\mathbf{D} = \mathbf{X} + \mathbf{E} + \epsilon \tag{1}$$

where $\mathbf{X}$ is a low-rank component, $\mathbf{E}$ is a sparse component and ε is a noise component. In aCGH data analysis, the low-rank component corresponds to the recurrent CNVs. The sparse component corresponds to individual-specific CNVs or gross measurement errors that sparsely appear at different locations for different samples. The noise component corresponds to the small perturbation of the intensity value at each probe, which is often modeled by i.i.d. Gaussian distribution with a zero mean.

To achieve the decomposition, the following minimization problem is considered:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E},\epsilon} \frac{1}{2}||\epsilon||_F^2 + \alpha\,\mathrm{rank}(\mathbf{X}) + \beta||\mathbf{E}||_0 \\ &\text{s.t. } \mathbf{D}=\mathbf{X}+\mathbf{E}+\epsilon \end{aligned} \tag{2}$$

where $|\epsilon||_F = \sqrt{\sum_{i,j}\epsilon_{ij}^2}$ is the Frobenious norm and $|\mathbf{E}||_0$ is the ℓ_0-norm that counts the number of non-zero values in $\mathbf{E}$. The solution to (2) will give a penalized maximum likelihood estimate with respect to the variables $\mathbf{X}, \mathbf{E}, \epsilon$.

The proposed model in (2) is intractable because both the rank operator and the ℓ_0-norm are combinatorial operators, which make (2) a NP-hard problem. The traditional work-around for such problems is to use the convex relaxation. Therefore, we replace the rank() by the nuclear norm, which is defined as $|\mathbf{X}||_* = \sum_{i=1}^{r}\sigma_i$, where $\sigma_1, \cdots, \sigma_r$ are the

singular values of $\mathbf{X}$. It is the tightest convex surrogate to the rank operator (Fazel, 2002) and has been widely used for low-rank matrix recovery (Candès *et al.*, 2011). We also replace the ℓ_0-norm by the ℓ_1-norm. The ℓ_1-norm is defined as $|\mathbf{X}||_1 = \sum_{i,j} X_{ij}$. The ℓ_1 relaxation has proven to be a powerful technique for sparse signal recovery (Tropp, 2006).

In addition, we like to introduce a smoothness penalty on each row of $\mathbf{X}$ to reflect the prior that the recovered profile should be piecewise constant. The total-variation norm, which is defined as $||\mathbf{x}||_{\mathrm{TV}} = \sum_{i=2}^{p} |x_i - x_{i-1}|$, is adopted in the proposed model.

Finally, the problem to be solved reads as follows:

$$\min_{\mathbf{X},\mathbf{E}} \frac{1}{2}||\mathbf{D}-\mathbf{X}-\mathbf{E}||_F^2 + \alpha_1||\mathbf{X}||_* + \alpha_2 \sum_{i=1}^{n} ||\mathbf{x}_i||_{\mathrm{TV}} + \beta||\mathbf{E}||_1 \tag{3}$$

where $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$.

The equality constraint in (2) is eliminated by replacing ε with $\mathbf{D}-\mathbf{X}-\mathbf{E}$. The optimization in (3) is convex. Thus, the global optimal solution can be found. As the recovered $\mathbf{X}$ is the piecewise-constant and low-rank approximation of $\mathbf{D}$, we name our model PLA. We will introduce the algorithm in the next subsection.

2.2 Algorithm

Although the problem in (3) is convex, it cannot be solved directly using generic convex optimization software such as CVX (Grant and Boyd, 2008) because of the large size of our problem. In this article, we propose an efficient and scalable algorithm based on the alternating direction method of multipliers (Boyd, 2010).

The first step is to separate the two non-smooth functions of $\mathbf{X}$ by introducing an auxiliary variable $\mathbf{Z}$ and rewrite (3) as follows:

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E},\mathbf{Z}} \frac{1}{2}||\mathbf{D}-\mathbf{X}-\mathbf{E}||_F^2 + \alpha_1||\mathbf{X}||_* + \alpha_2 \sum_{i=1}^{n} ||\mathbf{z}_i||_{\mathrm{TV}} + \beta||\mathbf{E}||_1 \\ &\text{s.t. } \mathbf{X}=\mathbf{Z} \end{aligned} \tag{4}$$

