Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 9436582, 11 pages
https://doi.org/10.1155/2021/9436582

Research Article

An Efficient Algorithm for the Detection of Outliers in Mislabeled Omics Data

Hongwei Sun,[1,2] **Jiu Wang**,[1] **Zhongwen Zhang**,[1] **Naibao Hu**,[1] **and Tong Wang**[2]

[1]*Department of Health Statistics, School of Public Health and Management, Binzhou Medical University, Yantai City, Shandong 264003, China*
[2]*Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan City, Shanxi 030001, China*

Correspondence should be addressed to Hongwei Sun; hwsun2000@163.com and Tong Wang; tongwang@sxmu.edu.cn

Received 13 May 2021; Accepted 30 November 2021; Published 22 December 2021

Academic Editor: Po-Hsiang Tsui

High dimensionality and noise have made it difficult to detect related biomarkers in omics data. Through previous study, penalized maximum trimmed likelihood estimation is effective in identifying mislabeled samples in high-dimensional data with mislabeled error. However, the algorithm commonly used in these studies is the concentration step (C-step), and the C-step algorithm that is applied to robust penalized regression does not ensure that the criterion function is gradually optimized iteratively, because the regularized parameters change during the iteration. This makes the C-step algorithm runs very slowly, especially when dealing with high-dimensional omics data. The AR-Cstep (C-step combined with an acceptance-rejection scheme) algorithm is proposed. In simulation experiments, the AR-Cstep algorithm converged faster (the average computation time was only 2% of that of the C-step algorithm) and was more accurate in terms of variable selection and outlier identification than the C-step algorithm. The two algorithms were further compared on triple negative breast cancer (TNBC) RNA-seq data. AR-Cstep can solve the problem of the C-step not converging and ensures that the iterative process is in the direction that improves criterion function. As an improvement of the C-step algorithm, the AR-Cstep algorithm can be extended to other robust models with regularized parameters.

1. Introduction

The first challenge presented by omics data is the high dimension, which far exceeds the sample size. The second challenge is the presence of noise in the omics data. This noise may be caused by misdiagnosis, mislabelling, recording errors, technical problems in the laboratory, or sample heterogeneity [1, 2]. Penalized regression is a common method to solve the problem of variable selection and prediction for a high-dimensional dataset. It has been applied to omics data such as gene expression (4), GWAS [3], and DNA methylation [4]. However, the outliers in the data make the estimation of penalized regression inaccurate, so biomarkers cannot be properly screened. Additionally, the identification and further investigation of these outliers can correct the errors during the experiment or investigation. Therefore, it is very important to develop robust statistical methods for penalized regression.

A robust estimation method, least trimmed square (LTS), was proposed by Rousseeuw [5]. LTS is highly robust to outliers in both the response and predictors. It is effective for identifying outliers and can solve the problem of the masking phenomenon caused by the coexistence of multiple outliers [5, 6]. Alfons et al. [6] applied LTS to LASSO-type penalized linear regression to solve the problem of robust high-dimensional variable selection when the dependent variable is quantitative data. Kurnaz et al. [7] applied LTS to elastic net- (EN-) type penalized linear and logistic regression to solve the problem of robust high-dimensional variable selection when the dependent variable is quantitative and binary data (enetLTS).

Both studies adopted the concentration step (C-step) in the FAST-LTS algorithm proposed by Rousseeuw and Van Driessen [8]. The basic ideas were an inequality involving order statistics and sums of squared residuals. This inequality guarantees that the criterion function declines monotonically

as the iteration progresses. However, when it is applied to penalized regression based on trimming, the inequality does not necessarily hold due to the change of the regularized parameters. Thus, the criterion function cannot be guaranteed to decrease. Through our previous simulation study [9], we have found that enetLTS is effective in identifying mislabeled samples in high-dimensional data with mislabeled error. However, it is also found that for a dataset with $n = 500$, $p = 1{,}000$, and an outlier ratio of 10%, it takes nearly 2 hours (Intel Core i7-6500U @2.50GHz) to run enetLTS once. For the omics data in real data analysis with $n = 924$ and $p = 19690$, enetLTS running time is about 77.8 hours (Intel Xeon Silver 4112 @2.60GHZ), which obviously does not meet the requirements for efficient data processing.

Therefore, the C-step algorithm needs to be improved to adapt to high-dimensional data. In this study, the AR-Cstep algorithm is proposed to solve the estimation of robust penalized regression based on trimming, which combines the C-step algorithm with the acceptance-rejection algorithm proposed by Chakraborty and Chaudhuri [10] . Two algorithms are compared in terms of variable selection and outlier identification accuracy and computation speed in simulation study. An RNA-seq dataset for triple negative breast cancer (TNBC) [1] that contains 28 samples with discordant labels obtained from different tests (immunohistochemical (IHC) method or fluorescence in situ hybridization (FISH)) is used to illustrate the application of the two algorithms.

The structure of this paper is as follows: In results section, simulation experiments are described that compare the MTL-EN (elastic net-type maximum trimmed likelihood) estimation using the AR-Cstep algorithm with enetLTS. The results of enetLTS and MTL-EN applied to a triple negative breast cancer (TNBC) RNA-seq dataset are compared. Then, the results are discussed and concluded.

In this article, a robust penalized logistic regression model based on trimming is introduced in Section 2. And the AR-Cstep algorithm is proposed and described in Section 3. In Section 4, simulation experiments are described that compare the MTL-EN (elastic net-type maximum trimmed likelihood) estimation using the AR-Cstep algorithm with enetLTS. The results of enetLTS and MTL-EN applied to a triple negative breast cancer (TNBC) RNA-seq dataset are compared in Section 5. We conclude with a discussion in Section 6 and a conclusion in Section 7.

2. Robust Penalized Logistic Regression Model Based on Trimming

Kurnaz et al. [7] proposed an EN-type penalized logistic regression based on trimming.

$$\boldsymbol{\beta}\wedge^{\text{enetLTS}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{l=1}^{h} d\left(y_{i_l}, x_{i_l}'\boldsymbol{\beta}\right) + h\lambda P_{\alpha}(\boldsymbol{\beta}), \quad (1)$$

where $d(y_{i_1}, x_{i_1}'\boldsymbol{\beta}) \le d(y_{i_2}, x_{i_2}'\boldsymbol{\beta}) \le \cdots \le d(y_{i_n}, x_{i_n}'\boldsymbol{\beta})$, $i_l \in \{1, 2, \cdots, n\}$, where $d(y_{i_k}, x_{i_k}'\boldsymbol{\beta})$ is the ordered deviance. $h = \lfloor \delta n \rfloor$ ($\lfloor . \rfloor$ means rounding down to the nearest integer) and $\alpha \in [0.5, 1]$, where $1 - \delta$ is the trimmed portion. Compared with EN, enetLTS only retains h observations with the smallest deviances, whereas n-h least likely observations under the given model are excluded.

Robust penalized logistic regression model based on trimming was denoted as enetLTS (robust EN based on the LTS), and C-step algorithm was adopted. We denote it as $\boldsymbol{\beta}\wedge^{\text{enetLTS}}$ in this paper. The estimate of the same model obtained by the AR-Cstep algorithm is recorded as the EN-type maximum trimmed likelihood estimate $\boldsymbol{\beta}\wedge^{MTL-EN}$.

