Detection Copy Number Variants from NGS with Sparse and Smooth Constraints

Yue Zhang, Yiu-ming Cheung, Bo Xu, and Weifeng Su

Abstract—It is known that copy number variations (CNVs) are associated with complex diseases and particular tumor types, thus reliable identification of CNVs is of great potential value. Recent advances in next generation sequencing (NGS) data analysis have helped manifest the richness of CNV information. However, the performances of these methods are not consistent. Reliably finding CNVs in NGS data in an efficient way remains a challenging topic, worthy of further investigation. Accordingly, we tackle the problem by formulating CNVs identification into a quadratic optimization problem involving two constraints. By imposing the constraints of sparsity and smoothness, the reconstructed read depth signal from NGS is anticipated to fit the CNVs patterns more accurately. An efficient numerical solution tailored from alternating direction minimization (ADM) framework is elaborated. We demonstrate the advantages of the proposed method, namely ADM-CNV, by comparing it with six popular CNV detection methods using synthetic, simulated, and empirical sequencing data. It is shown that the proposed approach can successfully reconstruct CNV patterns from raw data, and achieve superior or comparable performance in detection of the CNVs compared to the existing counterparts.

Index Terms—Copy number variants, read depth, sparsity, total variation

1 Introduction

Copy number variants (CNVs), a common type of structural variation, involve a duplication or deletion of a DNA segment larger than one kbp. Rapid development of new techniques for uncovering the intricacies within the human genome has revealed CNVs to be potential candidates for explaining various phenotype differences and genetic diseases [1]. Recent studies have revealed a strong correlation between CNVs and classic Mendelian diseases, plus other, less well characterized conditions, including autism, schizophrenia, osteoporosis, and certain tumor types [2], [3], [4]. Various laboratory techniques have been developed to measure the DNA copy number. Traditional major approaches include array-based comparative genomic hybridization (arrayCGH) and single-nucleotide polymorphism (SNP) array methods [5]. The next-generation sequencing are adopted as a popular strategy for genotyping and has included comprehensive characterization of CNVs by generating hundreds of millions of short reads in a single run [6]. The NGS could achieve higher coverage and resolution, thus allowing for more accurate estimation of copy numbers and detection of breakpoints with high throughput, than arrayCGH does. This has dramatically increased our capability of detecting CNVs. However, due to the complexity of the genome and the short read lengths associated with NGS technology, there are still many challenges associated with the analysis of NGS data for CNVs.

There are multiple approaches to having the copy number profiles from sequenced data. One of the major and popular approaches is based on counting of depth of coverage (DOC, also known as read-depth (RD) methods) [7], [8]. The DOC methods count the number of reads that fall in each prespecified window of a certain size [7], [9], [10] provided that the sequencing process follows a Poisson distribution with its mean value being proportional to the number of copies. The CNV calling is then obtained by detecting deviations between the number of reads mapped to a chromosome window with its expectation. However, the empirical sequencing process itself will inevitably introduce certain biases, such as non-uniform GC-content and low map-ability ratios. Therefore, the observed copy numbers of the genome appear to have spurious signals due to over- or under-sampling. While a number of successful approaches have been presented for copy number calling, there is still a paucity of methodology for analyzing it through systematical framework to achieve high accuracy and nice robustness.

In this paper, we tackle the problem by formulating the RD signal reconstruction problem into a quadratic minimization problem involving two constraints. Our proposed alternating direction minimization-total variational (ADM-CNV) method offers two main contributions as opposed to the previous methods: 1) The RD signal reconstruction is characterized by a well-designed model to reveal intrinsic structure, and thus facilitating the subsequent statistical testing for CNV detection; 2) An efficient numerical method using a

- *Y. Zhang is with the Electrical and Information College, Jinan University, Zhuhai, China, and the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China. E-mail: yuezhang@comp.hkbu.edu.hk.*
- *Y.M. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, and the BNU-HKBU United International College, Zhuhai, China. E-mail: ymc@comp.hkbu.edu.hk.*
- *B. Xu is with the School of Computer Science and Technology, South China University of Technology, Guangzhou, China. E-mail: 446712385@qq.com.*
- *W. Su is with the BNU-HKBU United International College, and Zhuhai Key Laboratory of Agricultural Product Quality and Food Safety, Zhuhai, China. E-mail: wfsu@uic.edu.hk.*

Manuscript received 2 Oct. 2015; revised 15 Apr. 2016; accepted 26 Apr. 2016. Date of publication 3 May 2016; date of current version 4 Aug. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2016.2561933

classical alternating direction minimization framework is tailored to solve the ADM-CNV model. The key characteristic of the ADM method that makes it different from standard methods in solving total variation (TV) related minimization problems [8] is that it does not require matrix inversion, thus greatly saving computational cost. This is particularly useful in NGS sequencing data analysis, in which hundreds of thousands of signals need to be recovered.

The remainder of the paper is as follows: Section 2 overviews the existing models for analyzing copy number variants. In particular, the methods based on read-depth counting for CNV analysis are briefly reviewed. In Section 3, we reformulate the detection of copy number variants into signal reconstruction problem, which is constrained by multiple conditions to satisfy the characteristics of copy number distribution. An exact numerical solution is also presented and its theoretical convergence is analyzed in this section. Section 4 demonstrates the performance of the proposed model through extensive experiments on both synthetic and empirical sequencing data. Finally, concluding remarks are given in Section 5.

2 Preliminaries

The current methodology for reconstructing CNV information from observed data can be broadly divided into four categories: (i) depth of coverage (DOC, also known as read-depth (RD) methods) [7], [8], (ii) paired-end mapping (PEM, also known as read-pair methods) [11], (iii) split-read (SR) [12], [13], and (iv) assembly-based (AS) methods [14]. All of these methodologies, except for the last one, require initially mapping the sequenced reads to a known reference genome. Read depth analysis is particularly effective for exome sequence data, as it does not rely on sequencing into or near CNV breakpoints.

Most of the existing tools for CNV calling based on read depth can be broadly divided into two categories: parametric versus non-parametric schemes. The performance of parametric approaches heavily depends on the underlying statistical distributions. For example, Magi [15] used the stochastic process of a shifting level model to simulate multi-sequential samples. ExomeCNV [16] and CNV-seq [10], assume a Gaussian distribution of the read ratios. Also, they assume that the proportion of reads matching to a specific sample usually follows a binomial distribution whose success rate is determined by the genome-wide read count ratio between the test sample and the reference set, as well as the potential presence of CNVs. However, it has been shown that the binomial assumption is actually violated in practice due to technical variability, induced noise in library preparation, and capturing or sequencing error. ExomeDepth [17] uses a beta-binomial model to approximate the distribution of the read count ratio, and then employs statistical inference to detect the CNVs. CLImAT [18] takes tumor impurity and ploidy into consideration for identifying genomic aberrations. DeAnnCNV [19] is an online integrated tool to precisely detect and systematically annotates copy number variations from whole-exome sequencing (WES) data. Hidden Markov models (HMMs) and many variations thereof [20] have dominated the parametric approach. The unique characteristics of HMMs lie in its flexibility for handling several common complications, including variable single nucleotide polymorphism (SNP) frequencies, variable distances between adjacent SNPs, linkage disequilibriums, and relationships between study subjects.

Nonparametric methods attempt to reconstruct the CNV copy number from the observed read depth signal in concert with the application of subsequent statistical inference on the estimated signal to detect the CNVs. For simplicity, let us consider a plateau/basin in the reconstructed signal to be a duplication/deletion event. Therefore, the CN is assumed to be piece-wise constant function with two fundamental characteristics: sparsity because of few CNVs, and smoothness because of contingency positions possessing similar CNs. In such a configuration, the CNV detection is considered to be a piece-wise linear signal recovering problem. The signal is normalized around zero, thus possessing sparsity. Mathematically, let the parameter vector $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ quantify DNA levels at n successive SNPs. These levels are normalized such that $r_i = 0$ corresponds to the standard copy number 2, where SNP i is represented once each on the maternal and paternal chromosomes. After such configuration, the problem is reformulated as a signal reconstruction problem. A classical approach to addressing it is by the fussed lasso scheme [21] with its refinements [22]. It was also tailored to use in CNV detection on array CGH [23]. Most recently, the conditions that the fussed lasso could consistently recover the piecewise constant pattern have been investigated [24]. Numerical scheme for solving the fussed lasso typed problem includes approximation method of path coordinate descendent optimization [25], and linearization method within ADM framework [26], [27].

