Penalized weighted low-rank approximation for robust recovery of recurrent copy number variations

Xiaoli Gao

Abstract

Background: Copy number variation (CNV) analysis has become one of the most important research areas for understanding complex disease. With increasing resolution of array-based comparative genomic hybridization (aCGH) arrays, more and more raw copy number data are collected for multiple arrays. It is natural to realize the co-existence of both recurrent and individual-specific CNVs, together with the possible data contamination during the data generation process. Therefore, there is a great need for an efficient and robust statistical model for simultaneous recovery of both recurrent and individual-specific CNVs.

Result: We develop a penalized weighted low-rank approximation method (WPLA) for robust recovery of recurrent CNVs. In particular, we formulate multiple aCGH arrays into a realization of a hidden low-rank matrix with some random noises and let an additional weight matrix account for those individual-specific effects. Thus, we do not restrict the random noise to be normally distributed, or even homogeneous. We show its performance through three real datasets and twelve synthetic datasets from different types of recurrent CNV regions associated with either normal random errors or heavily contaminated errors.

Conclusion: Our numerical experiments have demonstrated that the WPLA can successfully recover the recurrent CNV patterns from raw data under different scenarios. Compared with two other recent methods, it performs the best regarding its ability to simultaneously detect both recurrent and individual-specific CNVs under normal random errors. More importantly, the WPLA is the only method which can effectively recover the recurrent CNVs region when the data is heavily contaminated.

Keywords: Copy number variation, Fused lasso, Low-rank approximation, Recurrent copy number variation, Penalized weighted approximation

Background

Copy-number variations (CNVs) are changes in the number of copies of DNA in some genome regions. The size of those variations including both deletion and amplification can vary from size of 1kb to a complete chromosome arm. CNVs in genomic DNA have been considered as a major source of genomic variation [1, 2] and can be linked to the susceptibility or resistance to certain disease such as cancer, Alzheimer and Parkinson’s disease [3–5].

Three main types of technologies have been developed to detect CNVs: array comparative genomic hybridization (aCGH) arrays [6, 7], SNP genotyping arrays [8, 9] and genome re-sequencing [10–13]. Among these technologies, aCGHs have remained the most frequently used methods for CNVs identification and genotyping of personal CNVs [9, 14, 15], because of their accuracy and cost-effectiveness [16]. A typical aCGH experiment includes DNA labeling, hybridization, and scanning. During the experiment, the fluorescence signal intensities from both the test and sample DNA at given probes are measured. After some appropriate preprocessing procedures including normalization, the raw DNA copy number data from an aCGH experiment is generally in the form of log2 ratios of those intensities between test and reference DNA samples. Thus, a probe with a positive (negative)
Finding common or recurrent CNA regions from a noisy data remains a challenge both computationally and conceptually, not mentioning the noisy data is contaminated by both individual-specific and non-Gaussian random errors. There are much less robust methods on CNV detection from individual samples or multiple samples [39, 40]. Very recently, researchers formulated the recurrent CNV detection into a matrix decomposition problem with hidden signals being low-rank [41, 42]. For example, to address the data contamination, [41] included a individual-specific effect matrix into the model and consider the observed CNV data matrix $D$ as an addition of three matrices: low-rank true recurrent CNVs matrix $X$, individual-specific effect matrix $E$, and random noise matrix $\epsilon$. Penalization optimization are adopted to recover $X$ and $E$ iteratively. In particular, a soft threshold operator [43] is adopted to update $E$. We denote this method as RPLA in this paper to differentiate it from ours. Mohammadi et al. [44] proposed another robust recurrent CNVs detection method using a Correntropy induced metric. We denote this method as CPLA in this paper. After using a Half-Quadratic technique [37, 45], CPLA is eventually reduced into a similar optimization problem as the one for RPLA. Instead of solving $E$ using a soft threshold operator, CPLA updates $E$ using a minimizer function $\delta$ during the iteration.

As explained in the last paragraph, both RPLA and CPLA introduce an individual-specific effect matrix $E$ in the model needed to be estimated. She and Owen [46] has demonstrated by linear regression analysis that a Lasso penalty on the mean shift parameter cannot reduce both the masking (outliers are not detected) and swamping (normal observations are incorrectly identified as outliers) effects for the outlier detection. This justifies some limitations of RPLA, where $E$ plays the same role as outliers in multiple regression in [46]. Additionally, the minimizer function $\rho$ used in CPLA does not encourage the sparsity of the matrix $E$. Thus, CPLA itself does not have any ability of detecting individual-specific CNVs. This phenomenon will be further addressed in Section ‘Results and discussion’.