3. Algorithm

3.1. C-Step Algorithm. Kurnaz et al. [7] adopted the C-step algorithm in enetLTS. This algorithm was described below.

Let $Q(H; \boldsymbol{\beta})$ be the criterion function of the penalized logistic regression based on the subsample $H \subseteq \{1, 2, \cdots, n\}$, where $|H| = h$. Thus,

$$Q(H; \boldsymbol{\beta}) = \sum_{i \in H} d\left(y_i, x_i'\boldsymbol{\beta}\right) + h \sum_{j=1}^{p} \lambda \left|\boldsymbol{\beta}_j\right|. \quad (2)$$

Additionally, $\widehat{\boldsymbol{\beta}}_H$ represents $\widehat{\boldsymbol{\beta}}_H = \arg\min_{\boldsymbol{\beta}} Q(H, \boldsymbol{\beta})$.

When the regularized parameters $\lambda = \lambda_1$ and $\alpha = \alpha_1$ are fixed, at the kth step of the iteration, H_k is the current subset with h observations, and $\widehat{\boldsymbol{\beta}}_{H_k}$ is the solution of the penalized logistic regression based on H_k. The negative log-likelihood functions corresponding to n observations can be derived from $\widehat{\boldsymbol{\beta}}_{H_k}$. The subsample H_{k+1} consists of the h smallest negative log-likelihood observations, that is,

$$H_{k+1} = \{i_1, i_2 \cdots, i_h\}, \quad (3)$$

where $d(y_{i_1}, x_{i_1}'\widehat{\boldsymbol{\beta}}_{H_k}) \le d(y_{i_2}, x_{i_2}'\widehat{\boldsymbol{\beta}}_{H_k}) \le \cdots \le d(y_{i_n}, x_{i_n}'\widehat{\boldsymbol{\beta}}_{H_k})$, $i_l \in \{1, 2, \cdots, n\}$.

Thus, $Q(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_k}) \le Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k})$ can be obtained. H_{k+1} is the subset that minimizes the criterion function under the solution $\widehat{\boldsymbol{\beta}}_{H_k}$. Then, penalized logistic regression is applied to subset H_{k+1}. If $\lambda = \lambda_1$ and $\alpha = \alpha_1$ are unchanged, we get the solution $\widehat{\boldsymbol{\beta}}_{H_{k+1}}$ which minimize the solution of criterion function under the regularization parameters λ_1 and α_1. Thus $Q(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_{k+1}}) \le Q(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_k})$ holds. Therefore, when $\lambda = \lambda_1$ is fixed,

$$Q\left(H_k; \widehat{\boldsymbol{\beta}}_{H_k}\right) \ge Q\left(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_k}\right) \ge Q\left(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_{k+1}}\right). \quad (4)$$

The definition of H_{k+1} makes the first equation hold. The definition of $\widehat{\boldsymbol{\beta}}_{H_{k+1}}$ makes the second inequality hold.

For the C-step algorithm, the candidate subset H_{k+2} is constructed by sorting out h samples with the smallest negative log-likelihood contribution to $Q(H_{k+1}; \widehat{\boldsymbol{\beta}}_{H_{k+1}})$. Then, the C-step algorithm continues until $Q_m = Q_{m-1}$.

Therefore, when $\lambda = \lambda_1$ and $\alpha = \alpha_1$ remain unchanged, as the number of iterations k increases, the criterion function decreases. Because the criterion function is nonnegative and the number of subsets with sample size h is limited, the C-step algorithm must converge to the subset with the smallest criterion function after a limited number of steps.

The C-step algorithm is described in Algorithm 1, where "continueCstep" is set so that the absolute value of the difference between the likelihood functions of two iterations is less than some small value.

However, when penalized regression is performed on the subset H_{k+1}, the regularized parameters λ and α are not fixed. The regularized parameters are usually determined by data, such as by cross-validation. The regularized parameters determined for penalized regression performed on two different subsets are often different, which leads to the second inequality of [11] not necessarily being true.

A way to solve the problem is to set all λ and α values firstly. For a certain combination of λ and α, perform the C-step algorithm until convergence. Then, compare the convergent subsets under different regularized parameters, and select the subset that minimizes the criterion function. If the number of λ values is 40 and that of α values is 20, there are 800 parameter combinations. This means running the C-step algorithm 800 times, which will undoubtedly make the algorithm very slow.

3.2. AR-Cstep Algorithm. In this study, the AR-Cstep algorithm is proposed to solve the estimation of the robust penalized regression based on trimming, which combines the C-step algorithm with the acceptance-rejection algorithm, which was proposed by Chakraborty and Chaudhuri [10].

3.2.1. Acceptance-Rejection Algorithm. The acceptance-rejection algorithm is similar to that of Metropolise-Hastings in MCMC. Let H_k represent the subset at the kth step of the iteration. Then, a randomly selected sample outside of H_k replaces one of the samples in H_k to form H_{cand}. The corresponding likelihood function is obtained after penalized regression is performed on H_{cand}. If the criterion function corresponding to H_{cand} is better than that corresponding to the current subset H_k, then H_{cand} is accepted as H_{k+1} with probability one, and $H_{k+1} = H_{\text{cand}}$. Otherwise, H_{cand} is accepted as H_{k+1} with a probability of $p < 1$, so that the algorithm can escape the local optimal value.

In the acceptance-rejection algorithm, the candidate sample at each step is randomly selected from the remaining samples other than the current subset H_k. Thus, whether the candidate subset can improve the criterion function better is completely random, which leads to the slower convergence of the iteration. The advantage of this algorithm is that, whether the criterion function corresponding to the candidate subset is better than that of the current subset is examined at each step. Moreover, the subset with the optimal criterion function up to the current step is recorded at each step.

3.2.2. AR-Cstep Algorithm. The changes of the regularized parameters λ and α make the C-step algorithm hardly gradually converge to the subset with the smallest criterion function. Suppose the current subset is H_k, and we obtain $\widehat{\boldsymbol{\beta}}_{H_k}$, and corresponding criterion function $Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k}; \lambda_1, \alpha_1)$ after the penalized regression is performed on H_k. The h smallest negative log-likelihood observations constitute the subset H_{cand}, so that $Q(H_{\text{cand}}; \widehat{\boldsymbol{\beta}}_{H_k}; \lambda_1, \alpha_1) \le Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k}; \lambda_1, \alpha_1)$ holds. Then, penalized regression is performed on H_{cand}, $\widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}$ is obtained, and the corresponding regularized parameters changed to λ_2 and α_2. The corresponding criterion function $Q(H_{\text{cand}}; \widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}; \lambda_2, \alpha_2)$ of H_{cand} is not necessarily less than $Q(H_{\text{cand}}; \widehat{\boldsymbol{\beta}}_{H_k}; \lambda_1, \alpha_1)$. The AR-Cstep algorithm adds the step of comparing the criterion function of the candidate subset H_{cand} with that of the current subset H_k. If $Q(H_{\text{cand}}; \widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}; \lambda_2, \alpha_2) > Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k}; \lambda_1, \alpha_1)$, to avoid falling into a local optimum, U is a random number that follows the Bernoulli distribution with p, where $p = e^{\tau_k(\log \ell(\boldsymbol{\beta}\wedge_{H_{\text{cand}}}, H_{\text{cand}}) - \log \ell(\boldsymbol{\beta}\wedge_{H_k}, H_k))}$. If $U = 1$, then $H_{k+1} = H_{\text{cand}}$. If $U = 0$, then $H_{k+1} = H_k$, that is, no replacement. The criterion function corresponding to the initial subset is recorded as the optimal subset, that is, $Q(H_{\text{opt}}; \widehat{\boldsymbol{\beta}}_{H_{\text{opt}}}) = Q(H_0; \widehat{\boldsymbol{\beta}}_{H_0})$, and $H_{\text{opt}} = H_0$. At each step of the iteration, the criterion function $Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k})$ is compared with $Q(H_{\text{opt}}; \widehat{\boldsymbol{\beta}}_{H_{\text{opt}}})$. If $Q(H_k; \widehat{\boldsymbol{\beta}}_{H_k}) < Q(H_{\text{opt}}; \widehat{\boldsymbol{\beta}}_{H_{opt}})$, then $Q(H_{\text{opt}}; \widehat{\boldsymbol{\beta}}_{H_{opt}}) = Q(H_k; \widehat{\boldsymbol{\beta}}_{H_{kt}})$ and $H_{\text{opt}} = H_k$. H_{opt} in the last step is the solution.