Next, the augmented lagrangian is introduced to eliminate the equality constraint in (4), which reads

$$\begin{aligned} \mathcal{L}(\mathbf{X},\mathbf{Z},\mathbf{Y}) = &\frac{1}{2}||\mathbf{D}-\mathbf{X}-\mathbf{E}||_F^2 + \alpha_1||\mathbf{X}||_* + \alpha_2 \sum_{i=1}^{n} ||\mathbf{z}_i||_{\mathrm{TV}} \\ &+ \beta||\mathbf{E}||_1 + <\mathbf{Y},\mathbf{X}-\mathbf{Z}> + \frac{\rho}{2}||\mathbf{X}-\mathbf{Z}||_F^2 \end{aligned} \tag{5}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, Y is the dual variable and ρ is an adaptively tuned parameter that controls the convergence of the algorithm. The final result of optimization will not be affected by ρ if it is chosen properly. Please refer to (Boyd, 2010) for more details.

To solve (4), the following updating steps are alternated until convergence

$$\widehat{\mathbf{X}} \leftarrow \arg\min_{\mathbf{X}} \mathcal{L}(\mathbf{X},\widehat{\mathbf{Z}},\widehat{\mathbf{E}},\widehat{\mathbf{Y}}) \tag{6}$$

$$\widehat{\mathbf{Z}} \leftarrow \arg\min_{\mathbf{Z}} \mathcal{L}(\widehat{\mathbf{X}},\mathbf{Z},\widehat{\mathbf{E}},\widehat{\mathbf{Y}}) \tag{7}$$

$$\widehat{\mathbf{E}} \leftarrow \arg\min_{\mathbf{E}} \mathcal{L}(\widehat{\mathbf{X}},\widehat{\mathbf{Z}},\mathbf{E},\widehat{\mathbf{Y}}) \tag{8}$$

$$\widehat{\mathbf{Y}} \leftarrow \widehat{\mathbf{Y}} + \rho(\widehat{\mathbf{X}}-\widehat{\mathbf{Z}}) \tag{9}$$

The iteration of above steps will finally converge to the solution of problem (4). The theoretical proof for the convergence can be found in (Boyd, 2010).

The remaining issue is how to solve the steps in (6) to (8). The problem in (6) can be reduced to

$$\min_{\mathbf{X}} \frac{1+\rho}{2}\left\|\frac{\mathbf{D}-\widehat{\mathbf{E}}+\rho(\widehat{\mathbf{Z}}-\frac{1}{\rho}\widehat{\mathbf{Y}})}{1+\rho}-\mathbf{X}\right\|_F^2 + \alpha_1||\mathbf{X}||_* \tag{10}$$

that becomes a nuclear-norm regularized least-squares problem and has the following closed-form solution (Cai *et al.*, 2010)

$$\widehat{\mathbf{X}} = \mathcal{D}_{\frac{\alpha_1}{1+\rho}}\left(\frac{\mathbf{D}-\widehat{\mathbf{E}}+\rho(\widehat{\mathbf{Z}}-\frac{1}{\rho}\widehat{\mathbf{Y}})}{1+\rho}\right) \tag{11}$$

where $\mathcal{D}_\lambda$ refers to the singular value thresholding (SVT)

$$\mathcal{D}_\lambda(\mathbf{M}) = \sum_{i=1}^{r} (\sigma_i - \lambda)_+ \mathbf{u}_i \mathbf{v}_i^T \tag{12}$$

Here, $(x)_+ = \max(x, 0)$. $\{\mathbf{u}_i\}$, $\{\mathbf{v}_i\}$ and $\{\sigma_i\}$ are the left singular vectors, the right singular vectors and the singular values of $\mathbf{M}$, respectively.

The problem in (7) can be rewritten as follows:

$$\min_{\mathbf{Z}} \frac{1}{2}||\widehat{\mathbf{X}}+\frac{1}{\rho}\widehat{\mathbf{Y}}-\mathbf{Z}||_F^2 + \frac{\alpha_2}{\rho}\sum_{i=1}^{n}||\mathbf{z}_i||_{\mathrm{TV}} \tag{13}$$

Apparently, each row of $\mathbf{Z}$ can be updated separately

$$\widehat{\mathbf{z}}_i = \arg\min_{\mathbf{z}} \frac{1}{2}||\widehat{\mathbf{b}}_i+\frac{1}{\rho}\widehat{\mathbf{y}}_i-\mathbf{z}||_2^2 + \frac{\alpha_2}{\rho}||\mathbf{z}||_{\mathrm{TV}} \tag{14}$$

Problem (14) is the fused lasso signal approximation problem that can be solved efficiently (Liu *et al.*, 2010).