In this paper, we propose a novel method for robust recovery of the recurrent CNVs using a penalized weighted low-rank approximation (WPLA). Instead of using a mean shift parameter to represent each individual effect, we consider to assign a weight parameter to each probe of every sample. Thus, all the individual effects are related to a weight matrix $W$, where a weight value of 1 indicates a normal probe for a normal sample without individual-specific effect, and a weight value less than 1 indicates possible individual-specific effect occurring at this probe. We propose to shrink all individual-specific effects in the direction of the recurrent effects by penalizing the weight matrix $W$. As a result, a robust detection
of recurrent CNVs is obtained by simultaneous identification of both individual-specific CNVs and recurrent ones.

Our proposed WPLA has the following two features:

- It can perform both the recurrent CNV and individual effect detection simultaneously and efficiently;
- It has strong robustness in terms of recurrent CNV detection. When the data is heavily contaminated, WPLA performs consistently better than the two aforementioned methods (CPLA and RPLA).

The rest of the paper is organized as follows. In Section ‘Methods’, we introduce our model formulation with some properties. We also provide its computation algorithm in this section. In Section ‘Results and discussion’, we demonstrate the performance of WPLA by both synthetic data analysis in multiple scenarios and two real data analysis. Finally, we conclude our paper with some discussions in Section ‘Conclusions’.

**Methods**

**Formulation**

Suppose we have an aCGH array data from \( p \) probes of \( n \) samples. Let \( d_{ij} \) be the observed log2 intensities at probe \( j \) of sample \( i \). Then \( d_{ij} \) is a realization of the true hidden signal \( x_{ij} \) and random error \( \epsilon_{ij} \),

\[
d_{ij} = x_{ij} + \epsilon_{ij}, \quad \text{for all } 1 \leq i \leq n, 1 \leq j \leq p,
\]

where \( \epsilon_{ij} \) is assumed to have mean 0 and variance \( \sigma_{ij}^2 = \sigma^2 / w_{ij}^2 \) for all \( i \) and \( j \). Here \( 0 < w_{ij} \leq 1 \) is a weight parameter at probe \( j \) of sample \( i \). A major relaxation of model (1) from existing recurrent CNVs detection methods is that we do not restrict all random errors to be homogeneously distributed with the same variance. In fact, the variance can go to infinity when \( w_{ij} \) goes to 0.

Let \( D = (d_{ij}), X = (x_{ij}) \) and \( \varepsilon = (\epsilon_{ij}) \) be three corresponding matrices from the observation data, true hidden signals, and random noises. We can write the recurrent CNV detection problem in (1) into a a multivariate regression type model,

\[
W \cdot (D - X) = \varepsilon,
\]

where \( A \cdot B \) represents an elementary-wise product between two matrices \( A \) and \( B \). We consider to recover the hidden signal matrix \( X \) under following recurrent CNV properties:

(P1) For each sample, the hidden log2 intensities tend to be the same at nearby probes. This property is incorporated into the model by assuming each row in the hidden signal matrix, \( x_i \), to be piecewise constant with only a few breakpoints.

(P2) Most samples include recurrent CNV, and the number of unique recurrent CNV regions is small. This property is incorporated into the model by assuming the hidden signal matrix \( X \) to be low-rank.

(P3) Most probes are observed with homogeneous random errors with mean 0, some of them may be contaminated and include individual-specific effects. This feature is incorporated into the model by assuming most \( w_{ij} \)s to be 1, except a few of them being smaller than 1.

We propose to recover the hidden signal \( X \) using a penalized weighted low-rank approximation. In particular, we aim to solve an optimization problem,

\[
(\hat{X}, \hat{W}) = \arg\min_{X, W} \left\{ \frac{1}{2} \| W \cdot (D - X) \|_F^2 + \alpha_1 \| X \|_* + \alpha_2 \sum_{i=1}^n \| x_i \|_{TV} + \beta \| \log(W) \|_1 \right\}
\]

subject to \( 0 < \min_{ij} w_{ij} \leq \max_{ij} w_{ij} \leq 1 \),

where \( \| A \|_F \) is the Frobenious norm. All three penalty terms in (3) are adopted to incorporate features P1–P3 for a multiple aCGH data as follows.

- The hidden signal total variation term \( \alpha_2 \| x_i \|_{TV} = \alpha_2 \sum_{j=2}^p |x_{ij} - x_{ij-1}| \) is to enforce a piecewise constant estimation of all \( x_i \) along the sequence for all \( 1 \leq i \leq n \), where \( \alpha_2 > 0 \) is a tuning parameter controlling the number of breakpoints among all \( n \) sequences. The larger \( \alpha_2 \) is, the less number of breakpoints are encouraged. This term is to realize the above feature P1.