To make the proportion of samples with $y = 1$ in the candidate subset H_{cand} consistent with that in the full set, the samples constituting the candidate subset H_{cand} are selected in the following manner. H_{cand} consists of h_1 observations with the smallest $d(y_i, x_i^{\prime}\widehat{\boldsymbol{\beta}}_{H_k})$ among observations with $y = 1$ (set a total of n_1 observations), and h_0 observations with the smallest $d(y_i, x_i^{\prime}\widehat{\boldsymbol{\beta}}_{H_k})$ among observations with $y = 0$ (set a total of n_0 observations), where $h_1 = \lfloor (n_1 + 1)\eta \rfloor$, and $\lfloor . \rfloor$ means round down. $1 - \eta$ is the trimming ratio and $h_0 = h - h_1$. In comparison with the acceptance-rejection algorithm, for which H_{cand} consists of samples selected randomly from the complementary set, H_{cand} of AR-Cstep is composed of observations with the smallest deviance; that is, each sample of H_{cand} contains information that improves the criterion function; hence, the algorithm converges to the subset with the optimal criterion function faster. The AR-Cstep algorithm is described in Algorithm 2.

The acceptance probability $p = e^{\tau_k(\log \ell(\boldsymbol{\beta}\wedge_{H_{\text{cand}}}, H_{\text{cand}}) - \log \ell(\boldsymbol{\beta}\wedge_{H_k}, H_k))}$. It is inversely proportional to the absolute value of the difference between the two likelihood functions $\log \ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$ and $\log \ell(\widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}, H_{\text{cand}})$. The acceptance probability p is also related to τ_k. According to $\tau_k \coloneqq \log(k + 1)/D$, the acceptance probability p is inversely proportional to k, which is the kth step of the iteration. Similar to the study of Farcomeni and Viviani [12], $D \coloneqq 0.1n(1 - \eta)$, and the acceptance probability p is inversely proportional to the sample size h of the subset. When other features remain unchanged, the larger the sample size h of the subset,

While (continueCstep)
do
Penalized logisitic regression is applied on the current subset H_k, and get

$$\widehat{\boldsymbol{\beta}}_{H_k} \coloneqq \arg\max_{\beta}\{\log \ell(\beta, H_k) - n\lambda\textstyle\sum_{j=1}^{p}|\beta_j|\}$$

For $i \in \{1, 2,\cdots,\mathrm{n}\}$, $d(y_i, x_i^T\beta)$ of every observation is got and observations are sorted according to their deviances.

$$d(y_{i_1}, x_{i_1}'\boldsymbol{\beta}) \le d(y_{i_2}, x_{i_2}'\boldsymbol{\beta}) \le \cdots \le d(y_{i_n}, x_{i_n}'\boldsymbol{\beta}), i_l \in \{1, 2,\cdots,n\}.$$

The h observations with smallest negative log-likelihood function are retained to form a subset H_{k+1}.
end

Algorithm 1: Description of C-step algorithm.

k represents the kth iteration, and r represents that the current subset has not been replaced after r iterations.
While (k<=$k$max & r<=2)
do

$$\widehat{\boldsymbol{\beta}}_{H_k} \coloneqq \mathrm{argmin}_{\beta}\textstyle\sum_{i_l \in H_k} d(y_{i_l}, x_{i_l}'\boldsymbol{\beta}) + h\lambda\sum_{j=1}^{p}|\beta_j|$$

Under $\widehat{\boldsymbol{\beta}}_{H_k}$, $d(y_i, x_i'\widehat{\boldsymbol{\beta}}_{H_k})$ corresponding to each sample is derived. The current criterion function is

$$Q(H_k\,;\widehat{\boldsymbol{\beta}}_{H_k}) = \textstyle\sum_{i_l \in H_k} d(y_{i_l}, x_{i_l}'\widehat{\boldsymbol{\beta}}_{H_k}))$$

Candidate subset $H_{cand} = \{i_1, i_2,\cdots,i_{h_0}\} \cup \{j_1, j_2,\cdots,j_{h_1}\}$, where

$$d(1, x_{i_1}'\widehat{\boldsymbol{\beta}}_{H_k}) \le d(1, x_{i_2}'\widehat{\boldsymbol{\beta}}_{H_k}) \le \cdots \le d(1, x_{i_{n_1}}'\widehat{\boldsymbol{\beta}}_{H_k}),\ i_k \text{ is the index of individuals with } y = 1.$$

$$d(0, x_{j_1}'\widehat{\boldsymbol{\beta}}_{H_k}) \le d(0, x_{j_2}'\widehat{\boldsymbol{\beta}}_{H_k}) \le \cdots \le d(0, x_{j_{n_0}}'\widehat{\boldsymbol{\beta}}_{H_k}),\ j_k \text{ is the index of individuals with } y = 0.$$

$$h_1 = \lfloor (n_1 + 1)\eta \rfloor,\ h_0 = h - h_1,\ n = n_0 + n_1 \circ$$

$$\widehat{\boldsymbol{\beta}}_{H_{cand}} \coloneqq \mathrm{argmin}_{\beta}\textstyle\sum_{i_l \in H_{cand}} d(y_{i_l}, x_{i_l}'\boldsymbol{\beta}) + h\lambda\sum_{j=1}^{p}|\beta_j|$$

Under $\widehat{\boldsymbol{\beta}}_{H_{cand}}$, $d(y_i, x_i'\widehat{\boldsymbol{\beta}}_{H_{cand}})$ corresponding to each sample is derived. The corresponding criterion function is

$$Q(H_{cand}\,;\widehat{\boldsymbol{\beta}}_{H_{cand}}) = \textstyle\sum_{i_l \in H_{cand}} d(y_{i_l}, x_{i_l}'\widehat{\boldsymbol{\beta}}_{H_{cand}}))$$

If $Q(H_{cand}\,;\widehat{\boldsymbol{\beta}}_{H_{cand}}) \le Q(H_k\,;\widehat{\boldsymbol{\beta}}_{H_k})$ then

$$H_{k+1} = H_{cand}$$

If $Q(H_{cand}\,;\widehat{\boldsymbol{\beta}}_{H_{cand}}) > Q(H_k\,;\widehat{\boldsymbol{\beta}}_{H_k})$ then

$$p = e^{\tau_k(\log \ell(\boldsymbol{\beta}\wedge_{H_{cand}}, H_{cand}) - \log \ell(\boldsymbol{\beta}\wedge_{H_k}, H_k))}$$

U is a random number that obeys the Bernoulli distribution with the parameter p.
if U=1 then

$$H_{k+1} = H_{cand}$$

else

$$H_{k+1} = H_k$$

end
end
end

Algorithm 2: Description of AR-Cstep algorithm.

the smaller the probability of being accepted. Additionally, if the current subset is not replaced after r iterations, the iteration process is stopped.

To ensure that the initial subset does not contain outliers, the sample size should be smaller. The initial subset consisted of six observations, three of which were randomly selected from groups $y = 1$ and $y = 0$, respectively. In order to make the algorithm reach the global optimal value, multiple initial subsets were selected.

