Computers in Biology and Medicine 141 (2022) 105154

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis

Juntao Li [a,*], Ke Liang [a], Xuekun Song [b]

[a] College of Mathematics and Information Science, Henan Normal University, Xinxiang, 453007, China
[b] College of Information Technology, Henan University of Chinese Medicine, Zhengzhou, 450046, China

ARTICLE INFO

Keywords:
Cancer diagnosis
Gene selection
Sparse group lasso
Acute leukemia

ABSTRACT

Cancer diagnosis based on gene expression profile data has attracted extensive attention in computational biology and medicine. It suffers from three challenges in practical applications: noise, gene grouping, and adaptive gene selection. This paper aims to solve the above problems by developing the logistic regression with adaptive sparse group lasso penalty (LR-ASGL). A noise information processing method for cancer gene expression profile data is first presented via robust principal component analysis. Genes are then divided into groups by performing weighted gene co-expression network analysis on the clean matrix. By approximating the relative value of the noise size, gene reliability criterion and robust evaluation criterion are proposed. Finally, LR-ASGL is presented for simultaneous cancer diagnosis and adaptive gene selection. The performance of the proposed method is compared with the other four methods in three simulation settings: Gaussian noise, uniformly distributed noise, and mixed noise. The acute leukemia data are adopted as an experimental example to demonstrate the advantages of LR-ASGL in prediction and gene selection.

1. Introduction

Cancer is a malignant tumor that seriously threatens human health and life. Timely detection of cancer in the early stages is crucial for patients [1]. Traditional techniques for cancer diagnosis are mainly based on medical imaging, e.g., ultrasound detection and computed tomography (CT). However, these diagnostic techniques cannot explain the pathogenesis from a microscopic perspective [2]. With the development of gene sequencing technology, many cancer gene expression profile data have emerged. By mining these high-throughput sequencing data, machine learning methods for classification have been successfully applied to distinguish cancer and normal samples [3–8]. These methods can identify biomarkers related to cancer, and analyzing these biomarkers can enable researchers to understand the pathogenic mechanism from a molecular perspective.

In the process of gene sequencing, cancer gene expression profile data inevitably contain noise information due to instrument electromagnetic interference and insufficient probe hybridization [9,10]. Noise may affect the classification performance of the model. To reduce noise, a few robust regression methods have emerged. Wang et al. provided LAD-lasso, which has good resistance to heavy-tailed errors or outliers in response [11]. Lambert-Lacroix and Zwald proposed a robust regression method that can solve the case of small errors [12]. Pannu and Billor presented a new robust functional variables (LAD-group lasso) selection method, which can reduce the influence of abnormal values on estimation [13]. Especially, noise in cancer gene expression profile data may be arbitrary or even non-specific distribution. However, the noise type is required to be known for these methods. Therefore, the above methods are not suitable for cancer gene expression profile data, and it is urgent to develop a new noise information processing technology. This is also the first motivation of this paper.

Since the relevant genes can be automatically selected when performing cancer diagnosis, machine learning methods for classification represented by sparse regression have attracted extensive attention in computational biology and medicine. Tibshirani initiated the lasso, which is a sparse linear regression model via shrinking some regression coefficients to zero [14]. Many lasso extension methods have emerged [15–18]. To select variables in groups, Yuan and Lin provided the group lasso (GL), which can identify a set of sparse sets by adopting the $L_{2,1}$ norm penalty [16]. Moreover, to generate group-wise and within-group sparsity, Simon et al. introduced the sparse group lasso (SGL), which can identify not only important groups but also important variables within

[*] Corresponding author.
E-mail addresses: juntaolimail@126.com (J. Li), keliangmail@163.com (K. Liang), bioinformatics_son@126.com (X. Song).

https://doi.org/10.1016/j.compbiomed.2021.105154
Received 22 October 2021; Received in revised form 14 December 2021; Accepted 15 December 2021
Available online 18 December 2021
0010-4825/© 2021 Elsevier Ltd. All rights reserved.

group [17]. There are some other group lasso penalty methods that have been developed [19–22].