- The nuclear norm term \( \alpha_1 \| X \|_* = \alpha_1 \sum_{i=1}^n \sigma_i \) is adopted to realize the above feature P2 and obtain a reduced rank estimation of \( X \). Here \( \sigma_i \) for \( 1 \leq i \leq r \) are all \( r \) singular values of \( X \) and \( \alpha_1 > 0 \) is the tuning parameter controlling the effective rank of matrix \( X \). The larger \( \alpha_1 \) is, the lower rank of \( X \) with stronger recurrent properties is encouraged.

- The last \( \ell_1 \) norm penalty \( \beta \| \log(W) \|_1 = \beta \sum_{i=1}^n \sum_{j=1}^p | \log(w_{ij}) | \) is adopted to control the number of heterogeneous CNVs with individual effects. The larger \( \beta \) is, the less the individual effects is encouraged. Due to the fact that \( 1 - w_{ij} \approx \log(w_{ij}) \) when \( w_{ij} \) is close to 1, this term can be also replaced by an alternative \( \beta \| 1 - W \|_1 \), where \( 1 \) is a \( n \times p \) matrix with all elements being 1.

**Robust property**

The robust property of WPLA can be observed from its link with a redescending M-estimation. In particular, we
can associate the WPLA approximation of $X$ in (3) with a  
penalized redescending M-estimation approximation  
$$
\hat{X}_M = \arg \min_X \left\{ \sum_{i=1}^n \sum_{j=1}^p \rho \left| d_{ij} - x_{ij} \right| + \alpha_1 \| X \|_+ + \alpha_2 \sum_{i=1}^n \| x_i \|_{TV} \right\},
$$
where  
$$
\rho(t) = \begin{cases} 
\beta \log(t^2/\beta) + \beta & \text{if } |t| > \sqrt{\beta}, \\
\beta t^2 & \text{if } |t| \leq \sqrt{\beta}.
\end{cases}
$$
This $\rho(t)$ function in (5) produces strong robust property  
since $\frac{d \rho(t)}{dt}$ is approaching 0 when $t \to \infty$. An  
additional pdf file shows more details (see Section 1 in  
Additional file 1).

**Adaptive WPLA**

Considering the better performance of adaptive Lasso  
over Lasso [47], we also propose an adaptive WPLA by  
minimizing,

$$
\frac{1}{2} \| W \cdot (D - X) \|_F^2 + \alpha_1 \| X \|_+ + \alpha_2 \sum_{i=1}^n \| x_i \|_{TV} + \sum_{i=1}^n \sum_{j=1}^p \beta_{ij} | \log(w_{ij}) | \\
\text{subject to } 0 < \min_{ij} w_{ij} \leq \max_{ij} w_{ij} \leq 1,
$$

where $\beta_{ij} = \beta / \sqrt{| \log(w_{ij}^{(0)}) |}$ and $w_{ij}^{(0)}$ is an initial value  
of $w_{ij}$. In implementation, we can obtain $w_{ij}^{(0)}$'s from $x_{ij}$'s  
using (8) to be presented in Section ‘Algorithm’, where  
$x_{ij}^{(0)}$ is an initial estimate of $x_{ij}$. For example, we can obtain  
x_{ij}'s from either RPLA or CPLA. If 0 occurs at the denominator, we replace it by 0.001 by convention. In the  
next section, we will present more details on choice of initial  
values following the computation algorithm for (6).

**Algorithm**

Once $W$ is fixed, solving $X$ is a convex optimization  
problem. Due to the co-existence of both nuclear norm and  
total variation, instead of solving the optimization problem  
directly, we adopt the Alternating Direction Method  
of Multipliers (ADMM, [43]) in our algorithm. ADMM was also used in both RPLA and CPLA. We now divide  
the optimization problem in (3) in separate steps. Some  
more details on mathematical computations can be found in  
Additional file 1 (Section 3).

First, we rewrite the penalized objective function in  
model (3) as  
$$
L(X, W, Z, Y) = \frac{1}{2} \| W \cdot (D - X) \|_F^2 + \alpha_1 \| X \|_+ + \alpha_2 \sum_{i=1}^n \| z_i \|_{TV} + \beta \| \log(W) \|_1 + \langle Y, X - Z \rangle + \frac{\rho}{2} \| X - Z \|_F^2,
$$

where $Y$ is the dual variable matrix, $Z$ is an auxiliary variable  
matrix, and $\langle Y, X - Z \rangle = \text{Tr}(Y' (X - Z))$ denotes the  
inner product between $Y$ and $X - Z$. Here $\rho$ is a tuning  
parameter controlling the convergence of the algorithm ($\rho$  
is adaptively tuned during the iteration; one can refer Eq.  
(10) in Section 3.4.1 of [43] for more details).