First, the two-step iteration of AR-Cstep was performed on 500 initial subsets, and 500 updated subsets were obtained. Then, the 10 subsets with the smallest criterion function were retained. Then, AR-Cstep was performed on these 10 subsets until convergence. Among the 10 convergent subsets, the subset with the smallest criterion function was selected, denoted by H_{opt}. The penalized regression was performed on H_{opt}, and $\widehat{\boldsymbol{\beta}}_{\mathrm{opt}}$ was obtained.

3.2.3. Reweighted Step. In this article, we choose the subset of size $h = \lfloor n\eta \rfloor$ where $\eta = 0.75$. So $1 - \eta$ is the initial guess that less than 25% of outliers contained in the data. This is a rather conservative estimation of proportion of outliers. There may not be so many outliers in the data. Therefore, reweighted step is considered to detect outliers via $\widehat{\boldsymbol{\beta}}_{\mathrm{opt}}$. Then, these outliers are excluded, and a new subset H_{rwt} is

obtained. Then, EN-type penalized logistic regression is applied to H_{rwt} to get the solution $\widehat{\boldsymbol{\beta}}_{rwt}$. Usually, the size of H_{rwt} is larger than h, such that more samples can improve the performance of $\widehat{\boldsymbol{\beta}}_{rwt}$ compared to $\widehat{\boldsymbol{\beta}}_{\text{opt}}$. We called $\widehat{\boldsymbol{\beta}}_{rwt}$ reweighted MTL-EN (Rwt MTL-EN). To distinguish them, the unweighted $\widehat{\boldsymbol{\beta}}_{opt}$ is called Raw MTL-EN.

3.2.4. Choice of the Regularized Parameters and Standardization of Predictors. We select λ over a grid of values in the interval $(0, \lambda_{\max}]$ as discussed by Breheny and Huang [13].

$$\widehat{\lambda}_{\max} = \max_{j\in\{1,2,\cdots,p\}} n^{-1}\mathbf{X}_j'\mathbf{y}, \tag{5}$$

where $\mathbf{y}$ is the dependent variable and $\mathbf{X}_j$ is the jth independent variable. In iteration step of AR-Cstep, we take a grid with steps of size 0.05 $\widehat{\lambda}_{\max}$ and $\alpha = 0.5$ to reduce the computational burden. In the reweighted step, we take a grid with steps of size 0.01 $\widehat{\lambda}_{\max}$ of λ to derive the solution $\widehat{\boldsymbol{\beta}}_{\text{opt}}$ and $\widehat{\boldsymbol{\beta}}_{rwt}$. The choice of α is selected by cross-validation in the interval [0.1,1] with a step size of 0.1.

It would be better to standardize predictors before applying the penalized regression. Standardization mainly is aimed at eliminating the influence of dimension and quantity of a predictor. However, the mean and standard deviation computed from all sample are not robust with outliers. In the algorithm described above, penalized regression is applied to the subset in every iteration step of AR-Cstep. So we firstly, respectively, compute mean and standard deviation from subsamples. Then, we standardize all samples with this mean and standard deviation before applying penalized regression.

4. Simulation Study

4.1. Comparison of MTL-EN and enetLTS on Outlier Detection and Variable Selection. Simulation settings were the same as Sun et al. [9]. The parameter h of both enetLTS and MTL-EN was both set to $\lfloor 0.75n \rfloor$, which meant the trimmed rate is 25%. The parameters in Ensemble followed Lopes et al. [1].

In the simulation experiment, we compared the two methods enetLTS and MTL-EN using C-step and AR-Cstep algorithms, respectively. Through our previous research [9] and subsequent simulation experiments, we can see that enetLTS is good at identifying outliers. However, the FDR of its variable selection is high, and many unrelated variables are identified. When encountering mislabeled omics data, we can combine enetLTS with Ensemble. Running Ensemble on a subset of data after removing the outliers identified by enetLTS improved the variable selection accuracy. Then, we added the third method Ensemble to the simulation experiment. A detailed description of Ensemble is provided in our previous study [9].

The performances of the three methods are summarized in Figure 1.

The outlier detection accuracy of the three methods is shown in Figure 1. Here, we used two indicators Sn (sensitivity) and FPR (False Positive Rate) [14]. Sn represents the proportion of true misclassified individuals identified as misclassified ones among all true misclassified observations. FPR represents the proportion of individuals with correct labels that are wrongly categorized as misclassified ones.

The outliers identified by MTL-EN had the higher Sn than enetLTS. When the proportion of outliers were 10% and 15%, the gap between them further widened. MTL-EN FPRs were close to enetLTS. Ensemble has the lowest Sn and FPRs among the three methods. Therefore, MTL-EN had the best accuracy in identifying outliers.

The variable selection accuracy of the three methods is shown in Figure 1. PSR (Positive Selection Rate) indicates the proportion of true disease-related biomarkers identified in all true disease-related biomarkers. FDR (False Discovery Rate) represents the proportion of biomarkers that are not related to disease among all the screened biomarkers. A comprehensive indicator GM [15, 16] for the accuracy of variable selection was used, which is the geometric mean of PSR and (1 − FDR). High accuracy of variable selection is indicated by a high GM.

MTL-EN variable selection accuracy was very similar to enetLTS with high PSR and FDR. As also shown in our previous study [9], Ensemble had the highest variable selection accuracy with much low FDR; however, Ensemble missed some associated variables when the proportion of outliers was 10% or 15%.

In terms of variable selection, when there were a small proportion of outliers, Ensemble performed best. However, its accuracy was greatly decreased when the proportion of outliers was large. In terms of outlier detection, regardless of the portion of outliers, MTL-EN had the highest outlier detection accuracy among the three methods.

4.2. Combining with Ensemble to Improve the Accuracy of Variable Selection. In our previous study [9], we considered a two-step procedure when the proportion of outliers was relatively large. We found that it improved the variable selection accuracy by applying Ensemble on a subset with outliers identified by enetLTS removed. In this study, we also used MTL-EN to detect outliers and then applied Ensemble on the subset with outliers removed. The results of MTL-EN and enetLTS were compared by simulation, which is shown in Table 1.

From Table 1, compared with the results in the original data, the PSR of Ensemble raised from 0.533 to 0.644, and the GM was improved from 0.714 to 0.786 for subset after removing outliers identified by enetLTS. For subset with outliers identified by MTL-EN removed, the results of Ensemble were also improved with PSR increased from 0.533 to 0.708 and GM increased from 0.714 to 0.828. It can be seen that after removing the outliers identified by MTL-EN, the accuracy of Ensemble variable selection is the highest.