Before applying the aforementioned methods to cancer diagnosis and related gene selection, it is necessary to divide genes into groups. Note that cancer is a complex disease and its pathogenesis is not clear [23]. Hence, it is almost impossible to divide genes into groups according to their biological mechanisms and gene pathways. As a system biology method, weighted gene co-expression network analysis (WGCNA) can identify network modules with high topological overlap [24,25]. The second motivation of this paper is to further explore the interpretable grouping method on noisy cancer gene expression profile data.

The third motivation of this paper is to select cancer-related genes adaptively. By using univariate estimators to construct weights, Zou proposed the adaptive lasso, which encourages adaptive variable selection [26]. To select variable adaptively, Zheng et al. developed a robust adaptive lasso [27]. Following a similar idea, Wang and Leng presented the adaptive group lasso [28], and Fang et al. provided the adaptive sparse group lasso [29]. By evaluating the individual gene significance and the influence to improve the correlation of all the other pairwise genes in each group, Li et al. proposed the adaptive sparse group lasso [30]. With the emergence of many biological data, some additional information (like genomic annotation or external p-values) on the variables is available. Using such information, Van De Wiel et al. proposed an adaptive group regularized ridge regression [31]. To harness the information contained in this “hidden genome” of variants, Chakraborty et al. presented the multilevel multinomial logistic regression [32]. Furthermore, Yi et al. proposed the penalization-based estimation approach [33].

This paper focuses on acute leukemia diagnosis based on gene expression profile data. To solve the problems of noise, gene grouping, and adaptive gene selection, we give a logistic regression with adaptive sparse group lasso penalty (LR-ASGL) and develop its solving algorithm. A noise information processing method for cancer gene expression profile data is proposed. A robust evaluation criterion that is the product of gene significance and gene reliability is provided, based on which LR-ASGL is constructed, and the corresponding algorithm is developed. Simulation studies in three noise settings and acute leukemia experiments are provided to verify the effectiveness of noise processing and adaptive weighting techniques.

The rest of this paper is organized as follows. Problem statement and preliminaries are shown in Section 2. The logistic regression with adaptive sparse group lasso is proposed, and the solving algorithm is presented in Section 3. Simulation study is shown in Section 4. Diagnosis of acute leukemia is given in Section 5. The conclusion of this paper is given in Section 6.

2. Problem statement and preliminaries

2.1. Problem statement

This paper focuses on cancer diagnosis based on gene expression profile data. Due to instrument electromagnetic interference and insufficient probe hybridization, noise is inevitable in such data. Given cancer gene expression profile data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ denotes the expression levels of p genes for the ith sample, y_i represents a class label corresponding to x_i. For cancer diagnosis, $y_i = 1$ if the ith sample comes from cancer patients, otherwise, $y_i = 0$. From the perspective of statistical learning, the problem of cancer diagnosis based on gene expression profile data is to predict the sample label for a new patient. Generalized linear regression is often used to construct the following decision function

$$D(x) = \begin{cases} 1, \textbf{if } f(x) \geq T_0, \\ 0, \textbf{if } f(x) < T_0, \end{cases} \tag{1}$$

where $f(x) = \beta^T x + \beta_0$ is the regression function and T_0 is the threshold parameter. In fact, the selection of threshold T_0 will significantly affect the prediction accuracy (cancer diagnosis). According to the significance of logarithmic odds, T_0 is taken as 0.5.

As mentioned in the introduction, lasso and its extension methods have been successfully applied to analyzing high-dimensional data [16,17,30]. However, there are three problems when these methods are used to cancer diagnosis based on gene expression profile data: (1) effectively process the noise information; (2) reasonably divide genes into groups in advance; (3) adaptively select the related genes in groups. This paper is devoted to solving the three problems mentioned above by developing the logistic regression with adaptive sparse group lasso penalty.