The problem in (8) can be rewritten as follows:

$$\min_{\mathbf{E}} \frac{1}{2}||\mathbf{D}-\widehat{\mathbf{X}}-\mathbf{E}||_F^2 + \beta||\mathbf{E}||_1 \tag{15}$$

It admits a closed-form solution

$$\widehat{\mathbf{E}} = \mathcal{S}_\beta\left(\mathbf{D}-\widehat{\mathbf{X}}\right), \tag{16}$$

where $\mathcal{S}_\beta(\mathbf{M})_{ij} = \mathrm{sign}(M_{ij})(M_{ij}-\beta)_+$ refers to the elementwise soft-thresholding operator (Boyd, 2010).

Overall, the algorithm to optimize the proposed model in (3) is summarized in **Algorithm 1**. The convex program will give a global optimal solution independent of initialization.

Algorithm 1: The algorithm to solve PLA in (3).

(1) **Input: D**
(2) Initialize all variables to be zero.
(3) **repeat**
(4) Update $\mathbf{X}$ by solving (10) via singular value thresholding.
(5) Update $\mathbf{Z}$ by solving (13) via fused lasso solvers.
(6) Update $\mathbf{E}$ by solving (15) via soft thresholding.
(7) Update dual variable $\mathbf{Y}$ according to (9).
(8) **until** convergence
(9) **Output: X** and **E**

2.3 Parameter selection

We select the parameters α_1, α_2 and β in our model (3) based on the analysis in two similar models (Candès *et al.*, 2011; Zhou *et al.*, 2010). The proper values should depend on the size of the input matrix (n, p) and the standard variation of the noise σ.

The relative weight $\lambda = \beta/\alpha_1$ balances the two terms in $\alpha_1||\mathbf{X}||_* + \beta||\mathbf{E}||_1$ and consequently controls the rank of $\mathbf{X}$. The model in (Candès *et al.*, 2011) only includes these two terms. Candès *et al.* (2011) has proved that $\lambda = 1/\sqrt{m}$ gives a large probability of recovering $\mathbf{X}$ and $\mathbf{E}$ under their assumed conditions and stated that this value can be

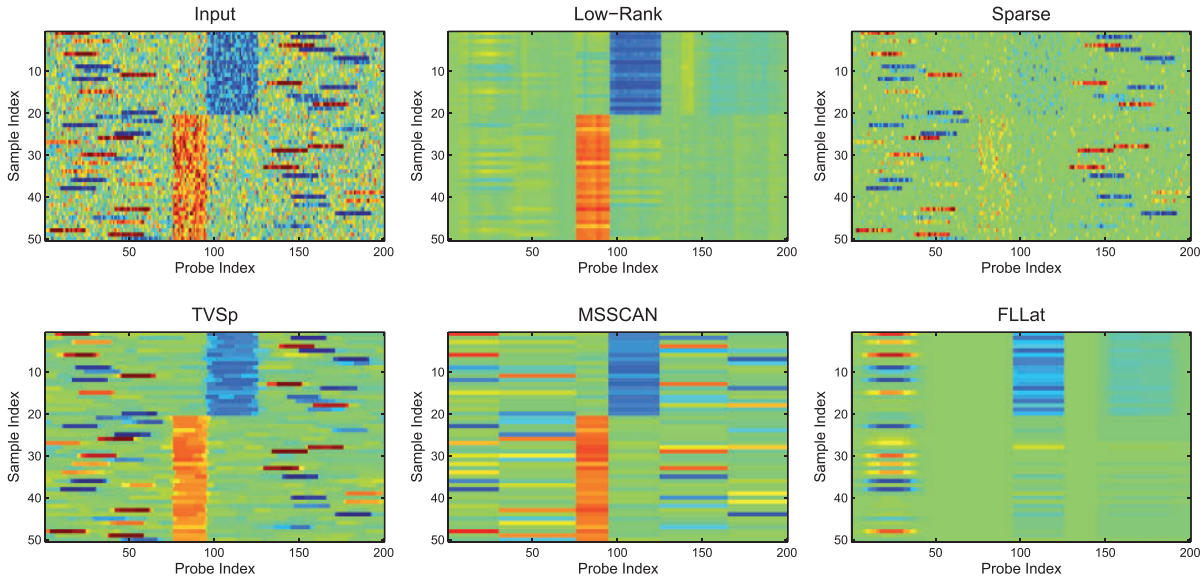

Fig. 2. An illustration of simulated experiments. Top row: from left to right are the synthetic dataset of Scenario 2, the low-rank component and the sparse component recovered by PLA, respectively. Bottom row: the results of three alternative methods TVSp (Zhou *et al.*, 2013), MSSCAN (Zhang *et al.*, 2010a) and FLLat (Nowak *et al.*, 2011)