4.3. The Computation Times of enetLTS and MTL-EN. From Table 2, the computation time of enetLTS is 39 times that of

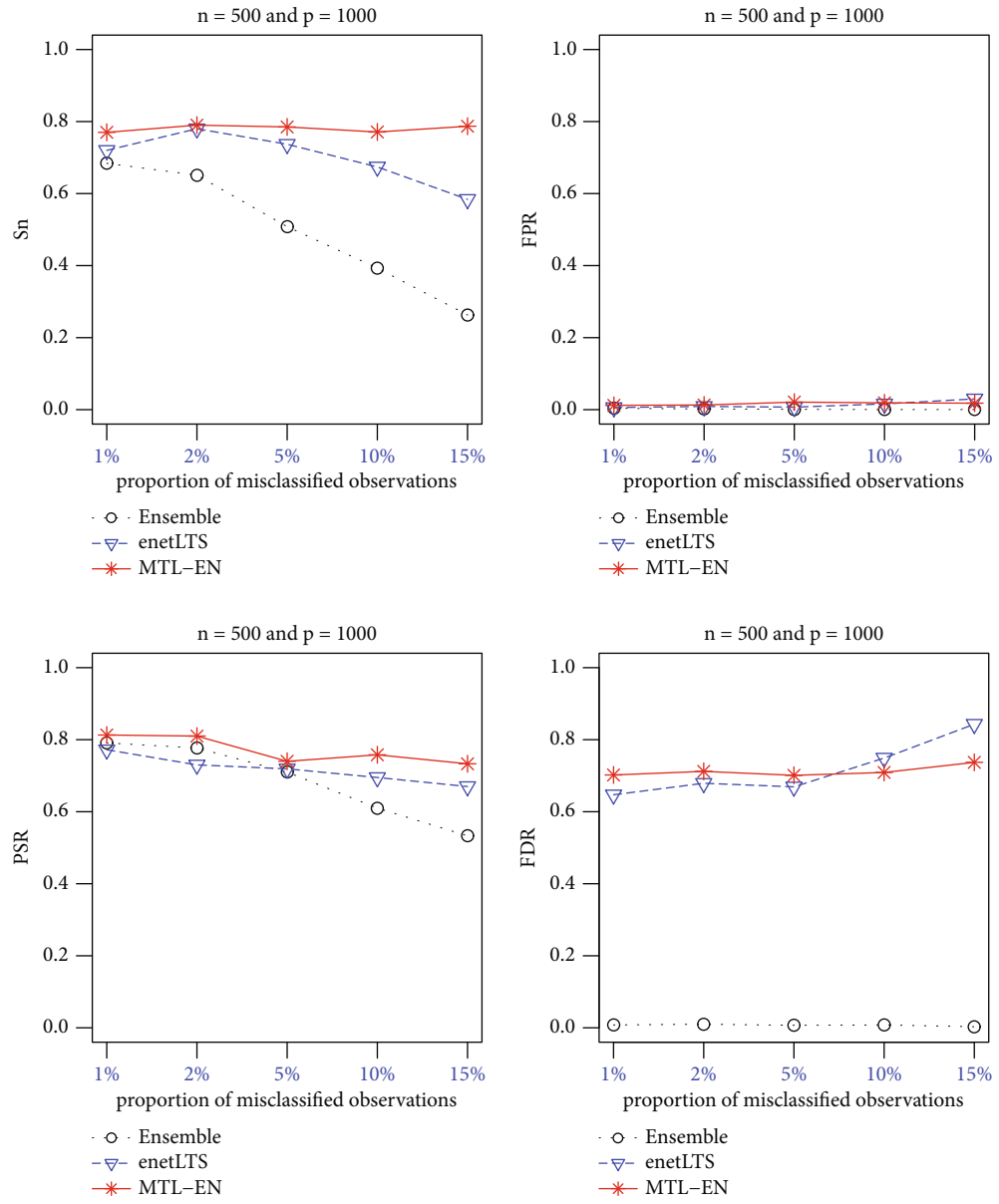

Figure 1: Results of MTL-EN, enetLTS, and Ensemble when $n = 500$ and $p = 1000$. Sn: sensitivity; FPR: False Positive Rate; PSR: Positive Selection Rate; FDR: False Discovery Rate.

Table 1: Results of Ensemble for the datasets with $n = 500$, $p = 1{,}000$, and $\varepsilon = 0.15$.

Data	Model size	PSR	FDR	GM#
Original data	16.06	0.533	0.003	0.714
Subset*	19.79	0.644	0.022	0.786
Subset**	21.75	0.708	0.021	0.828

*This subset is the original dataset after removing outliers identified by enetLTS. **This subset is the original dataset after removing outliers identified by MTL-EN. #GM: the geometric mean of PSR and (1-FDR).

Table 2: Computation times of MTL-EN and enetLTS for the datasets with $n = 500$, $p = 1{,}000$, and $\varepsilon = 0.1$.

Methods	Mean(s)
enetLTS	6489.06
MTL-EN	165.2

MTL-EN (Intel Core i7-6500U @2.50GHz); that is, the computation time of MTL-EN was 2% of that of enetLTS. This is because the C-step algorithm used by enetLTS does not take into account the regularized parameters that need to be determined at each step of the iteration. The criterion function cannot be guaranteed to gradually decrease, which makes the algorithm converge slowly. The AR-Cstep algorithm adopted by MTL-EN solves this problem well, which greatly improves the convergence speed.

5. Case Study

In the previous study [9], we compared the application of enetLTS, Ensemble, and Rlogreg on a TNBC dataset from the TCGA-BRCA data collection. The results showed that enetLTS identified 68 outliers, seven of which were individuals with inconsistent labels. After removing the outliers identified by enetLTS, the prediction accuracy of the three Ensemble models was improved, and the number of associated genes identified increased from 5 to 9. In this study, we applied MTL-EN to this TNBC dataset. The outliers identified by MTL-EN were compared with those by enetLTS, and we also compared the performances of Ensemble after removing the outliers identified by MTL-EN and enetLTS, respectively.

From Tables 3 and 4, among the 68 outliers identified by enetLTS, 3 of them were labeled as TNBC, which were also identified by MTL-EN; among them, 65 individuals with non-TNBC labels included 35 non-TNBC patients identified by MTL-EN. In other words, 38 of the 47 outliers with non-TNBC labels identified by MTL-EN were also identified by enetLTS. However, nine patients with TNBC labels were not identified by enetLTS. These 9 TNBC patients were highly expressed in one or more of the three genes, suggesting that they were likely to be non-TNBC patients or misclassified individuals. For example, TCGA-BH-A42U (HER2 38.37), TCGA-E2-A1L7 (ER 29.61, PR 22.98), TCGA-OL-A97C (PR 8.56), TCGA-A2-A1G6 (ER 23.90, PR 21.45, HER2 29.74), TCGA-A2-A0EQ (ER 2.13, HER2 30.15), TCGA-EW-A1OV (HER2 28.91), TCGA-OL-A5D6 (HER2 72.13), TCGA-C8-A26X (HER2 60.12), and TCGA-LL-A740 (HER2 68.56), with high expression in one or more of three receptors, were more likely not to be a TNBC patients; that is, his/her labels were probably wrong. Seven of the 47 outliers identified by MTL-EN were suspect individuals with inconsistent HER2 labels. Six of them were labeled as non-TNBC, which were also detected by enetLTS. The remaining one "TCGA-A2-A0EQ" was labeled as TNBC, which was not detected by enetLTS.

A total of 213 genes were identified by MTL-EN, and 40 genes with the largest absolute value are listed in Table 5. Among them, FOXA1 [17], ERBB2 [18], GRB7 [19], KRT16 [20], CXXC5 [21], FOXC1 [22], TFF3 [23], COL9A3 [24], FABP7 [25], CCNE1 [26], GZMB [27], and MIEN1 [28] were reported to be related to TNBC.

In our previous study [9], we combined the advantages of enetLTS and Ensemble and removed 68 outliers identified by enetLTS, then ran Ensemble on a subset (856 samples), to improve the accuracy of gene selection. In this study, we removed 47 misclassification samples detected by MTL-EN and then ran Ensemble in the remaining 877 samples. The results are shown in Tables 6 and 7.

From Table 6, for the subset with outliers detected by enetLTS removed, the prediction index MR of the three models in Ensemble was much lower than that on the original TNBC dataset; the MR of EN decreased from 0.012 to 0, the SPLS-DA MR reduced from 0.064 to 0.008, and the SGPLS MR reduced from 0.059 to 0.015. When Ensemble was run on a subset of 47 outliers identified by MTL-EN, the prediction accuracy MR of the three models in Ensemble also decreased greatly, to 0.001, 0.014, and 0.013, respectively.

Table 3: Number of misclassified observation that detected using enetLTS and Ensemble.

Method	Identified misclassification	Num of TNBC/non-TNBC*	Num of suspect TNBC/non-TNBC**
enetLTS	68	3/65	0/7
MTL-EN	47	12/35	1/6

*Number of identified misclassified observations with TNBC/non-TNBC labels. **Number of identified suspect individuals with inconsistent labels.