2.2. The related works

The least absolute shrinkage and selection operator (lasso) introduced by Tibshirani is a classical linear regression model which can be successfully used to perform cancer diagnosis on cancer gene expression profile data [14]. The regression coefficients of lasso can be solved via the following optimization problem:

$$\min_\beta \frac{1}{2} \parallel y - X\beta \|_2^2 + \lambda \parallel \beta \|_1, \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter, $X = (x_1, x_2, \ldots, x_n)^T \in R^{n \times p}$ represents matrix of features, y is a n-dimensional response vector. It should be noted that the constant β_0 in the regression funtion $f(x)$ is omitted in (2). In fact, the solution for β_0 is $\beta_0 = \bar{y} = \sum_{i=1}^n y_i / n$ and $\bar{y}$ is assumed to be zero without loss generality. We omitted the constant β_0 since it is a constant in our model and algorithm.

In order to select variables in groups, Yuan and Lin [16] proposed the following group lasso (GL) estimator:

$$\min_\beta \frac{1}{2} \parallel y - \sum\nolimits_{l=1}^m X^{(l)} \beta^{(l)} \|_2^2 + \lambda \sum\nolimits_{l=1}^m \sqrt{p_l} \parallel \beta^{(l)} \|_2, \tag{3}$$

where p_l represents the number of elements in lth group, $X^{(l)}$ and $\beta^{(l)}$ are the subsets of X and β corresponding to the lth group, respectively. Obviously, group lasso degenerates to the lasso if the size of each group is 1.

In order to generate sparsity within groups, Simon et al. [17] proposed the following sparse group lasso (SGL) estimator:

$$\min_\beta \frac{1}{2n} \parallel y - \sum\nolimits_{l=1}^m X^{(l)} \beta^{(l)} \|_2^2 + (1 - \alpha)\lambda \sum\nolimits_{l=1}^m \sqrt{p_l} \parallel \beta^{(l)} \|_2 + \alpha\lambda \parallel \beta \|_1, \tag{4}$$

where $0 \leq \alpha \leq 1$ and $\lambda \geq 0$ are regularization parameters. Particularly, sparse group lasso model degenerates to group lasso model if $\alpha = 0$ and lasso if $\alpha = 1$.

In order to adaptively select genes in groups, Li et al. [30] proposed the following adaptive sparse group lasso (ASGL-CMI) estimator:

$$\min_\beta \frac{1}{2n} \parallel y - \sum\nolimits_{l=1}^m X^{(l)} \beta^{(l)} \|_2^2 + (1 - \alpha)\lambda \sum\nolimits_{l=1}^m \sqrt{p_l} \parallel \beta^{(l)} \|_2 + \alpha\lambda \parallel W\beta \|_1, \tag{5}$$

where $W = diag\{w_1^1, w_2^1, \ldots, w_{p_l}^l\}$ is the weight matrix based on information theory. If W is the identity matrix, the model degenerates to the sparse group lasso.

3. Method

3.1. Noise information processing

Noise is inevitable in cancer gene expression profile data due to instrument electromagnetic interference and insufficient probe hybridization in the experimental process [9,10]. From the perspective of information processing, noise is considered harmful or useless. When noise is involved, it is a general and popular idea to remove or reduce it. A few robust regression models reducing the influence of noise have emerged when the noise type is known [11–13]. Note that noise in cancer gene expression profile data may be arbitrary or even non-specific distribution. Hence, it is necessary to develop new noise processing technology. Primarily, we argue that noise information helps build statistical models for cancer diagnosis based on gene expression profile data.

Let $X_{train} = (x_1, x_2, \ldots, x_{n_1})^T$ be the feature matrix of training set for cancer gene expression profile data, which is constructed by the expression levels of p genes of n_1 samples. Motivated by the fact that gene expression profile data has ultra-high dimensional property and only a small part of the values of expression are affected by noise, we decompose feature matrix X_{train} into a low-rank clean matrix $\widehat{D}$ and a sparse noise matrix $\widehat{E}$ via the robust principal component analysis (RPCA):

$$\begin{aligned}\langle \widehat{D}, \widehat{E} \rangle = \mathrm{argmin}_{D,E} \parallel D \|_* + \lambda \parallel E \|_1, \\ s.t. \quad D + E = X_{train},\end{aligned} \tag{6}$$

where λ is a hyperparameter, $\parallel E \parallel_1$ denotes the L_1 norm of E, $\parallel D \parallel_* = \sum_i \sigma_i(D)$ denotes the nuclear norm, i.e., the sum of its singular values. Among all possible decompositions, (6) is a convenient convex program called Principal Component Pursuit which minimizes a weighted combination of the nuclear norm and of the L_1 norm. We adopt the default value of parameter λ, which is suggested by Candes et al. [34]. The optimization problem (6) can be solved via R package *rpca*.