3 The Proposed Method

3.1 Problem Modeling

The proposed method for analyzing raw NGS data consists of four steps. In the first step, the short reads are aligned to a reference genome using standard tools such as MAQ [28] and Bowtie [29]. In the second step, the aligned reads are used to estimate the read depth signal $\boldsymbol{r}$ to measure its density of the aligned reads. Next, our proposed method is used to reconstruct the copy number from the raw ratio reads. Finally, statistical testing is borrowed from other methods to detect the suspicious CNVs.

Recent progress on detection of CNVs from read depth signal ratios has modeled this problem into a least-square minimization optimization, satisfying certain constraints. For example, CNV-TV [27] attempts to minimize the distance between the raw signal and the reconstructed one, coupled with constraints from a total variational (TV) term to penalize the similarities between successive reads. However, such a constraint is not sufficient for CN reconstruction because the CNV is very sparse.

We tackle the problem by modeling the CNV reconstruction into a least-square minimization problem, penalized by two constraints:

$$\min_{x_i} f(x_i) = \frac{1}{2}\sum_{i=1}^{n}(x_i - r_i)^2 + \lambda_1 \sum_{i=1}^{n} \|x_i - x_{i-1}\|_1 + \lambda_2 \sum_{i=1}^{n} \|x_i\|_1, \tag{1}$$

where r_i is the observed reads depths at the ith position, and x_i is the reconstructed copy number. The second term of Eq. (1) attempts to penalize the similarities between adjacent sites, whereas the third term aims to achieve a sparse signal representation.

We would like to emphasize here that the sparsity and smoothness may not be applicable to arm-level CNVs, which has a high frequency of occupying one chromosome arm exactly. However, the sensitivity of detecting focal amplifications and deletions will be greatly increased after filtering out the arm-level CNVs by using either amplitude or length thresholds [30]. An effective scheme is to firstly separate the two types of CNVs by using length threshold and then discriminate both of them individually.

As the second term in above equation is actually a TV term, we rewrite it as:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{r}\|_2^2 + \lambda_1\|\boldsymbol{D}\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{x}\|_1, \tag{2}$$

where $\boldsymbol{D}$ is the first order difference matrix, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$.

Minimization problems involving a TV term are common in disparate areas, including signal processing and image recovering [31]. Due to the non-differentiability of the TV term, Zhang et al. [32], [33] proposed to use an l_2 norm to approximate the TV norm, and then to use a majority minimization framework to solve the problem. Such an approach will typically lead to explicit solutions, and is thus very efficient at a sacrifice of accuracy. To obtain a sparse solution attributing to the TV norm, a normal practice is to use the Lasso numerical method to solve the problem [27], [34]. However, a common drawback of them is their highly demanding computational cost due to the necessity of calculating a matrix inversion or introducing too much slack variables.

Recent researches [31], [34] have successfully demonstrated the efficiency of ADM in solving image restoration problems involving TV norms. The numerical solution show surprisingly fast speed and high accuracy. The image restoration ADMs are variants of the classical Augmented Lagrangian method for optimization problems with separable structures and linear constraints, and they have been intensively studied in the optimization community. Inspired by this idea, we reformulated the problem in Eq. (2) into an optimization problem with favorably separable structures, thus allowing it to be efficiently solved using ADMs. The separable structure decomposition is particularly useful in sequencing data analysis due both to its large/huge dimensionality, and to the large amount of data involved; plus it dramatically reduces computational cost by overcoming the matrix inversion burden, required by Lasso and majority minimization schemes [31], [34].

3.2 Fast Numerical Solution

Mathematically, let $\boldsymbol{\omega} = \lambda_1\boldsymbol{D}\boldsymbol{x}, z = \lambda_2\boldsymbol{x}$, $\boldsymbol{B} = \begin{pmatrix}\lambda_1\boldsymbol{D}\\ \lambda_2\boldsymbol{I}\end{pmatrix}$, and $\boldsymbol{y} = \begin{pmatrix}\boldsymbol{\omega}\\ \boldsymbol{z}\end{pmatrix}$ with $\boldsymbol{I}$ being the identity matrix. The minimization problem (2) could be equivalently rewritten as,

$$\begin{aligned}&\arg\min\|y\|_1\\ &\text{subject to:}\\ &\boldsymbol{x}\in\mathcal{R}^n\\ &\boldsymbol{\omega} = \lambda_1\boldsymbol{D}\boldsymbol{x}\\ &\boldsymbol{z} = \lambda_2\boldsymbol{x}\\ &\|\boldsymbol{x}-\boldsymbol{r}\|_2^2 \le \alpha.\end{aligned}$$

The augmented Lagrangian of the minimization problem is given by:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) = \chi_S(\boldsymbol{x}) + \|\boldsymbol{y}\|_1 + <\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}> + \frac{1}{2}\gamma\|\boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}\|_2^2, \tag{3}$$

where the parameters $\boldsymbol{\xi}$ and γ are Lagrangian multiples. The function $\chi_S(\boldsymbol{x})$ is the usual indication function on the set $S = \{\boldsymbol{x}\in R^n, \|\boldsymbol{x}-\boldsymbol{r}\|_2^2 \le \alpha\}$ for a predefined parameter α.

Let $f_1(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$, $f_2(\boldsymbol{y}) = \|\boldsymbol{y}\|_1$, the minimization problem falls into the standard framework of Alternating Direction Method (ADM) with the following notations:.

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\xi}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{y}) + <\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}> + \frac{1}{2}\gamma\|\boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}\|_2^2. \tag{4}$$

This new formulation allows it to be decoupled into two separate sub-problems with variables $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and, therefore, to be solved in an iterative manner.

Step 1: Find

$$\begin{aligned}&\boldsymbol{x}^{k+1} \in \arg\min <\boldsymbol{\xi}, \boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}^k> + \frac{1}{2}\gamma\|\boldsymbol{B}\boldsymbol{x}-\boldsymbol{y}^k\|_2^2,\\ &\text{subject to:}\\ &\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \le \alpha.\end{aligned} \tag{5}$$

With an algebra transformation, minimization of Eq. (5) is equivalent to the following constrained least square problem:

$$\begin{aligned}&\boldsymbol{x}^{k+1} \in \arg\min\frac{1}{2}\|\boldsymbol{B}\boldsymbol{x} + \frac{\boldsymbol{\xi}}{\gamma} - \boldsymbol{y}^k\|_2^2,\\ &\text{subject to:}\\ &\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \le \alpha.\end{aligned}$$

Let $\hat{\boldsymbol{x}} = \boldsymbol{x}-\boldsymbol{x}_0$, then the above minimization can be rewritten as:

$$\hat{\boldsymbol{x}}^{k+1} \in \arg\min\frac{1}{2}\|\boldsymbol{B}\hat{\boldsymbol{x}} + \boldsymbol{B}\boldsymbol{x}_0 + \frac{\boldsymbol{\xi}}{\gamma} - \boldsymbol{y}^k\|_2^2 + \delta\|\hat{\boldsymbol{x}}\|^2,$$

where $\delta\in[0,+\infty)$ is a Lagrange multiplier. This is equivalent to solving the least square problem of $\min_{\boldsymbol{x}}\|\hat{\boldsymbol{B}}\hat{\boldsymbol{x}}-\hat{\boldsymbol{c}}\|^2$, where $\hat{\boldsymbol{B}} = \begin{pmatrix}\boldsymbol{B}\\ \delta\boldsymbol{I}_n\end{pmatrix}$, $\hat{\boldsymbol{c}} = \begin{pmatrix}\boldsymbol{c}\\ 0\end{pmatrix}$, and $\boldsymbol{c} = -(\boldsymbol{B}\boldsymbol{x}_0 + \frac{\boldsymbol{\xi}}{\gamma} - \boldsymbol{y}^k)$.