For subset with 68 outliers detected by enetLTS removed, the intersection of variables selected using the three Ensemble models increased from five to nine genes, namely, CA12 [29], GABRP [30], VGLL1 [31], AGR2 [32], GATA3 [17], FOXA1 [17], TFF3 [23], AGR3 [33], and KRT16 [20], were reported to be related to TNBC.

From Table 7, for subset with 47 outliers detected by MTL-EN removed, the intersection of variables selected using the three Ensemble models was 12 genes. Among them, ESR1, one of three key variables, and FOXC1 [22], AGR2 [32], FOXA1 [17], TFF3 [23], TFF1 [34], AGR3 [33], KRT6B [35], and KRT16 [20] have been reported to be related to TNBC. KLK6 [36], FDCSP [37], and PPP1R14C [38] have been reported to be related to other types of tumors. Their association with TNBC needs further study.

6. Discussion

Through our previous research [9], we have found that in high-dimensional data with mislabeled error, robust trimmed penalized regression is a recommended method in identifying mislabeled samples. However, the C-step algorithm to implement this method (enetLTS) is too slow to meet the requirement of data analysis for high-dimensional omics data. The reason is that for LTS without regularized parameters, the inequality that guarantees the convergence of the C-step algorithm is established. However, for the robust trimmed penalized regression with regularized parameters, the inequality does not necessarily hold due to the change of the regularized parameters.

In the AR-Cstep algorithm, penalized regression is repeatedly performed on the subset at each step to concentrate on the individuals who fit the model best gradually; that is, the idea of the C-step algorithm is still adopted. However, AR-Cstep can solve the problem of the C-step algorithm not converging because the regularized parameters change during the iteration. A comparison of the likelihood function of the current subset and that of the candidate subset is used to determine whether to replace the current subset with the candidate subset in AR-Cstep, thereby ensuring that the iterative process is in the direction that improves the criterion function. To avoid falling into a local optimum, the Metropolis-type probabilistic acceptance-rejection algorithm is combined.

Table 4: Forty-seven misclassified observations detected using MTL-EN for the TNBC dataset[#].

ID	ESR	PGR	HER2	HER2_level	HER2_status	HER2_FISH	y	Perres
TCGA-E9-A22G	0.44 (-)	0.02 (-)	15.32		+		Non-TNBC	32.54
TCGA-A2-A3Y0	2.18 (+)	0.03 (-)	11.34	1+	-		Non-TNBC	29.71
TCGA-A2-A04U	0.02 (-)	0.02 (-)	9.64	1+	-	+	Non-TNBC	22.86
TCGA-BH-A1EW	29.98 (-)	18.90 (-)	42.47		-		TNBC	18.89
TCGA-GM-A2DI	23.49 (-)	12.05 (-)	20.30			-	TNBC	14.73
TCGA-S3-AA0Z	16.67 (+)	0.07 (+)	33.07	1+	Equiv	-	Non-TNBC	14.58
TCGA-AN-A0FJ	0.08 (+)	0.04 (-)	14.28	1+	+		Non-TNBC	14.03
TCGA-BH-A5IZ	5.12 (+)	0.03 (-)	28.08		-	-	Non-TNBC	13.87
TCGA-OL-A5S0	0.09 (+)	0.06 (-)	31.92			+	Non-TNBC	13.45
TCGA-E9-A1ND	1.44 (-)	0.05 (-)	13.05		+		Non-TNBC	13.05
TCGA-B6-A0IJ	1.18 (+)	0.46 (+)	11.12				Non-TNBC	11.92
TCGA-AR-A251	1.57 (+)	0.10 (-)	14.02	2+	Equiv	-	Non-TNBC	10.70
TCGA-D8-A1JM	5.00 (+)	0.01 (-)	21.85	1+	-		Non-TNBC	10.52
TCGA-E2-A1II	0.14 (-)	0.19 (+)	10.73	1+	-		Non-TNBC	10.51
TCGA-A2-A1G6*	23.90 (-)	21.45 (-)	29.74	1+	-		TNBC	9.62
TCGA-A2-A0YJ	0.09 (+)	0.03 (-)	240.24	0	-		Non-TNBC	9.52
TCGA-LL-A5YP	0.16 (+)	0.05 (-)	15.10	1+	-	+	Non-TNBC	9.23
TCGA-E9-A1NC	0.11 (-)	0.07 (+)	15.91		+		Non-TNBC	8.98
TCGA-AC-A62X	0.19 (+)	0.02 (-)	28.53				Non-TNBC	8.93
TCGA-A7-A13E	0.82 (+)	0.06 (-)	46.08	2+	Equiv	-	Non-TNBC	8.77
TCGA-C8-A3M7	4.27 (-)	0.76 (-)	25.47		-		TNBC	8.71
TCGA-AR-A1AJ	1.47 (+)	0.07 (-)	9.74		-		Non-TNBC	8.53
TCGA-BH-A0DL	6.99 (+)	0.04 (-)	9.92		-		Non-TNBC	7.85
TCGA-E2-A1L7*	29.61 (-)	22.98 (-)	10.33		-		TNBC	7.35
TCGA-AR-A1AH	0.03 (+)	0.03 (-)	34.12		-		Non-TNBC	7.31
TCGA-E2-A14Y	0.67 (+)	0.03 (+)	487.90	2+	Equiv	+	Non-TNBC	7.11
TCGA-LL-A8F5	1.08 (+)	0.04 (-)	11.86	1+	-		Non-TNBC	6.96
TCGA-OL-A97C*	16.25 (-)	8.56 (-)	24.04			-	TNBC	6.86
TCGA-A7-A13D	0.52 (-)	0.81 (+)	42.28	2+	Equiv	-	Non-TNBC	6.73
TCGA-AR-A0TP	0.04 (+)	0.03 (-)	13.39		-		Non-TNBC	6.53
TCGA-LL-A6FR	0.33 (-)	0.04 (+)	32.13	2+	Equiv	+	Non-TNBC	6.19
TCGA-A2-A25F	0.62 (-)	0.23 (+)	5.19		-		Non-TNBC	5.86
TCGA-AO-A0JL	0.63 (-)	0.08 (-)	63.60	1+	-	+	Non-TNBC	5.45
TCGA-A2-A1G1	0.53 (-)	0.17 (-)	819.76	2+	Equiv	+	Non-TNBC	5.28
TCGA-BH-A42U*	9.19 (-)	1.83 (-)	38.37		-		TNBC	4.99
TCGA-AN-A0FX	1.13 (-)	0.64 (-)	24.02	1+	+		Non-TNBC	4.75
TCGA-D8-A1XW	0.32 (-)	0.11 (+)	21.03	1+	-		Non-TNBC	4.57
TCGA-AR-A24Q	1.00 (+)	0.36 (-)	20.67		-		Non-TNBC	4.52
TCGA-A1-A0SB	3.16 (+)	0.03 (-)	32.35		-		Non-TNBC	4.47
TCGA-A2-A4RX	0.68 (+)	0.93 (+)	26.64	1+	-		Non-TNBC	3.18
TCGA-AN-A0FL	0.09 (-)	1.07 (-)	15.07	1+	+		Non-TNBC	3.01
TCGA-A2-A0EQ*	2.13 (-)	0.04 (-)	30.15	3+	+	-	TNBC	2.63
TCGA-EW-A1OV*	0.23 (-)	0.03 (-)	28.91		-	-	TNBC	1.83
TCGA-OL-A5D6*	0.35 (-)	0.20 (-)	72.13			-	TNBC	1.69
TCGA-C8-A26X*	0.42 (-)	0.13 (-)	60.12	1+	-		TNBC	1.62
TCGA-LL-A740*	0.30 (-)	0.12 (-)	68.56	2+	Equiv	-	TNBC	1.48
TCGA-BH-A6R9	0.59 (-)	0.25 (+)	8.18		-		Non-TNBC	0.99

#Including the expression values, IHC, and FISH tests of ER, PR, and HER2 (individuals highlighted in bold are suspect individuals). *Outliers detected by MTL-EN but not by enetLTS. **Perres: the abstract value of Pearson residual.