Remark 1. According to the traditional viewpoint of removing or reducing noise, the sparse noise matrix $\widehat{E}$ should stay away from the model after completing the RPCA, i.e., the statistical model should be built by only using the low-rank clean matrix $\widehat{D}$ in which the main structural information of X_{train} is contained. In this paper, we argue that noise information contained in sparse noise matrix $\widehat{E}$ is useful. We use not only clean matrix $\widehat{D}$ but also noise matrix $\widehat{E}$ to build statistical models. In fact, noise information is used to assess the reliability of the gene, which is incorporated into the penalty term of the statistical model (see subsection 3.3 and 3.4).

Remark 2. As shown in the literature [35], it is valid to decompose feature matrix X_{train} by using RPCA under the assumption that noise is sparse. This noise sparsity hypothesis is reasonable because of the ultra-high dimensional characteristics of cancer gene expression profile data. It should be pointed out that the obtained sparse noise matrix $\widehat{E}$ is not equal to the real noise matrix even though noise information is contained in it. We approximate the relative value of the noise size by using the metric ratio of the corresponding column vector of the noise matrix and the clean matrix (see subsection 3.3).

3.2. WGCNA on clean data

It is an urgent problem to divide genes into groups for cancer diagnosis based on gene expression profile data. Generally speaking, cancer is caused by the interaction of multiple genes rather than a single gene. If genes are not divided into groups (or divided into groups randomly), then the cooperation effect of genes will be ignored (or no reasonable explanation). It is natural to divide genes into groups by using the real gene pathway. However, there are very few known gene pathways. Considering the advantages of weighted gene co-expression network analysis (WGCNA) [24,25], we divide genes into groups by performing WGCNA on the low-rank clean matrix $\widehat{D}$. The brief steps of the algorithm are as follows:

Step 1 Divide feature matrix X_{train} into two parts: clean matrix $\widehat{D}$ and noise matrix $\widehat{E}$;

Step 2 Choose appropriate soft threshold power τ and construct the weighted gene co-expression network by using low-rank clean matrix $\widehat{D}$;

Step 3 Identify the modules of the weighted gene co-expression network;

Step 4 Divide genes into m different groups according to the identified modules;

Step 5 Divide $\widehat{D}$ into m submatrices $\widehat{D}^{(1)}$, $\widehat{D}^{(2)}$, ⋯, $\widehat{D}^{(m)}$ according to m gene groups.

According to the obtained m gene groups, the sparse noise matrix $\widehat{E}$ obtained by RPCA can be easily divided into m submatrices $\widehat{E}^{(1)}$, $\widehat{E}^{(2)}$, ⋯, $\widehat{E}^{(m)}$. The proposed grouping method is different from that of ASGL-CMI proposed by Li et al. [30], WGCNA is performed on the feature matrix X_{train} with noise for the latter. Here, we perform WGCNA on low-rank clean matrix $\widehat{D}$, which can reduce the influence of noise on gene modules division. It should be noted that an appropriate soft threshold power τ should be selected in Step 2. The screening principle of τ is to make the network with the scale-free network characteristics. Without losing too much network connectivity, it is best to choose a τ that is greater than some threshold (e.g., scale free topology model fit index $R^2 > 0.8$) [24,25].

3.3. Gene significance and reliability

After dividing genes into groups, we focus on evaluating the gene significance within each group and the gene reliability for each gene in the following subsection.