If $\hat{X}$ is obtained, the optimization function for solving $W$  
(7) becomes  
$$
L(\hat{X}, W) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p w_{ij} (d_{ij} - \tilde{x}_{ij})^2 + \beta \sum_{i=1}^n \sum_{j=1}^p | \log(w_{ij}) |.
$$

Thus we can update $w_{ij}$ from  
$$
\begin{cases} 
\sqrt{\beta} & \text{if } |d_{ij} - \tilde{x}_{ij}| > \sqrt{\beta} \\
1 & \text{Otherwise.}
\end{cases}
$$

If $(W, \tilde{Z}, \tilde{Y})$ is obtained, solving $X$ from (7) becomes  
$$
\arg \min_{X} L(X, \tilde{W}, \tilde{Z}, \tilde{Y})
$$

An additional pdf file shows more detailed computation  
(see Section 3 in Additional file 1).

It turns out that updating $X$ in (9) becomes a matrix  
completion problem in a trace regression problem [48].  
Therefore, the solution of $X$ can be expressed explicitly.  
Denote $a_{ij} = (\rho + \tilde{w}_{ij}^2)^{1/2}$ and $b_{ij} = a_{ij}^{-2} (\tilde{w}_{ij}^2 d_{ij} + \rho \tilde{z}_{ij} - \tilde{y}_{ij})$.  
Let $A = \sum_{i=1}^n \sum_{j=1}^p A_{ij}$, where $A_{ij}$ is a $n \times p$ matrix with all  
zero elements except $a_{ij}$ being at row $i$ and column $j$. Let $B$ be a $n \times p$ matrix consists of all $b_{ij}$s. Then (9) is equivalent to  
$$
\arg \min_{X} \left\{ L(X, \tilde{W}, \tilde{Z}, \tilde{Y}) \right\}
$$

where $C = (1 + \rho)^{-1} \sum_{i=1}^n \sum_{j=1}^p a_{ij} b_{ij} A_{ij}$  
and $\alpha' = \alpha_1 / (1 + \rho)$.

An additional pdf file shows more detailed derivation  
(see Section 3 in Additional file 1). Thus solving $X$ reduces
to a soft thresholding of singular values in the singular value decomposition of $C$. In particular, let $\{u_i\}, \{v_i\}$ and $\{\sigma_i\}$ be the left singular vectors, the right singular vectors and the singular values of $C$, respectively. We can update $X$ from

$$X = \sum_{i=1}^r (\sigma_i - \alpha_i') + u_i v_i^T,$$

(11)

where $r$ is the rank of $C$ and $(a)_+ = \max\{a, 0\}.

If $(\tilde{X}, \tilde{W}, \tilde{Y})$ is obtained, solving $Z$ from (7) becomes

$$\arg\min_Z \left\{ L(\tilde{X}, \tilde{W}, Z, \tilde{Y}) \right\}$$

$$= \arg\min_Z \left\{ \frac{1}{2} \| \tilde{X} + \tilde{Y} - Z \|_F^2 + \frac{\alpha_2}{\rho} \sum_{i=1}^n \| z_i \|_{TV} \right\}$$

(12)

Thus for each $1 \leq i \leq n$, we can update $z_i$ using the fused lasso signal approximation algorithm in [49].

We summarize the above iterations in the following Algorithm 1.

**Algorithm 1** The algorithm of WPLA in (3)

1. **Input** $D$
2. Initialize $W$ and $X$
3. **While** not converged do
   - update $W$ from (8)
   - update $X$ from (11)
   - update $Z$ from (12) using the fused lasso signal approximator (FLSA)
   - update $Y \leftarrow Y + \rho (X - Z)$
4. **end while**
5. **Output** $X$ and $W$

The objective function in (3) is a bi-convexity optimization problem. Two matrices $W$ and $X$ are alternately updated with the other held fixed until reaching convergence. However, the initial points $X^{(0)}$ and $W^{(0)}$ may affect the final solution. Rousseeuw and Driessen [50] suggests a multi-start iterative strategy in general. In our applications, we found initial values generated from both RPLA and CPLA work very well.