adjusted slightly to obtain the best results in specific applications. m is the larger dimension of the input matrix. In our problem, $m = p$, i.e. the number of probes. In the synthesized experiments, we simply set $\beta = \alpha_1/\sqrt{p}$. On real datasets, the recurrent CNVs rarely form a perfectly low-rank matrix, and we use $\beta = 2\alpha_1/\sqrt{p}$ to keep sufficient variations in $\mathbf{X}$.

The parameter α_1 serves as a threshold in the SVT step in (11). It should be large enough to threshold out the noise but not too large to over-shrink the signal (Zhou *et al.*, 2010). A proper value is $\alpha_1 = (\sqrt{n} + \sqrt{p})\sigma$, which is the expected ℓ_2-norm of a $n \times p$ random matrix with entries sampled from $\mathcal{N}(0, \sigma^2)$. As CNVs are sparse in the data, we can estimate σ from the data by the median-absolute-deviation estimator (Meer *et al.*, 1991)

$$\hat{\sigma} = 1.48 \text{ median}\{|\mathbf{D} - \text{median}(\mathbf{D})|\} \tag{17}$$

As α_2 only controls the smoothness of the recovered profiles, we empirically set $\alpha_2 = 0.01\alpha_1$.

3 RESULTS

3.1 Synthetic datasets

3.1.1 Accuracy comparison In Rueda and Diaz-Uriarte (2010), six scenarios of recurrent CNVs are discussed (illustrated in Fig. 1). We adopt these six scenarios to generate synthetic data. For each scenario, 50 samples of aCGH profiles with a length of 200 probes are generated. The default signal value for no variation is set as 0. The recurrent variations are located in the interval from Probe 76 to Probe 125 with the patterns identical to the six scenarios given in Figure 1. The $\log_2$ ratio is 1 for a gain and -1 for a loss. For each sample, an individual-specific variation with a length of 20 probes is added at a random location that does not overlap with the recurrent region. The $\log_2$ ratio of each individual-specific variation is randomly sampled from $\{-2, -1, 1, 2\}$. Finally, we add i.d.d Gaussian noises to all probes.

The synthetic dataset of Scenario 2 is illustrated in the first panel of Figure 2. Each row of the matrix is a profile of a sample. There are two recurrent variation regions. The first one is located at Probes $76 \sim 95$, where 30 samples have gains. The second one is located at Probes $96 \sim 125$, where 20 samples have losses. The task is to recover the recurrent patterns from this matrix.

Our results are given in the second and third images in Figure 2. The recurrent CNVs are clearly presented in the low-rank component, while the individual-specific CNVs are mostly included in the sparse component. The rank of the recovered low-rank component is 6, which is slightly larger than the truth. We used the default parameter setting without any prior or tuning. The results of three closely related methods are given in the bottom row. TVSp (Zhou *et al.*, 2013) aims to detect all CNVs. Therefore, it cannot separate recurrent variations from individual-specific variations. MSSCAN (Zhang *et al.*, 2010a) attempts to find the common change points across samples and uses the mean intensity in each segment to reconstruct the profiles. Consequently, the segmentation process is largely influenced by the individual-specific variations resulting in unfaithful reconstruction of recurrent patterns. FLLat (Nowak *et al.*, 2011) tries to recover the common features across profiles. However, the recovery is corrupted because FLLat adopts a least-squares loss to fit the model without considering the individual-specific signals.

To quantitatively evaluate these methods, we calculate the true-positive rate (TPR) and false-positive rate (FPR) of recurrent CNV identification and plot the receiver operating characteristic curves (TPR versus FPR) under six scenarios with a noise level of $\sigma = 1$. The receiver operating characteristics (ROC) curves are shown in Figure 3. A curve closer to the top and left borders indicates a better performance. To illustrate the importance of smoothness penalty in PLA, we also give the result of