Table 5: Top 40 genes selected by MTL-EN for the TNBC dataset.

Upregulated	COX7B2 (0.14), LBP (0.12), SLC15A1 (0.11), B3GNT5 (0.10), A2ML1 (0.10), FOXC1 (0.09), COL9A3 (0.09), KRT16 (0.09), FDCSP (0.09), FABP7 (0.09), AADAT (0.09), VSNL1 (0.09), KLK6 (0.09), PPP1R14C (0.08), GZMB (0.07), CCNE1 (0.07), FAM171A1 (0.07)
Downregulted	AGR3 (-0.24), CA12 (-0.20), AGR2 (-0.19), MLPH (-0.17), ESR1 (-0.15), TBC1D9 (-0.13), FOXA1 (-0.12), TFF1 (-0.12), ERBB2 (-0.11), GRB7 (-0.10), STARD3 (-0.10), PGAP3 (-0.10), TFF3 (-0.10), CXXC5 (-0.10), GATA3 (-0.10), ACOX2 (-0.09), ASPN (-0.09), MIEN1 (-0.08), SPDEF (-0.08), CHAD (-0.08), EEF1A2 (-0.08), CMBL (-0.08), SRARP (-0.07)

Table 6: Results of Ensemble three models for the original TNBC data and subset with outliers removed.

Dataset	EN		SPLS-DA		SGPLS	
	Model size**	MR#	Model size	MR	Model size	MR
Original data	175	0.012	22	0.064	33	0.059
Subset*	83	0.000	87	0.008	16	0.015
Subset##	49	0.001	38	0.014	55	0.013

*This subset is the original dataset after removing 68 outliers identified by enetLTS. ##This subset is the original dataset after removing 47 outliers identified by MTL-EN. **Model size: number of variables; #MR: misclassification rate.

Table 7: Genes selected by Ensemble for the TNBC subset*.

FOXC1, ESR1, AGR2, FOXA1, TFF3, TFF1, KLK6, AGR3, FDCSP, KRT6B, KRT16, PPP1R14C

*This subset is the original dataset after removing 47 outliers identified by MTL-EN.

Through simulation experiments, it is found that MTL-EN using AR-Cstep algorithm was more accurate than enetLTS using C-step algorithm in outlier identification. In particular, the accuracy of Ensemble variable selection on the subset after removing outliers identified by MTL-EN was higher than the result of Ensemble running on the subset after removing outliers identified by enetLTS. The AR-Cstep algorithm adopted by MTL-EN greatly improved the convergence speed; that is, the computation time of MTL-EN was 2% of that of enetLTS.

If a misclassified sample identified by a certain method is labeled as non-TNBC, it means that the expression of the key genes ER, PR, or HER2 is false positive in this patient. Similarly, if a misclassified sample identified is labeled as TNBC, it implies that the expression of ER, PR, or HER2 is a false negative in the patient. In the analysis of the TNBC dataset, there are 153 individuals labeled as TNBC in this TNBC dataset. There are 3 samples identified by enetLTS that were labeled as TNBC patients with false negative rate 2% (3/153). Twelve individuals labeled as TNBC patients were identified as mislabeled samples by MTL-EN with false negative rate 7.8% (12/153). In the TNBC dataset, IHC test of ER and PR was adopted for all patients. For HER2 detection, the results of IHC were for 507 patients. According to previous studies, the false negative rates of IHC test for ER, PR, and HER2 were not low, 15.1% ~21.8% for ER [39], 6.8% (4/58) for PR [40], and 6.2% (4/65) for HER2 [41], respectively. Therefore, the false negative misclassified samples identified by MTL-EN were more likely to be close to the reality than enetLTS.

A large class of computational problems in robust statistics can be formulated as the selection of the optimal subset of data based on some criterion function [10]. AR-Cstep algorithm, as the improvement of C-step algorithm, can be extended to other robust models with regularized parameters. It is an effective algorithm for finding the most suitable subset of regularized models, such as robust Adaptive LASSO, Group LASSO, SCAD, and MCP. The AR-Cstep algorithm can be extended to other generalized linear models, such as penalized multiclass logistic regression and penalized Poisson regression.

7. Conclusion

AR-Cstep can solve the problem of the C-step algorithm not converging because the regularized parameters change during the iteration. It provides an idea for developing the efficient algorithm of robust penalized regression based on trimming. The AR-Cstep algorithm can be extended to other robust models with regularized parameters. In practice, MTL-EN using AR-Cstep algorithm is the recommended method for mislabeled sample identification in omics data because of its high accuracy and high operation speed. When the proportion of mislabeled samples is relatively low and ≤5%, Ensemble can be used for variable selection. When the proportion of mislabeled samples is >5%, Ensemble can be used for variable selection on a subset of data after removing mislabeled samples identified by MTL-EN.

Data Availability

Code is available on Github (https://github.com/hwsun2000/AR-Cstep). The BRCA RNA-Seq FPKM dataset was imported using the "brca.data" R package (https://github.com/averissimo/brca.data/releases/download/1.0/brca.data_1.0.tar.gz).

Disclosure

The funders played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interests.

Acknowledgments

This research work was funded by the National Natural Science Foundation for Young Scholars of China (Grant No. 81502891) and the National Natural Science Foundation of China (Grant No. 81872715).

References

[1] M. B. Lopes, A. Verissimo, E. Carrasquinha, S. Casimiro, N. Beerenwinkel, and S. Vinga, "Ensemble outlier detection and gene selection in triple-negative breast cancer data," *BMC Bioinformatics*, vol. 19, no. 1, p. 168, 2018.

[2] C. Wu and S. Ma, "A selective review of robust variable selection with applications in bioinformatics," *Briefings in Bioinformatics*, vol. 16, no. 5, pp. 873–883, 2015.

[3] K. L. Ayers and H. J. Cordell, "SNP selection in genome-wide and candidate gene studies via penalized logistic regression," *Genetic Epidemiology*, vol. 34, no. 8, pp. 879–891, 2010.

[4] H. Sun and S. Wang, "Penalized logistic regression for high-dimensional DNA methylation data with case-control studies," *Bioinformatics*, vol. 28, no. 10, pp. 1368–1375, 2012.

[5] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[6] A. Alfons, C. Croux, and S. Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *The Annals of Applied Statistics*, vol. 7, no. 1, pp. 226–248, 2013.

[7] F. S. Kurnaz, I. Hoffmann, and P. Filzmoser, "Robust and sparse estimation methods for high-dimensional linear and logistic regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 172, pp. 211–222, 2018.

[8] P. J. Rousseeuw and K. Van Driessen, "Computing LTS regression for large data sets," *Data Mining and Knowledge Discovery*, vol. 12, no. 1, pp. 29–45, 2006.

[9] H. Sun, Y. Cui, H. Wang, H. Liu, and T. Wang, "Comparison of methods for the detection of outliers and associated biomarkers in mislabeled omics data," *BMC Bioinformatics*, vol. 21, no. 1, p. 357, 2020.

[10] B. Chakraborty and P. Chaudhuri, "On an optimization problem in robust statistics," *Journal of Computational and Graphical Statistics*, vol. 17, no. 3, pp. 683–702, 2008.