**Tuning parameter selection**

Tuning parameter selection is always challenging in penalized optimization. It is computationally expensive to

| Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|--------|---------|---------|---------|---------|---------|---------|---------|---------|
| 1-10   | +       | +       |         |         |         |         |         |         |
| 11-20  | +       | +       |         |         |         |         |         |         |
| 21-30  | +       | +       |         |         |         |         |         |         |
| 31-40  | +       | +       |         |         |         |         |         |         |
| 41-50  | +       | +       |         |         |         |         |         |         |
| 51-60  | +       | +       |         |         |         |         |         |         |
| 61-70  | +       | +       |         |         |         |         |         |         |
| 71-80  | +       | +       |         |         |         |         |         |         |
| 81-90  | +       | +       |         |         |         |         |         |         |
| 91-100 | +       | +       |         |         |         |         |         |         |

| Case 2 & 8 | Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|------------|--------|---------|---------|---------|---------|---------|---------|---------|
| 1-10       | -      | -       |         |         |         |         |         |         |
| 11-20      | -      | -       |         |         |         |         |         |         |
| 21-30      | +      | +       |         |         |         |         |         |         |
| 31-40      | +      | +       |         |         |         |         |         |         |
| 41-50      | +      | +       |         |         |         |         |         |         |
| 51-60      | +      | +       |         |         |         |         |         |         |
| 61-70      | +      | +       |         |         |         |         |         |         |
| 71-80      | +      | +       |         |         |         |         |         |         |

| Case 3 & 9 | Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|------------|--------|---------|---------|---------|---------|---------|---------|---------|
| 1-10       | -      | -       |         |         |         |         |         |         |
| 11-20      | -      | -       |         |         |         |         |         |         |
| 21-30      | +      | +       |         |         |         |         |         |         |
| 31-40      | +      | +       |         |         |         |         |         |         |
| 41-50      | +      | +       |         |         |         |         |         |         |

| Case 4 & 10 | Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|-------------|--------|---------|---------|---------|---------|---------|---------|---------|
| 1-10        | +      | +       |         |         |         |         |         |         |
| 11-20       | +      | +       |         |         |         |         |         |         |
| 21-30       | +      | +       |         |         |         |         |         |         |
| 31-40       | +      | +       |         |         |         |         |         |         |
| 41-50       | +      | +       |         |         |         |         |         |         |

| Case 5 & 11 | Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|-------------|--------|---------|---------|---------|---------|---------|---------|---------|
| 1-10        | +      | +       |         |         |         |         |         |         |
| 11-20       | +      | +       |         |         |         |         |         |         |
| 21-30       | +      | +       |         |         |         |         |         |         |
| 31-40       | +      | +       |         |         |         |         |         |         |
| 41-50       | +      | +       |         |         |         |         |         |         |

| Case 6 & 12 | Sample | [76, 85] | [86, 95] | [96, 105] | [106, 125] | [76, 85] | [86, 95] | [96, 105] | [106, 125] |
|-------------|--------|---------|---------|---------|---------|---------|---------|---------|
| 1-10        | +      | +       |         |         |         |         |         |         |
| 11-20       | +      | +       |         |         |         |         |         |         |
| 21-30       | +      | +       |         |         |         |         |         |         |
| 31-40       | +      | +       |         |         |         |         |         |         |
| 41-50       | +      | +       |         |         |         |         |         |         |
extensively search an optimal \((\alpha_1, \alpha_2, \beta)\) in (7). Here we propose some strategies on selecting three types tuning parameters. This strategy works surprisingly well during the implementation of both synthetic data and two real data analysis.

1) Type 1 tuning parameters including \(\alpha_1\) and \(\alpha_2\) control the hidden recurrent copy number structure. Between them, \(\alpha_1\) is the most important one. We follow the discussions in [41] and let 
\[
\alpha_1 = \left(\sqrt{n} + \sqrt{p}\right) \hat{\sigma}, \quad \text{where} \quad \hat{\sigma} = 1.48 \text{MAD}, \quad \text{where} \quad \text{MAD} = \text{median}(|\mathbf{D} - \text{median}(\mathbf{D})|).
\]
After \(\alpha_1\) is determined, we fix \(\alpha_2 = 0.1 \alpha_1\).