The analytical solution is explicitly given by:

$$\begin{aligned}\hat{\boldsymbol{x}}^* &= (\hat{\boldsymbol{B}}^T\hat{\boldsymbol{B}})^{-1}\hat{\boldsymbol{B}}^T\hat{c}\\ &= (\boldsymbol{B}^T\boldsymbol{B} + \delta^2\boldsymbol{I})^{-1}\boldsymbol{B}^T c\\ &= (\lambda_1^2\boldsymbol{D}^T\boldsymbol{D} + \lambda_2^2\boldsymbol{I} + \delta^2\boldsymbol{I})^{-1}\boldsymbol{B}^T c.\end{aligned}$$

One may note that the matrix D is circulant and thus could be diagonalized by Fourier transform as $D = F^T K F$, where F is 2-D discrete Fourier transform (DFT) and K is a diagonal matrix containing the DFT coefficients of the difference operator D positive definite and thus its inverse can be computed efficiently by singular value decompositions. It follows that

$$\hat{\boldsymbol{x}}^* = \boldsymbol{F}^T(\lambda_1^2 \boldsymbol{K}^T \boldsymbol{K} + \lambda_2^2 \boldsymbol{I} + \delta^2 \boldsymbol{I})^{-1} \boldsymbol{F} \boldsymbol{B}^T c. \tag{6}$$

Step 2: Find

$$\boldsymbol{y}^{k+1} = \arg\min f_2(\boldsymbol{y}) - < \boldsymbol{\xi}^k, \boldsymbol{y} > + \frac{1}{2}\gamma \|\boldsymbol{B}\boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2. \tag{7}$$

Its solution is given by Moreau proximity operator, which is simply a soft threshold as follows:

$$\begin{aligned} \boldsymbol{y}^{k+1} &= \arg\min f_2(\boldsymbol{y}) - < \boldsymbol{\xi}^k, \boldsymbol{y} > + \frac{1}{2}\gamma \|\boldsymbol{B}\boldsymbol{x}^{k+1} - \boldsymbol{y}\|_2^2 \\ &= \arg\min f_2(\boldsymbol{y}) + \frac{\gamma}{2}\|\boldsymbol{y} - (\boldsymbol{B}\boldsymbol{x}^{k+1} + \frac{\boldsymbol{\xi}^k}{\gamma})\|^2 \\ &= \text{shrinkage}_{1/\gamma}[(\boldsymbol{B}\boldsymbol{x}^{k+1} + \frac{\boldsymbol{\xi}^k}{\gamma})]. \end{aligned}$$

Therefore, the analytical solution $\boldsymbol{y}^*$ is given by,

$$\boldsymbol{y}^* = \boldsymbol{y}_0 - \min(\frac{1}{\gamma}, |\boldsymbol{y}_0|)\frac{\boldsymbol{y}_0}{|\boldsymbol{y}_0|}, \tag{8}$$

where $\boldsymbol{y}_0 = (\boldsymbol{B}\boldsymbol{x}^{k+1} + \frac{\boldsymbol{\xi}^k}{\gamma})$.

We then update $\boldsymbol{\xi}^k$ by

$$\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}^k + [\boldsymbol{y}^{k+1} - \begin{pmatrix} \lambda_1 \boldsymbol{D}\boldsymbol{x}^{k+1} \\ \lambda_2 \boldsymbol{x} \end{pmatrix}]. \tag{9}$$

The above steps are updated iteratively until convergence to achieve the final optimal solution of $\boldsymbol{x}$.

The prominent characteristic of the aforementioned numerical solution is its separable structure that allows for finding a solution quickly. After simple algebra operations, explicit analytical solution for the two sub-problems, Eqs. (5) and (7) are obtained. In both sub-problems, their solutions involve low-cost calculations and thus could be solved efficiently.

3.3 Convergence and Complexity Analysis

The convergence of ADM-CNV is addressed by the following corollary of a theorem by *Eckstein-Bertsekas* [35], which is provided in Appendix.

Corollary 1. *The aforementioned algorithm for the problem of (4) converges to a minimizer.*

Proof. The proposed ADM-CNV is an instance in Theorem 1, where $f_1(\boldsymbol{x}) = \chi_S(\boldsymbol{x})$ and $f_2(\boldsymbol{y}) = \|\boldsymbol{y}\|_1$ are closed, proper and convex. The matrix $B = \begin{pmatrix} \lambda_1 \boldsymbol{D} \\ \lambda_2 \boldsymbol{I} \end{pmatrix}$ has full column rank. According to Theorem 1, ADM-CNV is convergent to a minimizer of the objective function.

The optimal solution is obtained through alternative updating the sequence of $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{\xi}^k\}$, whose solution is given in Eqs. (6) and (8), respectively. In Eq. (6), the inverse of diagonal matrix $(\lambda_1^2 \boldsymbol{K}^T \boldsymbol{K} + \lambda_2^2 \boldsymbol{I} + \boldsymbol{I})$ costs $O(n)$. The products by $\boldsymbol{F}, \boldsymbol{F}^T$ with the inverse diagonal matrix need $O(n\log n)$. Similarly, the product of $\hat{\boldsymbol{B}}^T \hat{c}$ costs $O(n\log n)$ too. Thus, the total computation cost in Eq. (6) is $O(n\log n)$.

In Eq. (8), the soft thresholding operator $\min(\frac{1}{\gamma}, |\boldsymbol{y}_0|)\frac{\boldsymbol{y}_0}{|\boldsymbol{y}_0|}$ costs $O(n)$. In total, the computational cost of ADM-CNV scales as $O(n \log n)$. □

3.4 Hyper-Parameters Pruning

Another challenging problem remains with model (2) is that of tuning the hyper-parameters of λ_1, λ_2. The parameters of λ_1, λ_2 have a profound influence on the reconstructed CNV signal. Since λ_1 controls the signal difference penalty, a larger value will lead to a smoother signal. In comparison, the value of λ_2 controls the sparsity of the CNV, a larger value will lead to a curve with most of its elements being zero. Although the two parameters could be set by a rule of thumb, an automatic strategy for choosing the parameters is desirable. Parameter pruning can be viewed as a model selection problem. There are a few commonly used model selection schemes, including the least angle regression [36] and Schwarz information criterion (SIC) [37]. Unfortunately, such model selection schemes do not return encouraging results in our framework, possibly due to the bias introduced by sequencing or our minimization process.

The parameter λ_2 controls the sparsity of the reconstructed ratios and thus serves as a thresholding operator. It should be large enough to smear out the noise, but not so large that over-smoothing the signal occurs [38]. A classical approach is to use the well-known hard threshold scheme in wavelet shrinkage [39],

$$\lambda_2 = \sigma\sqrt{\ln(N)},$$

where N is the sample size and σ is the standard variance of the noises that are assumed to follow $N(0, \sigma^2)$. Since CNVs are sparse in the data, we can estimate σ from the data using the median-absolute-deviation estimator, $\hat{\sigma} = 1.48\{median\{\boldsymbol{r} - median(\boldsymbol{r})\}\}$.

As the relative weight $\frac{\lambda_1}{\lambda_2}$ balances the two terms in $\lambda_1\|D\boldsymbol{x}\|_1 + \lambda_2\|\boldsymbol{x}\|_1$ and controls the recovered read ratios, we empirically set $\lambda_1 = 20\lambda_2$.

To further investigate the influence of the ratio of the two penalty parameters over the model performance, we conducted extensive experiments by changing the ratio from 15 to 50. Our experiments manifested that, when the ratio of the two parameters is less than 20, it has minor influences on the model performance. However, if the ratio is larger than 20, the smoothing term will impose heavy constraints such that the resulted CN loses variation. Consequently, small CNVs will be smeared out and resulted in a low precision value.