[11] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, pp. 427–443, 2004.

[12] A. Farcomeni and S. Viviani, "Robust estimation for the Cox regression model based on trimming," *Biometrical Journal*, vol. 53, no. 6, pp. 956–973, 2011.

[13] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, vol. 5, no. 1, p. 232, 2011.

[14] L. D. Maxim, R. Niebo, and M. J. Utell, "Screening tests: a review with examples," *Inhalation Toxicology*, vol. 26, no. 13, pp. 811–828, 2014.

[15] N. Ternes, F. Rotolo, and S. Michiels, "Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models," *Statistics in Medicine*, vol. 35, no. 15, pp. 2561–2573, 2016.

[16] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, 2011.

[17] X. Dai, R. Ma, X. Zhao, and F. Zhou, "Epigenetic profiles capturing breast cancer stemness for triple negative breast cancer control," *Epigenomics*, vol. 11, no. 16, pp. 1811–1825, 2019.

[18] R. Wang, Z. Huang, C. Qian et al., "LncRNA WEE2-AS1 promotes proliferation and inhibits apoptosis in triple negative breast cancer cells via regulating miR-32-5p/TOB1 axis," *Biochemical and Biophysical Research Communications*, vol. 526, no. 4, pp. 1005–1012, 2020.

[19] M. J. Gunzburg, K. Kulkarni, G. M. Watson et al., "Unexpected involvement of staple leads to redesign of selective bicyclic peptide inhibitor of Grb7," *Scientific Reports*, vol. 6, no. 1, article 27060, 2016.

[20] K. D. Yu, R. Zhu, M. Zhan et al., "Identification of prognosis-relevant subgroups in patients with chemoresistant triple-negative breast cancer," *Clinical Cancer Research*, vol. 19, no. 10, pp. 2723–2733, 2013.

[21] L. Fang, Y. Wang, Y. Gao, and X. Chen, "Overexpression of CXXC5 is a strong poor prognostic factor in ER+ breast cancer," *Oncology Letters*, vol. 16, no. 1, pp. 395–401, 2018.

[22] H. Pan, Z. Peng, J. Lin, X. Ren, G. Zhang, and Y. Cui, "Forkhead box C1 boosts triple-negative breast cancer metastasis through activating the transcription of chemokine receptor-4," *Cancer Science*, vol. 109, no. 12, pp. 3794–3804, 2018.

[23] G. G. Jinesh, E. R. Flores, and A. S. Brohl, "Chromosome 19 miRNA cluster and CEBPB expression specifically mark and potentially drive triple negative breast cancers," *PLoS One*, vol. 13, no. 10, article e0206008, 2018.

[24] X. Lv, M. He, Y. Zhao et al., "Identification of potential key genes and pathways predicting pathogenesis and prognosis for triple-negative breast cancer," *Cancer Cell International*, vol. 19, no. 1, p. 172, 2019.

[25] R. Z. Liu, K. Graham, D. D. Glubrecht, R. Lai, J. R. Mackey, and R. Godbout, "A fatty acid-binding protein 7/RXRβ pathway enhances survival and proliferation in triple-negative breast cancer," *The Journal of Pathology*, vol. 228, no. 3, pp. 310–321, 2012.

[26] R. Yang, L. Xing, X. Zheng, Y. Sun, X. Wang, and J. Chen, "The circRNA circAGFG1 acts as a sponge of miR-195-5p to promote triple-negative breast cancer progression through regulating CCNE1 expression," *Molecular Cancer*, vol. 18, no. 1, p. 4, 2019.

[27] J. Pérez-Pena, J. Tibor Fekete, R. Páez et al., "A transcriptomic immunologic signature predicts favorable outcome in neoadjuvant chemotherapy treated triple negative breast tumors," *Frontiers in Immunology*, vol. 10, p. 2802, 2019.

[28] X. Yu, W. Xiao, H. Song, Y. Jin, J. Xu, and X. Liu, "CircRNA_100876 sponges miR-136 to promote proliferation and

metastasis of gastric cancer by upregulating MIEN1 expression," *Gene*, vol. 748, article 144678, 2020.

[29] Y. Wang, H. Li, J. Ma et al., "Integrated bioinformatics data analysis reveals prognostic significance of SIDT1 in triple-negative breast cancer," *Oncotargets and Therapy*, vol. Volume 12, pp. 8401–8410, 2019.

[30] V. B. Wali, G. A. Patwardhan, V. Pelekanou et al., "Identification and validation of a novel biologics target in triple negative breast cancer," *Scientific Reports*, vol. 9, no. 1, p. 14934, 2019.

[31] M. Castilla, M. López-García, M. R. Atienza et al., "VGLL1 expression is associated with a triple-negative basal-like phenotype in breast cancer," *Endocrine-Related Cancer*, vol. 21, no. 4, pp. 587–599, 2014.

[32] P. Segaert, M. B. Lopes, S. Casimiro, S. Vinga, and P. J. Rousseeuw, "Robust identification of target genes and outliers in triple-negative breast cancer data," *Statistical Methods in Medical Research*, vol. 28, no. 10-11, pp. 3042–3056, 2019.

[33] A. Umesh, J. Park, J. Shima et al., "Identification of AGR3 as a potential biomarker though public genomic data analysis of triple-negative (TN) versus triple-positive (TP) breast cancer (BC)," *Clinical Oncology*, vol. 30, 27_suppl, p. 31, 2012.

[34] J. Yi, L. Ren, D. Li et al., "Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers," *Biomedicine & Pharmacotherapy*, vol. 124, article 109827, 2020.

[35] G. M. Sizemore, S. T. Sizemore, D. D. Seachrist, and R. A. Keri, "GABA (A) receptor pi (GABRP) stimulates basal-like breast cancer cell migration through activation of extracellular-regulated kinase 1/2 (ERK1/2)," *The Journal of Biological Chemistry*, vol. 289, no. 35, pp. 24102–24113, 2014.

[36] A. Sananes, I. Cohen, A. Shahar et al., "A potent, proteolysis-resistant inhibitor of kallikrein-related peptidase 6 (KLK6) for cancer therapy, developed by combinatorial engineering," *The Journal of Biological Chemistry*, vol. 293, no. 33, pp. 12663–12680, 2018.

[37] A. Shergalis, A. Bankhead, U. Luesakul, N. Muangsin, and N. Neamati, "Current challenges and opportunities in treating glioblastoma," *Pharmacological Reviews*, vol. 70, no. 3, pp. 412–445, 2018.

[38] J. Grey, D. Jones, L. Wilson et al., "Differential regulation of the androgen receptor by protein phosphatase regulatory subunits," *Oncotarget*, vol. 9, no. 3, pp. 3922–3935, 2018.

[39] Q. Li, A. C. Eklund, N. Juul et al., "Minimising immunohistochemical false negative ER classification using a complementary 23 gene expression signature of ER status," *PLoS One*, vol. 5, no. 12, article e15031, 2010.

[40] G. B. Fakhri, R. S. Akel, M. K. Khalil, D. A. Mukherji, F. I. Boulos, and A. H. Tfayli, "Concordance between immunohistochemistry and microarray gene expression profiling for estrogen receptor, progesterone receptor, and HER2 receptor statuses in breast cancer patients in Lebanon," *International Journal of Breast Cancer*, vol. 2018, Article ID 8530318, 6 pages, 2018.

[41] S. F. Wu, Y. Y. Liu, X. D. Liu, Y. Jiang, and X. Zeng, "HER2 gene status and mRNA expression in immunohistochemistry 1+ breast cancer," *Zhonghua bing li xue za zhi = Chinese Journal of Pathology*, vol. 47, no. 7, p. 522, 2018.