2) Type 2 tuning parameter includes \(\beta\), controlling the ratio of individual CNVs (outliers). Refer to (Gao X, Fang Y: Penalized weighted least squares for outlier detection and robust regression, Under revision), we provide a Bayesian understanding of the weight parameter penalty term in a multivariate regression framework, where the robust estimation of the coefficients vector is a posterior mode if \(v = 1/w\) when \(v\) has a Type I Pareto prior distribution \(\pi(v) \propto v^{1-\beta} I(v \geq 1)\), where \(\beta_0 = 1\) is the uniform prior and \(\beta_0 = 2\) is the Jeffrey’s prior. This motivates us to select the tuning parameter \(\beta = \hat{\sigma}^2 \beta_0\) with \(\beta_0 = 1\) or 2, where \(\hat{\sigma}\) is a robust measurement of the noises’ variability. For example, we let \(\hat{\sigma} = 1.4826 \text{MAD}\) in real implementation [51]. An additional pdf file shows more details on this Bayesian understanding (see Section 2 in Additional file 1).

3) Type 3 tuning parameter \(\rho\) is the parameter controls the convergence of the algorithm. We let \(\rho = 0.1 \sigma_D\), where \(\sigma_D\) is the maximum singular value of matrix \(\mathbf{D}\), and adaptively tuned during the iteration following [43]. On the one hand, if the primal residual (the maximal singular value of \(\mathbf{X} - \mathbf{Z}\)) is 10 times of the dual residual (the maximal singular value of \(\rho(\mathbf{Z} - \tilde{\mathbf{Z}})\)), then \(\rho\) is doubled. On the

---

**Fig. 1** Sample synthetic data and its low-rank output. Column 1: Case 3; Column 2: Case 9; Column 3: Case 10; Column 4: Case 12. Row 1: Input observations; Row 2: WPLA recovery; Row 3: RPLA recovery; Row 4: CPLA recovery.
other hand, if the dual residual is 10 times of the primal residual, then $\rho$ is halved. We keep updating $\rho$ during the iteration steps in Algorithm 1 until converge.

Results and discussion

Results

Synthetic data sets

We generate multiple synthetic data sets with 50 samples and 300 probes from

$$D = X + E + \varepsilon.$$  \hspace{1cm} (13)

This means that the synthetic data is consistent with the mean shift model studied in RPLA, and our WPLA model is actually misspecified under (13). However, we will show that WPLA still performs the best among all these three methods in all numerical experiments.

We consider twelve different combinations of six recurrent CNV scenarios ($X$) discussed in [31] and two types of random errors ($\varepsilon$). All individual sparse signals ($E$) are generated similar to [41]. We summarize all details as follows.

---

**Fig. 2** Recurrent CNVs Detection Result for Case 1-6, where random noise is Gaussian. Row 1-6 are for Case 1-6, respectively. Column 1: ROC curves (the higher, the better); Column 2: FDR curves (the lower, the better). Black: WPLA output; Red: RPLA output; Blue: CPLA output.
(Recurrent CNVs: X) Six different types recurrent CNV regions are listed in Table 1. Here all gains (+) and losses (-) have the true signal value of 1 and -1, respectively.

(Individual CNVs: E) Each sample includes an individual CNV region with a length of 20 probes. This region is randomly located outside of recurrent regions, with intensities randomly selected from {-2, -1, 1, 2}.

(Random error: ε) Two types of noises are considered for all probes: 1) Case 1-6 have Gaussian noises with \( \sigma = 0.3 \); 2) Case 7-12 have contaminated noises from \( 0.5t(1) \), where \( t(1) \) indicates a \( t \) distribution with degrees of freedom of 1. We take \( t(1) \) distribution as a heavily contaminated example since \( t(1) \) distribution is often used for quantifying the thicker tails of genetic data as mentioned in [52, 53]. In addition, no finite variance for \( t(1) \). Thus Case 7-12 are corresponding heavily contaminated scenarios parallel to Case 1-6.

In Fig. 1, we provide some heat maps of 4 synthetic data sets for Case 3 (Column 1), Case 9 (Column 2), Fig. 3 Recurrent CNVs Detection Result for Case 7-12, where data is heavily contaminated. Row 1-6 are for Case 7-12, respectively. Column 1: ROC curves (the higher, the better); Column 2: FDR curves (the lower, the better). Black: WPLA output; Red: RPLA output; Blue: CPLA output. Some of CPLA curves disappear since CPLA does not produce any recurrent CNVs region when data is heavily contaminated.
Case 10 (Column 3), and Case 12 (Column 4). The input data matrix is plotted in Row 1, with corresponding recurrent CNV regions recovery from WPLA, RPLA and CPLA plotted in Row 2, Row 3, and Row 4, respectively. Under normal noises with individual-specific CNVs in Case 3, both WPLA and CPLA provide almost perfect mappings of the true hidden low-rank matrix $X$. Although RPLA provides a slightly more noisy output than WPLA and CPLA, it can still recover the recurrent CNV region reasonably well. However, under a parallel heavily contaminated case in Case 9, both CPLA and RPLA lose their abilities of detecting recurrent CNVs. On the one hand, CPLA provides a zero estimation for $X$ and treat the entire data as random noises. On the other hand, RPLA provides a considerably noisy estimation of $X$. Compared with both RPLA and CPLA, WPLA provides a much more efficient recovery of those recurrent CNV regions, although there may have some false positives around those underlying CNV regions (See all plots at column 2 from Fig. 1). Similarly, WPLA shows dramatic improvement from RPLA and CPLA in detecting recurrent CNVs under Case 10 and 12, where the underlying recurrent CNV regions are much more complicated.