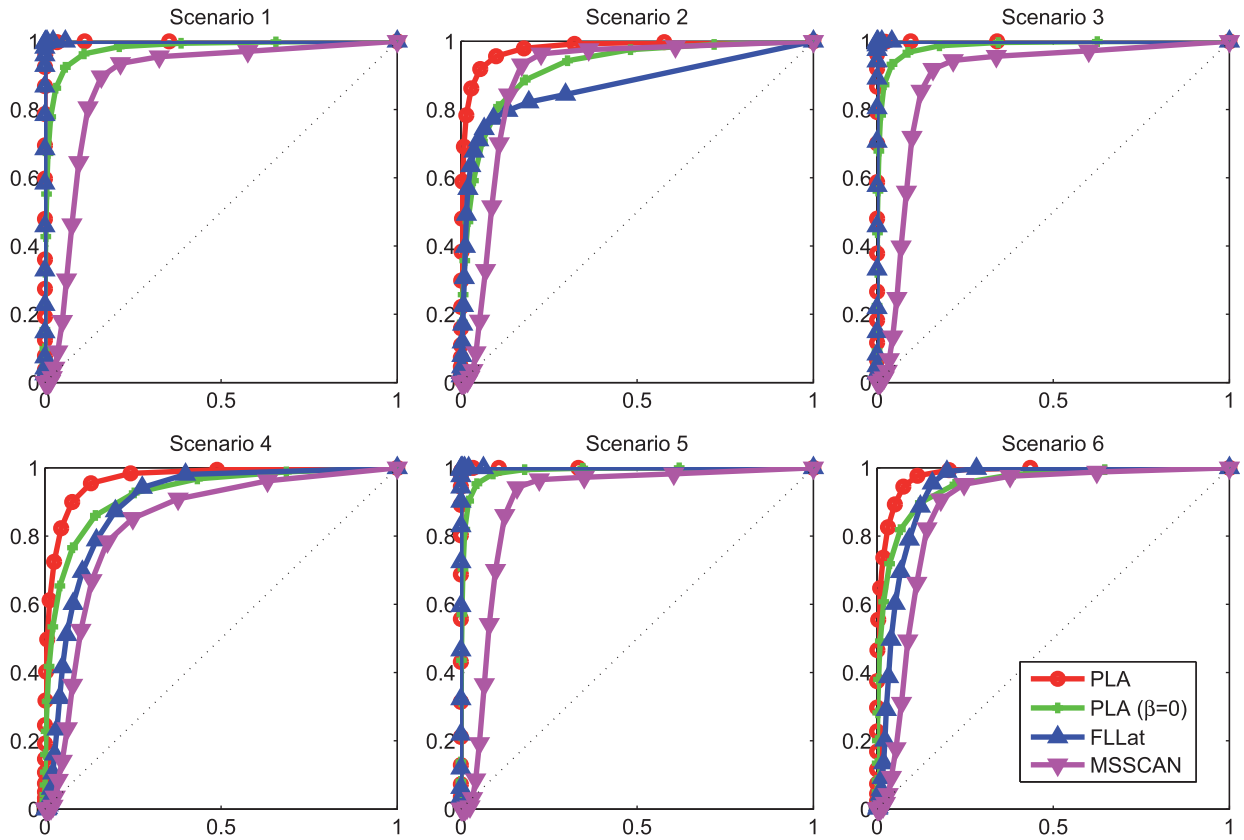

Fig. 3. The ROC curves of recurrent CNV identification on synthesized datasets of six scenarios. The y-axis and x-axis of each plot represent the TPR and the FPR, respectively

PLA without smoothness by setting $\beta = 0$ in (3). The ROC curves are produced by thresholding the recovered matrix with different thresholds.

For Scenarios 1, 3 and 5, both PLA and FLLat achieve nearly perfect results. For Scenarios 2, 4 and 6, the performance of them drops while PLA apparently outperforms FLLat. The difference between these two groups of scenarios is that there is only one recurrent region in Scenario 1, 3 and 5, while the other three scenarios involve more complicated recurrent patterns. When raw aCGH data contains two or more recurrent regions, FLLat is more prone to be influenced by the individual-specific variations. The expected number of recurrent regions is given as an input in FLLat, while it is not required in our algorithm. In practice, it is unrealistic to have such information. MSSCAN cannot differentiate between the recurrent and individual-specific CNVs during the segmentation step. Therefore, the output segmented matrix will be influenced by the individual-specific CNVs in a way similar to what is shown in Figure 2. Consequently, the corresponding ROC of MSSCAN is lower compared with PLA. It can also be observed that, in all six scenarios, removing the smoothness constraint will degrade the performance of PLA, which tells us that it is important to consider the smoothness constraint in the recurrent CNV identification.