4 Experimental Results

We conducted the experiments to demonstrate the performance of the proposed model. The testing datasets were categorized into three types: synthetic, simulated and empirical data. Due to the proximity of the proposed

method to CNV-TV [37], we firstly designed an experiment with synthetic data to illustrate the superiority of our proposed model over CNV-TV. In the second experiment, human genome data were employed to serve as fundamental testing data. CNVs with different length and copy numbers were introduced artificially into the human genome data to simulate real sequencing process. The proposed ADM-CNV method and six representative CNV detection methods were compared to evaluate their performances. Finally, experiment on an empirical data of chromosome 21 of NA19240 (Yoruba female) was comparatively studied. To save space, the details of experimental set-up are provided in Section 1 of the Supplementary File, which can be found on the Computer Society Digital Library at http://doi.10.1109/TCBB.2016.2561933.

4.1 Synthetic Studies by Comparing with CNV-TV

Random read depth data was created to simulate the reads from Illumina platform. At most positions in the data, the ratio was normalized to have zero mean value. At certain sites, read ratios following Poisson distributions with various mean values were added to simulate the CNVs. To test the robustness of the methods, the Gaussian noises with zero mean value and three different variances are further added to the data. The synthetic data are shown in the first column of Fig. 1. From the top to the bottom, the degraded noise level tends to larger. The results after using ADM-CNV and CNV-TV are shown in the second column of Fig. 1. The results by CNV-TV are denoted in red. The detected CN curve after ADM-CNV, highlighted in solid black, accurately characterizes the distribution of the read depth, denoted in blue. One may observe that the detected CNV in red line performed less satisfactorily than the proposed method did, especially when the noises level is high, in which only the most prominent CNV was detected as shown in Fig. 1a.

4.2 Comparative Studies on Simulated Data

To demonstrate the performance of the proposed ADM-CNV quantitatively, we simulated the empirical sequencing process to produce artificial data by introducing various CNVs manually. The simulation steps are briefly summarized as follows:

Step 1: Extract a real sequence with length 2 Mbp from chromosome 21 as the control genome. It was concatenated with its duplication, yielding the reference genome of length 4 Mbp;

Step 2: Introduce CNV with copy number c and single copy length d artificially to generate the test genome. In our experiments, d was set to be 6 kbp and c being tested at different values of 3 and 6;

Step 3: Sample the test genome to simulate the single-end short reads. The location of each read is uniformly distributed at the given genome (genomic coordinate from 1 to 4e6), and each read has length 35 bps to agree with the Illumina platform. The sample coverage is set at 1. In total, we generate $\frac{1*4e^6}{35} \approx 114286$ short reads. The reference genome is also sampled and handled to obtain the read depth ratio.

Step 4: Align the short reads to the first half of the reference genome with Bowtie2 [29] (the second half is only the duplication of the first half). Since a read may align to multiple locations, only the uniquely mapped reads are used for testing to minimize the bias brought by misalignment;

Step 5: Use a non-overlapping window of size 800 bps to calculate the read depth ratio. Then, we call ADM-CNV and other popular CNV detection methods to evaluate their performance. The result by each method is compared with the ground truth.

We compare the proposed ADM-CNV with five other popular CNV detection methods, including CNV-seq [10], FREEC [40], readDepth [41], CNVnator [9] and CNV-TV [27]. In the model of CNV-TV, the detection of CNV is modeled as a change-point detection scheme from the read depth signal, penalized with a total variational term. FREEC adopted a similar idea, yet estimation is obtained by using a non-overlapping sliding windows (raw CNP). CNV-seq and readDepth are based on statistical testing model. The former conducts the statistical confidence assessment of observed copy number ratios modeled as Gaussian ratio distribution. The latter one models the distribution of reads, which are uniquely mapping to the genome by negative-binomial distribution. CNVnator is based on mean-shift tracking to broaden the range of discovered CNVs.

For each pair of the two parameters (copy length d and copy number c), the experiment was repeated for 30 times. The results of CNV detection by each method are shown in Fig. 2. The horizontal axis denotes the experiment index, ranging from 1 to 30, while the vertical axis is the genomic coordinate. The blue line represents the detected CNV regions while the red dots denote the ground truth. The performance of each method was quantified by the accuracy of the detected CN matching with the ground truth. The better the matching is, the better performance of the CNV detection method is. Among the six methods, readDepth achieved superior performance by matching perfectly to the ground truth. CNVnator was less satisfactory in that it failed to detect true CNV twice and produced a large false positive in the last experiment. FREEC and CNV-TV resulted in inaccurate CNVs during the 30 experiments. In comparison, the proposed ADM-CNV and CNV-seq achieved very nice performance by almost perfectly matching to the ground truth. The ADM-CNV achieved slightly over-performance than CNV-seq by having smaller false positives. Such visual observations were further validated by quantitatively comparisons. The true positive ratio (truly detected CNVs located within the ground truth) over false positive ratio (falsely detected CNVs within the ground truth) for the 30 experiments was calculated and shown in Table 1. The performance of the ADM-CNV was suboptimal to the readDepth, but was comparable or better than the others. It performed robustly and steadily across the 30 repetitions.

4.3 Comparative Studies on Real Sequencing Data

We tested ADM-CNV on an empirical NGS data by comparison with the aforementioned five methods, CNV-TV [27], FREEC [40], CNV-seq [10], CNVnator [9], readDepth [41] and a recently proposed method of CLImAT [18], to demonstrate