![Graphs](image-url)
To further evaluate the performance of WPLA in terms of recurrent CNV detection, we compute both the true positive rate (TPR) and the false positive rate (FPR), and estimate the false discovery rate (the false positive number out of total number of recurrent CNVs detected). In particular, we report both the receiver operation characteristic (ROC) curves (TPR vs FPR) and the false discovery rate (FDR = FP/(TP+FP)) based upon a sequence of cut-off values for CNVs. ROC and FDR curves from Case 1-6 and Case 7-12 are plotted and compared in Figs. 2 and 3, respectively. In each figure, all left panels are ROC curves (the higher, the better), all right panels are FDR curves (the lower, the better).

We also plot both ROC and FDR curves regarding the individual-specific CNVs detection for Case 1-6 in Fig. 4. The comparison between WPLA and RPLA or CPLA under other cases exhibit the similar patterns and are omitted due to the space limit.

---

**Fig. 5** Pollack data analysis results. Row 1: Input observation matrix; Row 2: recurrent CNVs low-rank matrix recovery from WPLA; Row 3: recurrent CNVs low-rank matrix recovery from RPLA; Row 4: recurrent CNVs low-rank matrix recovery from CPLA; Row 5: recurrent CNVs frequency output from WPLA; Row 6: recurrent CNVs frequency output from RPLA; Row 7: recurrent CNVs frequency output from CPLA
We have the following findings. 1) When the random noise is Gaussian, all three methods perform well in terms of the recurrent CNVs detection; there is still some advantages of WPLA and CPLA over RPLA by producing smaller false discovery rate. 2) Among all three methods, CPLA performs the worst in terms of individual CNVs detection. This is not surprising since CPLA is only designed for robust detection of recurrent CNVs, not for individual CNVs. 3) When the data is heavily contaminated in addition to the existence of individual CNVs, WPLA beats both two other methods considerably by producing much higher ROC curves and much lower FDR curves. In this situation, CPLA loses control of detecting recurrent CNVs completely and treat the input data as noise most of the time. Therefore, some of CPLA curves disappear in Fig. 3 since CPLA does not produce

![Fig. 6 Bekhouche data analysis results. Row 1: Input observation matrix; Row 2: recurrent CNVs low-rank matrix recovery from WPLA; Row 3: recurrent CNVs low-rank matrix recovery from RPLA; Row 4: recurrent CNVs low-rank matrix recovery from CPLA; Row 5: recurrent CNVs frequency output from WPLA; Row 6: recurrent CNVs frequency output from RPLA; Row 7: recurrent CNVs frequency output from CPLA]
any recurrent CNVs region when data is heavily contaminated. Overall, WPLA has strong robust properties when data is heavily contaminated; it is as efficient as existing robust methods when the random noise is normally distributed. In addition, WPLA has the ability of simultaneous detection of both recurrent and individual CNVs.

Real applications
We apply WPLA to three independent real data sets: chromosome 17 from Pollack data [54], chromosome 17 from Bekhouche data [55], and chromosome 11 from Wang data [16]. The first two data sets were also analyzed by [41]. The Pollack data consists of log2 intensities at 382 probes from 44 breast tumors samples, while the Bekhouche data is a much larger data set including 7727 probes from 173 breast tumors. Compared with the other two data sets, Wang data is the newest with the highest resolution, and has much smaller sample size with only 12 pig samples: one Asian wild boar population, six Chinese indigenous breeds, and two European commercial breeds.