3.1.2 Detection limit To illustrate the limit of population frequency for a variation being detected as a recurrent variation by our model, we simulate a dataset with 50 samples and a single variation region. The number of samples carrying the variation and the length of the variation are varying. To see whether the variation signal is detected, we calculate the relative difference between two matrices by $||\mathbf{X} - \mathbf{X}_0||_F / ||\mathbf{X}_0||_F$, where $\mathbf{X}$ is the recovered low-rank matrix and $\mathbf{X}_0$ is the simulated variation signal. A smaller value indicates a better approximation to the variation signal by the recovered low-rank component. The results are summarized in Figure 4. When the signal is strong (covering sufficient number of probes), it can be detected even if the recurrent frequency is small. For instance, the variation with a length of 10 probes is well detected if more than five samples carry it. When the signal is abrupt (covering few probes), it is unlikely to be detected even if the frequency is large because the signal is more likely to be included in the sparse component and the smoothness constraint will also prevent such signal from being detected. In practice, a spike-like signal is unlikely to be a true CNV. The results also depend on the relative weight between the low-rank term and the sparse term in our formulation. We used the default parameter setting in this experiment.

3.2 Real applications

To illustrate the applicability of our method in real cases, we have applied PLA to analyze two independent breast tumor datasets. We focused on the analysis of Chromosome 17, which has many frequently altered regions (Bekhouche *et al.*,

2011; Pollack *et al.*, 2002). The first dataset is from Pollack *et al.* (2002), which consists of aCGH profiles of 44 breast tumors over 382 probes for Chromosome 17. The second dataset is from Bekhouche *et al.* (2011) with a much higher resolution. It contains aCGH profiles of 173 breast tumors over 7727 unique probes for Chromosome 17. To remove the wave bias in each profile, we subtract the raw $\log_2$ ratios by local median values. The window size for the median calculation is one-fourth of the chromosome length.

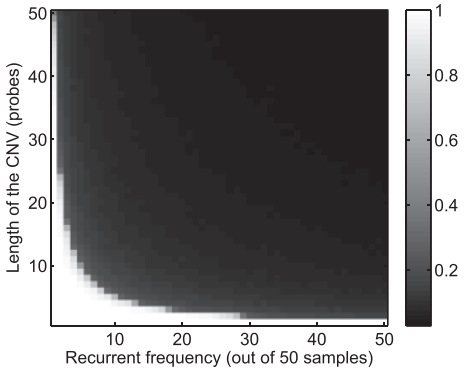

Fig. 4. A plot illustrating the conditions under which a variation is detected. A smaller value indicates a better detection

The heat maps of input data are shown in the top panels in Figure 5, followed by the low-rank and sparse components recovered by PLA. As we can see, the recurrent patterns are clearly presented in the low-rank components, while the individual-specific variations are mostly included in the sparse components. The rank of the low-rank component is 11 in Figure 5a and 55 in Figure 5b. To summarize the results, we calculated the frequency of gains G_i and losses L_i at each probe i by

$$G_j = \frac{1}{n}\sum_i \mathbf{1}(X_{ij} > T), L_j = \frac{1}{n}\sum_i \mathbf{1}(X_{ij} < -T) \qquad (18)$$

where T is a threshold that is set be 0.25 and $\mathbf{1}$ denotes the indicator function. The results are given in the fourth row in Figure 5. The recurrent regions can be clearly identified from the frequency plots, which are mainly located in the chromosome regions 17q11.2, 17q12, 17q21.3-q22 and 17q25. These identified regions match the results from both references (Bekhouche *et al.*, 2011; Pollack *et al.*, 2002). Many breast cancer-related genes are located in these regions. For example, genes ERBB2 and C17orf37 are located around Probe 3460 in Figure 5b, where a high peak appears in the frequency plot. Gene ERBB2 status is often used to indicate grades or stages of breast tumors. Gene C17orf37 is abundantly expressed in breast cancer and is claimed to be a tumor biomarker (Evans *et al.*, 2006). For simple