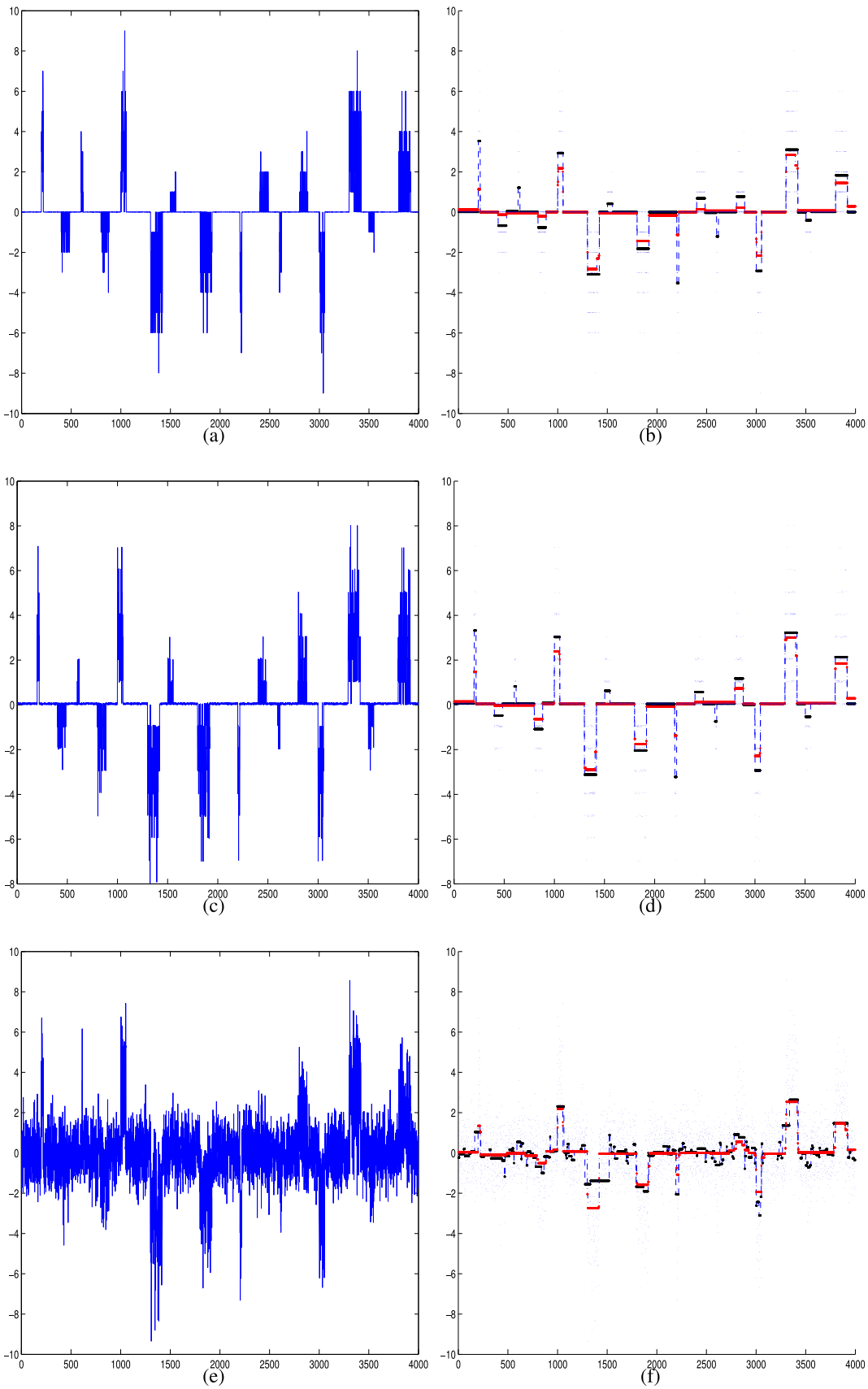

Fig. 1. Experiment on simulated read ratios following a Poisson distribution with various mean values, followed by Gaussian noise degraded by various variances. The reads with Gaussian noise with (a) small value of variance (0.01); (b) intermediate value of variance (0.1); and high variance (1) are shown in the first column. In comparison, the corresponding detected CNV curve are shown in the second column with the raw reads overlapped. The results by the proposed ADM-CNV are highlighted in black, while the one after CNV-TV is represented in red color. The proposed ADM-CNV can accurately detect all of the CNVs in different scenarios.

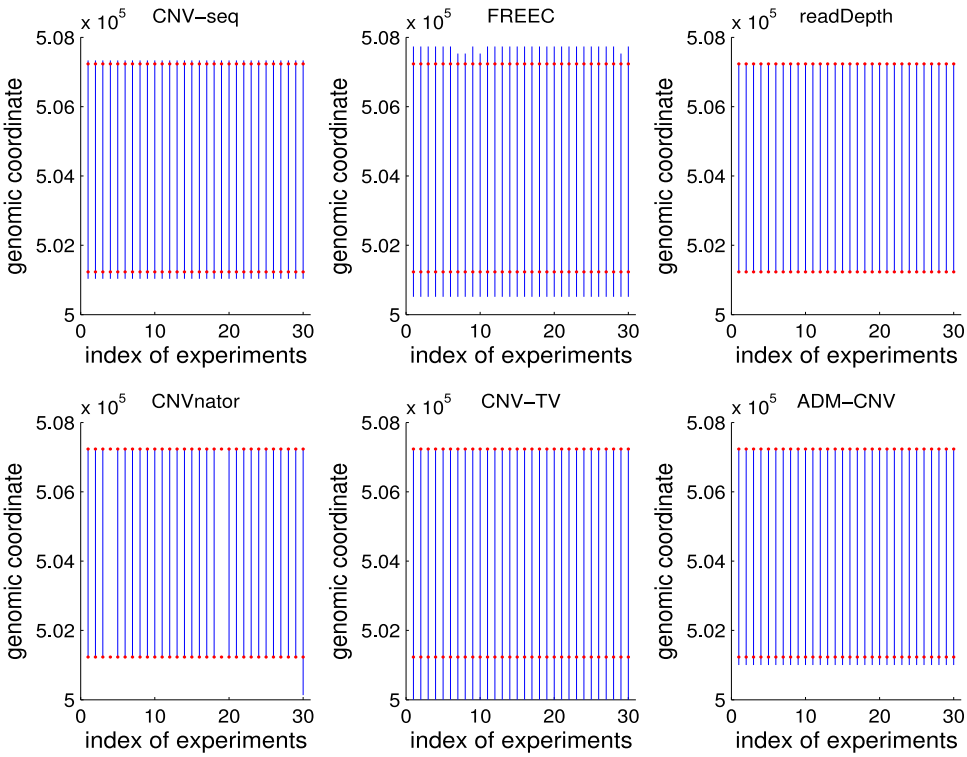

Fig. 2. A sample CNV detection result with $d = 6e3$, $cp = 6$ by repeating the detection scheme 30 times. The horizontal and vertical axis represent the experiment index and the genomic coordinate, respectively. The fraction between the true positive ratio and false positive ratio is formatted as (TPR/FPR) and is shown in Table 1. The detected CNVs are highlighted in blue lines, while the ground truth, given by the starting and ending position with known CNVs at the reference genome is in red dots.

its performance on empirical NGS data. The tested mapped reads data (BAM files) was downloaded from the 1,000 Genomes Project at (http://ftp.1000genomes.ebi.ac.uk/). The raw data was firstly aligned to a reference genome by alignment tools of Bowtie2 [29]. The reference samples were obtained by randomly sampled at the reference genome to obtain a .hit file. The location of each read is uniformly distributed at the given genome (genomic coordinate from 1 to the length of the chromosome 21), and each read has a length of 35 bps to agree with the Illumina platform. The sample coverage is set to be same as that of the bam file. Then the copy number signal was calculated by dividing the read count of each window's test sample by reference sample. Finally, the potential CNVs were detected from the read depth signal by the tested methods.

The DGV database of genomic variants was downloaded from http:// projects.tcag.ca /variation/ to mark all of the discovered CNVs reported in the literature, and these known CNVs were chosen to serve as ground truth. Here, we used the beta version of DGV, from which CNVs can be retrieved with respect to the sample, the platform, and the study [42]. The option of filter query was *external sample id* = *NA19240*, *chromosome* = *21*, $assembly = NCBI36/hg18$, *variant type* = *CNV*. To calculate the CNV detection power at a tolerable genome wide false discovery rate, the genome was artificially divided into sliding windows at 800bp with overlaps at 200bp.

The proposed ADM-CNV was applied on the data and the reconstructed CN (highlighted blue color) as well as the detected CNV (highlighted in red) were shown in Fig. 3. For better visualization, a short segment interval of 45.7M-46.1M (highlighted in green) was zoomed-in and displayed in Fig. 4. The two horizontal lines represents the global thresholds for the CNV, and the detected CNVs are highlighted in red. The aforementioned six methods were also tested to compare with ADM-CNV. The results were analyzed by Venn diagram and summarized in Table 2. This table quantifies the number of overlapping regions between those detected by the seven methods. All of the detected CNVs were validated in DGV. For clarity, only those blocks with numbers greater than 1,000 were used in the comparison. The integer 1 denotes there exists of a logical overlap between the ones by the tested methods and the reported ones in DGV, while 0 implies of zero overlapping. One may observe that ADM-CNV can detect as many CNV as 24,544 blocks, which were all identified in DGV. This number is dramatically larger than the ones detected by other methods except CLImAT. For example, CNV-TV can only detect 9,272 blocks, and CNVnator detected 14,435 blocks. The performance of CLImAT was superior by

TABLE 1
Number of True Positive/False Positive Detections for the 30 Trails in Simulation Experiment

Method	CNV-seq	FREEc	readDepth	CNVnator	CNV-TV	ADM-CNV
TPR/FPR	1/2.7e-4	0.994/6.3e-4	1/0	0.93/5e-3	0.997/6e-4	1/2.5e-4

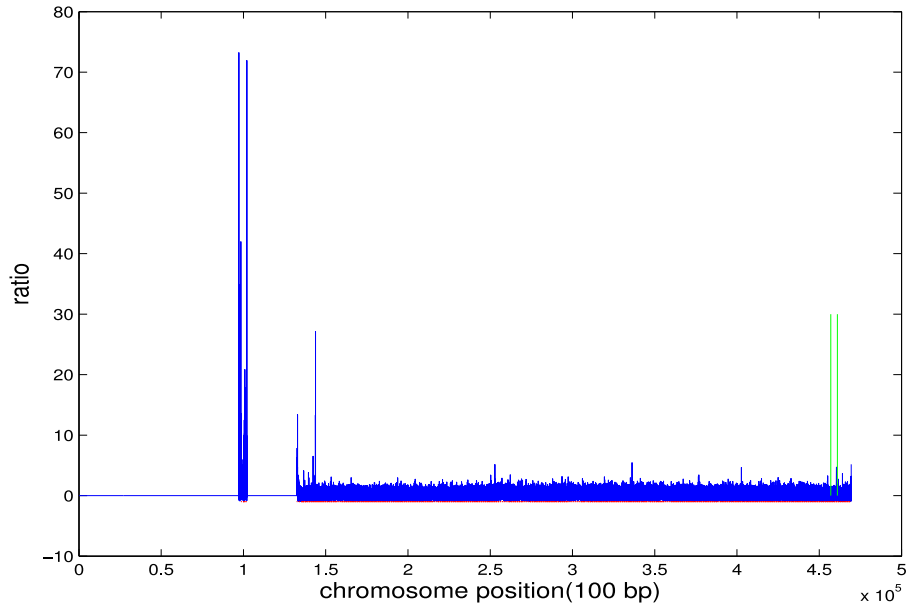

Fig. 3. Chromosome 21 from NA19240. The blue curve is the read depth signal. The CNV regions are shown in red dots, and the green lines indicate the bounds of the zoom region. A zoom between the region within the two vertical green lines is displayed in Fig. 4.