![Fig. 7](image_url) Wang data analysis results. Row 1: Input observation matrix; Row 2: recurrent CNVs low-rank matrix recovery from WPLA; Row 3: recurrent CNVs low-rank matrix recovery from RPLA; Row 4: recurrent CNVs low-rank matrix recovery from CPLA; Row 5: recurrent CNVs frequency output from WPLA; Row 6: recurrent CNVs frequency output from RPLA; Row 7: recurrent CNVs frequency output from CPLA.
Due to the small sample size and high resolution, we only analyze one segment of chromosome 11 (between 62,001,001 and 71,997,810), where three CNVs regions were confirmed by the quantitative real-time PCR analysis. Thus, Wang data consists of 3766 probes from 12 samples.

Figures 5, 6 and 7 give the heatmaps of the Pollack data, the Bekhouche data, and the Wang data and their corresponding recurrent CNVs analysis results, respectively. In each figure, we plot the original data in Row 1 and report the detected recurrent CNVs regions from WPLA, RPLA, CPLA in Row 2-4. Corresponding gains and losses ratios out of all samples profiles are also reported in Row 5-7 in each figure, where a CNV is claimed if the estimation $|\hat{x}_{ij}| > 0.225$.

For the Pollack data and Wang data, three methods produce more discrepant CNV regions, where results from WPLA are smoother than two other methods. For the Bekhouche data, all three methods produce more consistent recurrent CNV regions. Overall, among all detected recurrent CNVs regions, WPLA turns to produce higher frequency ratio among all samples, which is reasonable for recurrent CNVs detection. For example, WPLA detects the recurrent CNVs region where gene ERBB2 and C17orf37 are located (at around probe 3460). This result is consistent with scientific discoveries since both of those two genes are claimed to be related breast cancer [41, 56]. It is worthwhile to point it out that WPLA detects this recurrent region at a much higher frequency ratio than both RPLA and CPLA.

Discussion
All numerical experiments have been done under fixed tuning parameters. There is still some potential improvement if all parameters are optimally tuned. However, the computation must be much more expensive. It is worthwhile to investigate some more effective ways of tuning parameter selection methods in future studies.

Conclusions
In this paper, we propose a novel robust method for recurrent copy number variation detection. This method is unique by assigning a weight parameter to each probe of every sample. Thus, all the individual effects are related to a weight matrix $W$, which is estimated data adaptively, together with the low-rank approximation. As a result, a robust detection of recurrent CNVs is obtained by shrinking all weights from some small values (because of individual-specific effects) to 1 (no individual effects). This proposed method has two important results: efficient detection of both recurrent CNVs and individual-specific CNVs, strong robustness in dealing with severe data contamination.

We have applied the proposed method to three real data sets and twelve synthetic data sets generated from six different types of recurrent CNVs associated with either normal random errors or heavily contaminated errors. The numerical experiment has demonstrated its superior performance of recovering recurrent CNV patterns from raw data under different scenarios. Compared with two other recent methods, it has the best ability of simultaneous detection of both recurrent and individual-specific CNVs under normal random errors. More importantly, the proposed method is the only one which can effectively recover the recurrent CNVs region when the data is heavily contaminated.

Additional file

Additional file 1: Supplementary Material. Supplementary Material is also available online under the name of “wccna_suppl2.pdf” and PDF format. This Supplementary Material provides additional proofs and some more mathematical details associated with the proposed method. In particular, there are three sections in the Supplementary Material. Section 1 is on the link between WPLA and a redescending M-estimation; Section 2 is on Bayesian understanding of WPLA; Section 3 is on more detailed derivation of some equations in Algorithm 1. (PDF 185 kb)

Competing interests
The author declares that there is no competing interests.

Authors’ contributions
XG conducted the entire investigation, including the model development and computation and the manuscript preparation. The author has read and approved the manuscript.

Acknowledgements
The author gratefully acknowledges Simons Foundation (#359337) and UNC Greensboro (New Faculty Grant) for their support in this Project.

Received: 22 September 2015 Accepted: 23 November 2015
Published online: 10 December 2015

References
1. Iafrate AJ, Feuk L, Rivera MN, Listernik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36:949–51. doi:10.1038/ng14169.
2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin S, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305:529–34. doi:10.1126/science.10989185683.
3. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1(62) doi:10.1186/gm62.
4. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007;39:37–42.
5. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature. 2006;444:444–54.
6. Yliraja B, van den Jussel P, Carvalho B, Brakenhoff RH, Meijer GA. BAC to the future! or oligonucleotides: a perspective for microarray comparative genomic hybridization (array CGH). Nucleic Acids Res. 2006;34:445–50.
7. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet. 2005;37:11–17.
8. Huang J, Wei W, Chen J, Zhang J, Liu G, Qi X, et al. CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinforma. 2006;7(83) doi:10.1186/1471-2105-7-83.
9. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007;39:16–21.