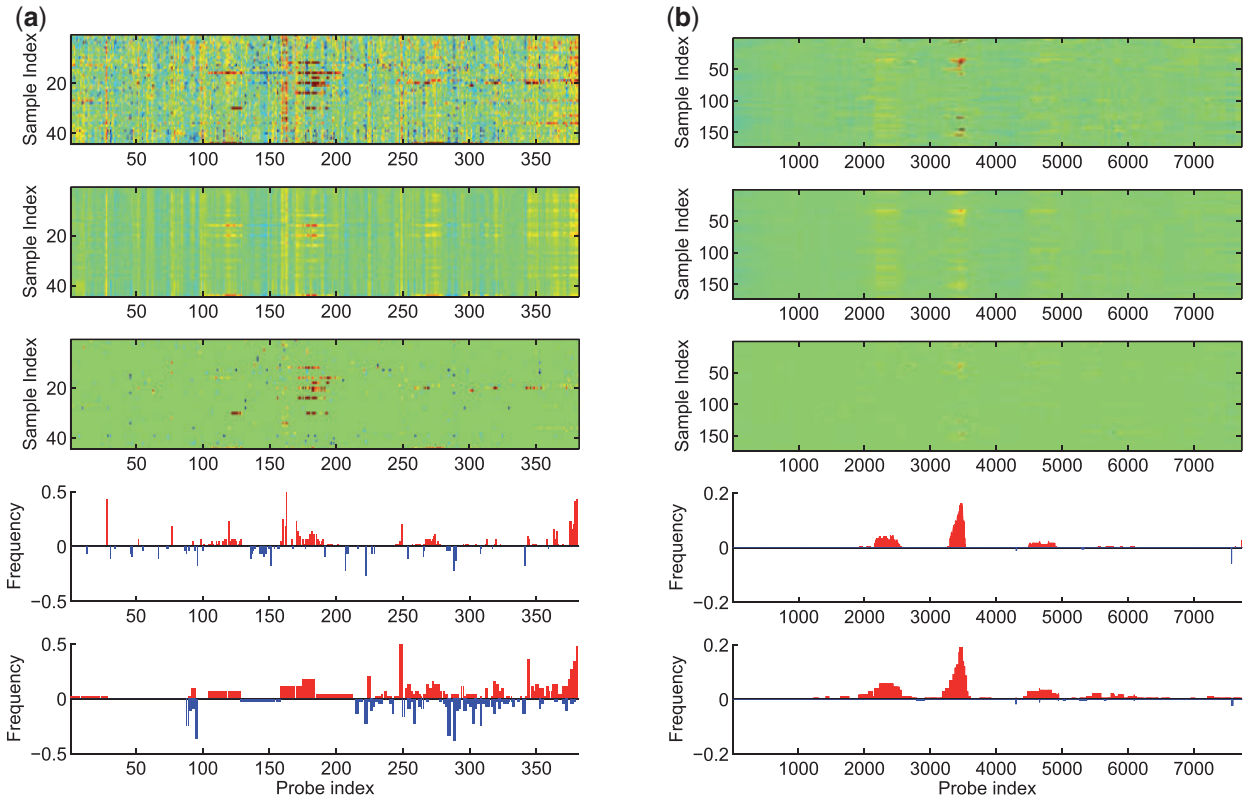

Fig. 5. The experimental results on the aCGH datasets from (**a**) Pollack *et al.* (2002) and (**b**) Bekhouche *et al.* (2011). Only Chromosome 17 is shown here. From top to bottom are the heat map of the input matrix, the recovered low-rank component, the sparse component, the recurrent frequency of alternations by thresholding the output of PLA and the recurrent frequency of alternations by thresholding the output of MSSCAN, respectively. For the bar plots, positive and negative y-values correspond to gains and losses, respectively

comparison, the results using MSSCAN are shown in the bottom panels. The frequency plot of MSSCAN is obtained by thresholding the segmentation result of MSSCAN. Compared with MSSCAN, our result is more sparse but accurately captures the recurrent regions that have been verified to have high correlation with breast cancer using gene expression (Bekhouche *et al*., 2011).

3.3 Computational cost

We test the algorithm on a desktop PC with a 3.4 GHz Intel i7 CPU and 8 GB RAM. For the dataset from Pollack *et al*. (2002) with 44 samples and 382 probes, PLA takes 0.4 s while MSSCAN takes 24.8 s. For the dataset from Bekhouche *et al*. (2011) with 173 samples and 7227 probes, PLA takes 146.4 s while MSSCAN takes 322.1 s. For the dataset from Bekhouche *et al*. (2011), the peak value of RAM usage of our program is 90 MB while that of MSSCAN is 170 MB.