detecting as many as 7,85,590 blocks. Such nice performance was attributed to its integrative analysis, including careful handling of data preprocessing. However, it also resulted in a high computational costs, as shown in Table S1. The last column in Table 2 summarized the overlapping block number detected by all the seven methods. It provided a consistency measurements of the results. CLImAT preformed superior by detecting as many blocks as 7,85,590. The proposed ADM-CNV ranked the second by detecting 24,544 blocks. One may also observed from the fifth column that both CLImAT and ADM-CNV found 24,508 common blocks, which was more than 90 percent of the total number of 24,544 by the ADM-CNV. It implies that the proposed method could detect highly consistent variations. Another interesting observation from the last column is that only 5,678 blocks were uniformly detected by all the seven methods. Such inconsistences show that it prone to have a large number of false positives in CNV detection.

F-score quantitative measurements were also calculated for each method. The F-score measures the overlap ratio

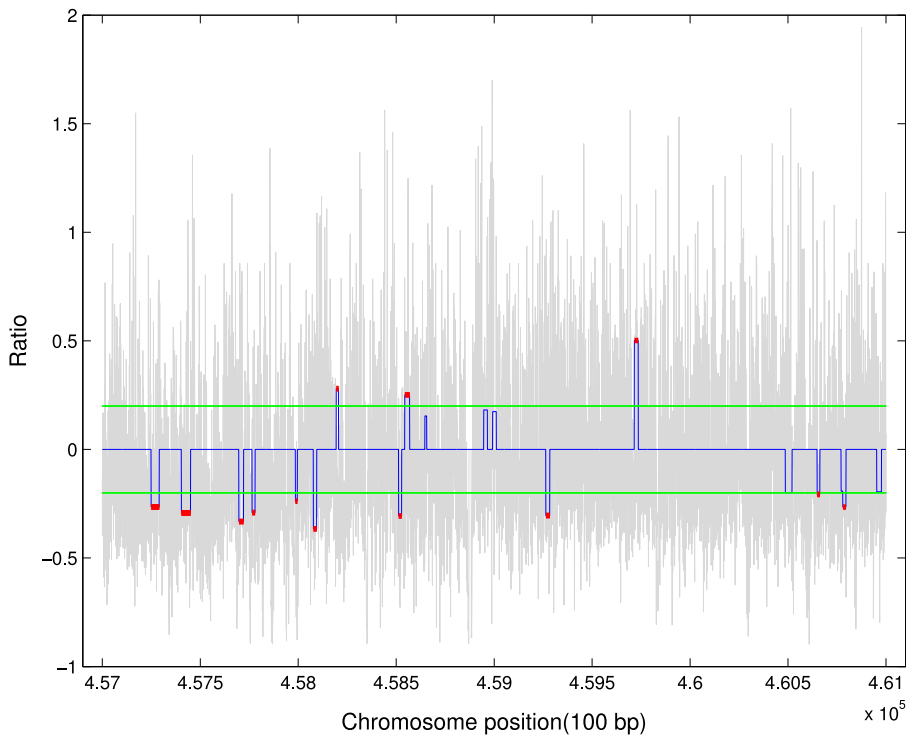

Fig. 4. The zoomed CNVs detected by ADM-CNV. The gray lines are the raw depth of the coverage data while the blue solid line is the reconstructed CN. The two horizontal green lines represent the global cutoff values for the copy number variation. Therefore, the ratio beyond the cutoff values is recognized as the CNV and is highlighted in red.

TABLE 2
A Summary of a Venn-Style Analysis Five-Way Tabulated Venn Diagram Obtained from Sample of NA19240

ADM-CNV	1	0	0	1	0	1	1	1	1	1	0	1
CNV-TV	0	1	0	0	0	0	1	1	0	1	1	1
FREEC	0	0	0	0	0	0	1	0	0	1	1	1
CNV-seq	0	0	0	0	0	0	0	1	1	1	1	1
CNVnator	0	0	1	0	0	0	1	1	1	1	1	1
readDepth	0	0	0	0	0	1	0	0	0	0	1	1
CLImAT	0	0	0	0	1	1	0	0	0	0	1	1
Number of Blocks	24544	9272	14435	24508	785590	11520	4865	8884	14135	8821	11800	5678

The integer 1 *represents common CNV blocks that can be detected by the corresponding method, while* 0 *means it is not detected. The proposed ADM-CNV could detect as many as* $24,564$ *blocks, all being marked in DGV. It ranked the second position among the seven methods.*

TABLE 3
F-Scores of Top 10 CNVs Detected by Each Method

CNVs	1	2	3	4	5	6	7	8	9	10	*Avg.* $\pm$ *std.*
CNV-TV	0.88	0.83	0.74	0.70	0.68	0.68	0.64	0.58	0.56	0.55	0.684 ± 0.011
FREEC	0.81	0.77	0.72	0.71	0.70	0.59	0.59	0.54	0.51	0.44	0.6380 ± 0.121
CNV-seq	0.88	0.83	0.74	0.69	0.68	0.68	0.64	0.58	0.56	0.55	0.683 ± 0.110
ADM-CNV	0.90	0.88	0.88	0.85	0.78	0.78	0.77	0.77	0.75	0.73	0.81 ± 0.03
readDepth	0.8845	0.8708	0.7956	0.7169	0.7109	0.6587	0.6564	0.6023	0.4618	0.4304	0.679 ± 0.153
CNVnator	0.9893	0.9723	0.9539	0.9323	0.8982	0.8919	0.8138	0.7589	0.7392	0.6040	0.8554 ± 0.0138
CLImAT	0.9158	0.8792	0.7514	0.7398	0.7120	0.6687	0.6451	0.51633	0.5018	0.4707	0.6397 ± 0.03074

For visual comparison, the best and the second best value were highlighted in red and blue, respectively.

between two intervals and thus takes values between 0 and 1. A low score means the overlap quality is poor, while a higher score implies a better overlap. The formula for calculating the F-score is $F = \frac{2PR}{P+R}$, where P is the precision (percent of detected CNVs that overlap with the ground truth CNVs from DGV), and R is the recall (percent of the ground truth CNVs which overlap with the detected CNVs). For visual comparison, the best and the second best value are highlighted in red and blue, respectively. Table 3 lists the top 10 F-score CNVs detected by each method. In most cases, the proposed method achieved superior or subsuperior performance in terms of F-scores. The method of CNVnator achieved the top ranking. The performance of the proposed method outperformed CLImAT except at the first column. Another prominent characteristic of the proposed ADM-CNV is its stability. Among the top 10 CNVs detected, the proposed method achieved the next best F-score for every CNV, and the smallest standard deviation.

To have a comprehensive understanding on the F-scores after each methods, the average F-scores distribution was further calculated. It reflects the averaged performance of the CNV detection methods. The CNVs detected by the aforementioned methods were categorized into 10 classes $(0-0.1, 0.1-0.2, \ldots, 0.9-1)$ by its F-score. The results, shown in Table 4, demonstrated that the performance of ADM-CNV ranked the third, suboptimal to CNVnator and CLImAT at the category of 0.9–1.0. It achieved superior performance than CNVnator at the category of 0.0–0.1. In the other categories, the ADM-CNV always outperformed the CNV-TV and had comparable performance in comparison with the others.