4 DISCUSSION

In this article, we propose a new method to identify recurrent CNVs via recovering a low-rank matrix from raw aCGH profiles. The proposed method models different scenarios in a single framework. With a convex formulation, our algorithm guarantees a global optimal solution. We demonstrate the ability of our method to separate recurrent CNVs from other variations in both simulated data and real data. The low-rank matrix output of our method can be used as an input to other recurrent region-finding algorithms based on permutation test or other statistical analyses.

After the decomposition, the individual-specific CNVs are mostly included in the sparse component, as shown in Figures 2 and 5. However, the sparse component may also contain noise and measurement error. To precisely detect individual-specific CNVs, we may have to apply a postprocessing step to the sparse component. Another possible approach is to first use existing algorithms (e.g. fused lasso and TVSp) to remove noise in aCGH profiles and then apply our model to decompose the processed profiles into recurrent CNVs (low rank) and individual-specific CNVs (sparse).

Funding: The Hong Kong RGC-Theme-based Research Scheme (T12-402/13N); the Hong Kong Baptist University (FRG2/13-14/005).

Conflict of Interest: none declared.

REFERENCES

Bekhouche,I. *et al*. (2011) High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PLoS One*, **6**, e16950.

Ben-Yaacov,E. and Eldar,Y. (2008) A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, **24**, i139–i145.

Beroukhim,R. *et al*. (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.

Boyd,S. (2010) Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends. Mach. Learn*., **3**, 1–122.

Cai,J. *et al*. (2010) A singular value thresholding algorithm for matrix completion. *SIAM J. Optim*., **20**, 1956–1982.

Candès,E. *et al*. (2011) Robust principal component analysis? *J. ACM*, **58**, 11.

Diskin,S. *et al*. (2006) Stac: a method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res*., **16**, 1149–1158.

Eckart,C. and Young,G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.

Evans,E.E. *et al*. (2006) C35 (c17orf37) is a novel tumor biomarker abundantly expressed in breast cancer. *Mol. Cancer Ther*., **5**, 2919–2930.

Fazel,M. (2002) *Matrix rank minimization with applications*. PhD Thesis, Stanford University, Stanford, CA.

Gonzalez,E. *et al*. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434–1440.

Grant,M. and Boyd,S. (2008) *CVX: Matlab software for disciplined convex programming*. http://cvxr.com/cvx/ (26 March 2014, date last accessed).

Guttman,M. *et al*. (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet*., **3**, e143.

Hastings,P. *et al*. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet*., **10**, 551–564.

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

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

Lee,J.A. and Lupski,J.R. (2006) Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, **52**, 103–121.

Lee,C. *et al*. (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat. Genet*., **39**, S48–S54.

Liu,J. *et al*. (2010) An efficient algorithm for a class of fused lasso problems. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 323–332. ACM.

Lupski,J.R. (2007) Genomic rearrangements and sporadic disease. *Nat. Genet*., **39**, S43–S47.

Marioni,J. *et al*. (2006) Biohmm: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Meer,P. *et al*. (1991) Robust regression methods for computer vision: a review. *Int. J. Comp. Vision*, **6**, 59–70.

Nowak,G. *et al*. (2011) A fused lasso latent feature model for analyzing multi-sample aCGH data. *Biostatistics*, **12**, 776–791.

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

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Picard,F. *et al*. (2011) Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics*, **12**, 413–428.

Pinkel,D. and Albertson,D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genetics*, **37**, S11–S17.

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

Rancoita,P. *et al*. (2009) Bayesian DNA copy number analysis. *BMC Bioinformatics*, **10**, 10.

Rueda,O. and Diaz-Uriarte,R. (2010) Finding recurrent copy number alteration regions: a review of methods. *Curr. Bioinform*., **5**, 1–17.

Shah,S.P. *et al*. (2007) Modeling recurrent DNA copy number alterations in array cgh data. *Bioinformatics*, **23**, i450–i458.

Stjernqvist,S. *et al*. (2007) Continuous-index hidden Markov modelling of array CGH copy number data. *Bioinformatics*, **23**, 1006–1014.

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

Tropp,J.A. (2006) Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory*, **52**, 1030–1051.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

Zhang,N. *et al*. (2010a) Detecting simultaneous changepoints in multiple sequences. *Biometrika*, **97**, 631–645.

Zhang,Q. *et al*. (2010b) CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics*, **26**, 464–469.

Zhou,Z. *et al*. (2010) Stable principal component pursuit. In: *Proceedings of the IEEE International Symposium on Information Theory, Austin, TX, USA*.

Zhou,X. *et al*. (2013) Multisample aCGH data analysis via total variation and spectral regularization. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **10**, 230–235.