We have illustrated that CNVnator achieved the best performance in Table 3. However, it obtained the second lowest (highlighted in red and blue) performance among all the methods when viewing its average distribution at class $0.0-0.1$ in Table 4. The main reason is that CNVnator could detected almost all deletions, but it generated large false positives for duplications. Such inaccuracy resulted in a low precision value. For the class $0.9-1.0$, one will find out that the method of CLImAT, the proposed method and CNVnator were the top three model. In summary, CLImAT ranked the top one by achieving even distribution of F-score. The proposed method performed suboptimal while CNV-TV performed less satisfactory among all the tested methods.

Finally, the proposed ADM-CNV requires three user-defined parameters, including cut-off value for breakpoint

TABLE 4
Average Distribution (in Percentage) of F-Scores of Detected CNVs

F-score	0.0–0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1.0
CNV-TV	0.7837	0.0898	0.0408	0.0204	0.0163	0.0204	0.0163	0.0041	0.0082	0.0
FREEC	0.5	0.14	0.05	0.05	0.05	0.1	0.0	0.1	0.024	0.0
CNV-seq	0.54	0.21	0.08	0.05	0.03	0.017	0.017	0.03	0.02	0.0
ADM-CNV	0.6571	0.0737	0.0641	0.0417	0.0481	0.0417	0.0288	0.0321	0.0096	0.0032
readDepth	0.6389	0.1389	0.0278	0.0417	0.0417	0.0	0.0417	0.0471	0.0278	0.0
CNVnator	0.7143	0.0476	0	0.0159	0.0476	0.0159	0.0159	0.00317	0.00467	0.0635
CLImAT	0.4483	0.0345	0.0690	0.0345	0.1034	0.0690	0.0690	0.1034	0.0345	0.0345

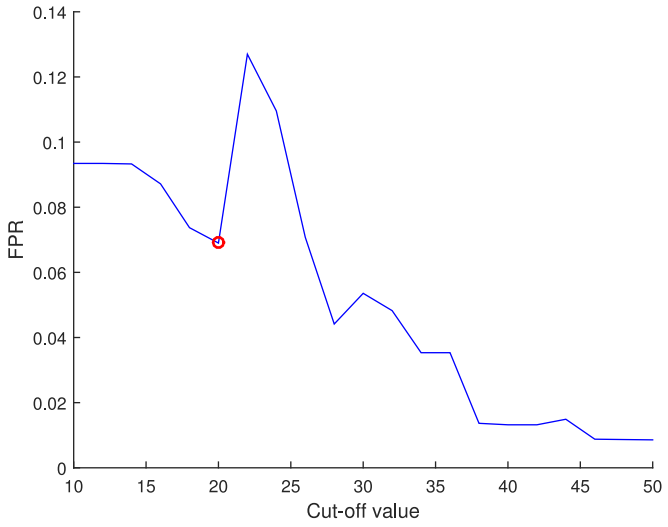

Fig. 5. False positive rates to detect copy-number alterations is considered as a function of the cut-off value. The red circle shows that when the cut-off value is 0.2, the genome-wide false positive rate is minimized. Thus, the cut-off value is set to be 0.2.

detection, the window size in which the read counts are calculated, and the overlap length of each sliding window. Through our experiments, these three parameters were optimized by a rule of thumb to minimize the false positive rate in detecting CNVs. A typical example is shown in Fig. 5. The false positive rates to detect copy number alterations was considered as a function of the cut-off value, and was plotted in Fig. 5. It can be observed that, when the cut-off value is 0.2, the genome-wide false positive rate is the smallest. Thus, the calling criteria was set at 0.2.

5 Concluding Remarks

We have proposed a model for identifying CNVs from raw NGS profiles by formulating the problem into a changing point detection optimization. The new formulation possesses a unique and fundamental characteristic distinct from its peers, in which the two constraints sparsity and smoothness of the reconstructed copy number are considered. An exact numerical solution for the convex formulation has been provided by using the classical ADM framework to guarantee a global optimal solution. Another superior characteristic of the numerical method lies in its efficiency in avoiding the matrix inversion operation, which is commonly used in signal recovering problems.

We have demonstrated the capability of the proposed method to separate CNVs from other variations in wide data types, including synthetic, simulated and empirical sequencing data. The RD ratios obtained by our method demonstrate sparsity and smoothness, thus accurately identifying CNVs.

Appendix: ADM and its Convergence

The ADM is designed to solve a generally well-structured optimization problem

$$\min f_1(\boldsymbol{x}) + f_2(\boldsymbol{B}\boldsymbol{x}), \tag{10}$$

where $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^p \to \mathbb{R}$ are convex functions. $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ is a transformation matrix with full column rank.

Given a Lagrangian multiplier $\boldsymbol{\xi}$, the augmented Lagrangian function for the above function is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\xi}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{y}) + \frac{\gamma}{2}\|\boldsymbol{B}\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{11}$$

The above minimization problem could be addressed solved by alternating solving the following sub-problems sequentially:

$$\begin{cases} \boldsymbol{x}^{k+1} & \in \arg\min \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^k, \boldsymbol{\xi}^k) \\ \boldsymbol{y}^{k+1} & \in \arg\min \mathcal{L}(\boldsymbol{x}^{k+1}, \boldsymbol{y}, \boldsymbol{\xi}^k) \\ \boldsymbol{\xi}^{k+1} & = \boldsymbol{\xi}^k - \beta(\boldsymbol{B}\boldsymbol{x}^{k+1} - \boldsymbol{y}^{k+1}). \end{cases} \tag{12}$$

Theorem 1 (Eckstein-Bertsekas [35]. *Consider problem (10), $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R}^p \to \mathbb{R}$ are closed, proper, convex functions. For arbitrary $\beta > 0$ and $\boldsymbol{y}_0, \lambda_0 \in \mathbb{R}^p$, if there exist two convergent sequences of $\{\eta^k \geq 0, k = 0, 1, \ldots\}$ and $\{\rho^k \geq 0, k = 0, 1, \ldots\}$, such that there are three sequences $\{\boldsymbol{x}^k \in \mathbb{R}^d, k = 0, 1, \ldots\}, \{\boldsymbol{y}^k \in \mathbb{R}^p, k = 0, 1, \ldots\}, \{\boldsymbol{\xi}^k \in \mathbb{R}^p, k = 0, 1, \ldots\}$ that satisfy*

$$\|\boldsymbol{x}^{k+1} - \arg\min_{\boldsymbol{x}} f_1(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{B}\boldsymbol{x} - \boldsymbol{y}^{k+1} - \boldsymbol{\xi}^k\|_2^2\| \leq \eta^k$$
$$\|\boldsymbol{y}^{k+1} - \arg\min_{\boldsymbol{y}} f_2(\boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{B}\boldsymbol{x}^{k+1} + \boldsymbol{y} - \boldsymbol{\xi}^k\|_2^2\| \leq \rho^k$$
$$\boldsymbol{\xi}^{k+1} = \boldsymbol{\xi}^k - \beta(\boldsymbol{B}\boldsymbol{x}^{k+1} - \boldsymbol{y}^{k+1}).$$

Then, if Eq. (10) has a solution $\boldsymbol{x}^*$, it follows that $\boldsymbol{x}^k \to \boldsymbol{x}^*$.

Acknowledgments

The work described in this paper was partially supported by grants from the National Nature Science Foundation of China (NSFC) with grant: 61272366 and UIC internal grant. Y. M. Cheung is a corresponding author.

References

[1] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Rev. Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[2] C. Lee, A. J. Iafrate, and A. R. Brothman, "Copy number variations and clinical cytogenetic diagnosis of constitutional disorders," *Nature Genetics*, vol. 39, no. 7, pp. S48–S54, 2007.

[3] H. Stefansson, D. Rujescu, S. Cichon, O. P. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, et al., "Large recurrent microdeletions associated with schizophrenia," *Nature*, vol. 455, no. 7210, pp. 232–236, 2008.

[4] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. Lee, J. Hicks, S. Spence, A. Lee, K. Puura, T. Lehtimaki, D. Ledbetter, P. Gregersen, J. Bregman, J. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. King, D. Skuse, D. Geschwind, T. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, no. 5823, pp. 445–449, 2007.

[5] D. Pinkel and D. G. Albertson, "Array comparative genomic hybridization and its applications in cancer," *Nature Genetics*, vol. 37, pp. S11–S17, 2005.

[6] M. L. Metzker, "Sequencing technologiesłthe next generation," *Nature Rev. Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[7] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Res.*, vol. 19, no. 9, pp. 1586–1592, 2009.

[8] J. Duan, J. Zhang, H. Deng, and Y. Wang, "Comparative studies of copy number variation detection methods for next generation sequencing technologies," *Plos One*, vol. 8, no. 3, p. e59128, 2013.

[9] A. Abyzov, A. Urban, M. Snyder, and M. Gerstein, "Cnvnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing," *Genome Res.*, vol. 21, no. 6, pp. 974–984, 2011.
[10] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinf.*, vol. 10, no. 1, p. 80, 2009.
[11] A. Magi, "Bioinformatics for next generation sequencing data," *Genes*, vol. 1, no. 2, pp. 294–307, 2010.
[12] A. Abyzov and M. Gerstein, "Age: Defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision," *Bioinf.*, vol. 27, no. 5, pp. 595–603, 2011.
[13] P. Medvedev, M. Stanciu, and M. Brudno, "Computational methods for discovering structural variation with next-generation sequencing," *Nature Methods*, vol. 6, no. 11, pp. S13–S20, 2009.
[14] C. Alkan, J. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. Kitzman, C. Baker, M. Malig, O. Mutlu, S. Sahinalp, R. Gibbs, and E. Eichler, "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nat. Genet.*, vol. 41, no. 10, pp. 1061–1067, 2009.
[15] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, "Detecting common copy number variants in high-throughput sequencing data by using jointslm algorithm," *Nucleic Acids Res.*, vol. 39, no. 10, p. e65, 2011.
[16] J. F. Sathirapongsasuti, H. Lee, B. A. J. Horst, G. Brunner, A. J. Cochran, S. Binder, J. Quackenbush, and S. F. Nelson, "Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv," *Bioinf.*, vol. 27, no. 19, pp. 2648–2654, 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/27/19/2648.abstract
[17] V. Plagnol, J. Curtis, M. Epstein, K. Y. Mok, E. Stebbings, S. Grigoriadou, N. W. Wood, S. Hambleton, S. O. Burns, A. J. Thrasher, D. Kumararatne, R. Doffinger, and S. Nejentsev, "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling," *Bioinf.*, vol. 28, no. 21, pp. 2747–2754, 2012.
[18] Y. Zhenhua, L. Yuanning, Y. Shen, M. Wang, and A. Li, "Climat: Accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data," *Sci. Found. China*, vol. 30, no. 18, pp. 2576–2583, 2015.
[19] Y. Zhang, Z. Yu, R. Ban, H. Zhang, F. Iqbal, A. Zhao, L. Ao, and Q. Shi, "Deanncnv: A tool for online detection and annotation of copy number variations from whole-exome sequencing data," *Nucleic Acids Res.*, vol. 43, no. W1, pp. 289–294, 2015.
[20] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, "Penncnv: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Res.*, vol. 17, no. 11, pp. 1665–1674, 2007.
[21] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Royal Statistical Soc.: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.
[22] A. Rinaldo, et al., "Properties and refinements of the fused lasso," *The Ann. Statists*, vol. 37, no. 5B, pp. 2922–2952, 2009.
[23] 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.
[24] J. Qian and J. Jia, "On stepwise pattern recovery of the fused lasso," *Comput. Statist. Data Anal.*, vol. 94, pp. 221–237, 2016.
[25] T. B. Arnold and R. J. Tibshirani, "Efficient implementations of the generalized lasso dual path algorithm," *J. Comput. Graphical Statists.*, vol. 25, no. 1, pp. 1–27, 2016.
[26] J. Biesinger and Y. Chen, "Solving generalized FLSA with ADMM algorithm for copy number variation detection in human," Research Report, pp. 1–7, 2011.
[27] J. Duan, J.-G. Zhang, H.-W. Deng, and Y.-P. Wang, "CNV-TV: A robust method to discover copy number variation from short sequencing reads," *BMC Bioinf.*, vol. 14, no. 1, p. 150, 2013.
[28] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al., "The sequence alignment/map format and samtools," *Bioinf.*, vol. 25, no. 16, pp. 2078–2079, 2009.
[29] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," *Genome Biol*, vol. 10, no. 3, p. R25, 2009.
[30] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, G. Getz, et al., "Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol.*, vol. 12, no. 4, p. R41, 2011.
[31] M. K. Ng, P. Weiss, and X. Yuan, "Solving constrained total-variation image restoration and reconstruction problems via alternating direction methods," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2710–2736, 2010.
[32] Z. Zhang, K. Lange, R. Ophoff, and C. Sabatti, "Reconstructing dna copy number by penalized estimation and imputation," *The Ann. Appl Statist.*, vol. 4, no. 4, p. 1749, 2010.
[33] Z. Zhang, K. Lange, and C. Sabatti, "Reconstructing dna copy number by joint segmentation of multiple sequences," *BMC Bioinf.*, vol. 13, no. 1, p. 205, 2012.
[34] X. Zhou, C. Yang, X. Wan, H. Zhao, and W. Yu, "Multisample acgh data analysis via total variation and spectral regularization," *IEEE/ACM Trans Comput. Biology Bioinf.*, vol. 10, no. 1, pp. 230–235, 2013.
[35] J. Eckstein and D. P. Bertsekas, "On the douglasłrachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 1-3, pp. 293–318, 1992.
[36] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., "Least angle regression," *The Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[37] J. Duan, J. Zhang, J. Lefante, H. Deng, and Y. Wang, "Detection of copy number variation from next generation sequencing data with total variation penalized least square optimization," in *Proc. IEEE Int. Conf. Bioinf. Biomed. Workshops*, 2011, pp. 3–12.
[38] X. Zhou, J. Liu, X. Wan, and W. Yu, "Piecewise-constant and low-rank approximation for identification of recurrent copy number variations," *Bioinf.*, vol. 30, no. 14, pp. 1943–1949, 2014.
[39] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.
[40] V. Boeva, A. Zinovyev, K. Bleakley, J. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization," *Bioinf.*, vol. 27, no. 2, pp. 268–269, 2011.
[41] C. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic, "Readdepth: A parallel r package for detecting copy number alterations from short sequencing reads," *Plos One*, vol. 6, no. 1, p. e16327, 2011.
[42] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome." *Nature Genetics*, vol. 36, no. 9, pp. 949–951, 2004.

Yue Zhang received the MSc degree from the Harbin Institute of Technology, China, in 2003, and is currently working toward the PhD degree at the Hong Kong Baptist University. She is a lecturer in the Electrical and Information College at Jinan University. Her research interests include bioinformatics and optimization.

Yiu-ming Cheung (SM'06) received the PhD degree from the Department of Computer Science and Engineering at the Chinese University of Hong Kong. He is a full professor in the Department of Computer Science at Hong Kong Baptist University. His current research interests including machine learning, pattern recognition, visual computing, and optimization. He is the founding chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. Also, he is now serving as an associate editor of the *IEEE Transactions on Neural Networks and Learning Systems*, *Knowledge and Information Systems*, and *International Journal of Pattern Recognition and Artificial Intelligence*, among others. He is a senior member of the IEEE and ACM. More details can be found at: http://www.comp.hkbu.edu.hk/∼ymc.

Bo Xu is currently working toward the master's degree from the School of Computer Science and Engineering at the South China University of Technology, Guangzhou, China. Her current research interests include machine learning, data mining, and bioinformatics. She was the recipient of the Anjubao research scholarship. She has coauthored more than five referred papers.

Weifeng Su received the BSc degree in computer science from the Petroleum University of China, the MSc degree from the Xiamen University of China, and the PhD degree from the Hong Kong University of Science and Technology in 1995, 2002, and 2007, respectively. He is an associate professor in the Computer Science and Technology programme at BNU-HKBU-UIC. His research interests include database, deep web, data mining, machine learning, and natural language processing.