Firstly, we evaluate the tth gene significance in the lth group on low-rank clean matrix $\widehat{D}$. Let $\widehat{D}_i^{(l)}, \widehat{D}_j^{(l)}, \widehat{D}_t^{(l)}$ denote the gene expression level of the ith, jth and tth gene in the lth group for low-rank clean matrix $\widehat{D}$, respectively. Let $I(\widehat{D}_i^{(l)}; \widehat{D}_j^{(l)})$ be the mutual information which measures the amount of information shared by $\widehat{D}_i^{(l)}$ and $\widehat{D}_j^{(l)}$, $I(\widehat{D}_i^{(l)}; \widehat{D}_j^{(l)} | \widehat{D}_t^{(l)})$ be the conditional mutual information of variables $\widehat{D}_i^{(l)}$ and $\widehat{D}_j^{(l)}$. Let $[*]_+ = \max\{0, *\}$. We adopt the following integrated evaluation criterion:

$$\widehat{s}_t^{(l)} = \frac{\delta_2}{\delta_1 + \delta_2} si_t^{(l)} + \frac{\delta_1}{\delta_1 + \delta_2} sc_t^{(l)}, \tag{7}$$

where δ_1 and δ_2 represent the upper bound of the $si_t^{(l)}$ and $sc_t^{(l)}$, $si_t^{(l)} = \frac{1}{(p_l - 1)^2} \sum_{\substack{i=1 \\ i \neq t}}^{p_l} \sum_{\substack{j=1 \\ j \neq t}}^{p_l} I(\widehat{D}_i^{(l)}; \widehat{D}_j^{(l)} | \widehat{D}_t^{(l)})$, $sc_t^{(l)} = \frac{1}{(p_l - 1)^2} \sum_{\substack{i=1 \\ i \neq t}}^{p_l} \sum_{\substack{j=1 \\ j \neq t}}^{p_l} \left[I\left(\widehat{D}_i^{(l)}; \widehat{D}_j^{(l)} | \widehat{D}_t^{(l)}\right) - I\left(\widehat{D}_i^{(l)}; \widehat{D}_j^{(l)}\right) \right]_+$. As shown in the literature [30]. $si_t^{(l)}$ evaluate the individual significance of the tth gene in the lth group and $sc_t^{(l)}$ evaluate the improved correlation of all the other $(p_l - 1)^2$ pairwise genes in the lth group. The gene significance evaluation criterion (7) can be calculated by using the MATLAB toolbox *MIToolbox*, which is available online at https://mloss.org/software/view/325/.

Secondly, we evaluate the gene reliability by using the sparse noise matrix $\widehat{E}$. Due to noise, not all genes are equally reliable for cancer gene expression profile data. In fact, noise information affects the reliability of genes, and important but unreliable genes should not contribute equally to the establishment of models with important and reliable genes. Therefore, it is necessary to consider the gene significance and the gene reliability. Following the above ideas, we attempt to assess the gene reliability by using the noise information contained in sparse noise matrix $\widehat{E}$. Note that the sparse noise matrix $\widehat{E}$ is not equal to the real noise matrix even though noise information is contained in it. Let $\widehat{D}_t^{(l)}$ and $\widehat{E}_t^{(l)}$ denote the column vector of matrix $\widehat{D}$ and $\widehat{E}$ corresponding to the tth gene in the lth group, respectively. We approximate the relative value of the noise size by using the metric ratio of the corresponding column vector of the noise matrix and the clean matrix, and propose the following gene reliability criterion:

$$\widehat{r}_t^{(l)} = \frac{\| \widehat{D}_t^{(l)} \|_1}{\| \widehat{E}_t^{(l)} \|_1 + \varepsilon}, \tag{8}$$

where ε is a sufficiently small real number. Note that $\| \widehat{E}_t^{(l)} \|_1$ may be zero due to the sparsity of $\widehat{E}$. Hence, we introduce ε in (8). Obviously, the value of gene reliability $\widehat{r}_t^{(l)}$ is large when the noise level of tth gene in lth group is low or tends to zero.

3.4. Statistical model

According to the biomedical view, cell canceration is usually not caused by all genes but by a small number of gene mutations. Some successful adaptive methods have been proposed to highlight the gene significance [26,28,30]. Since gene significance is introduced as weights of penalty term, these methods encourage adaptability in gene selection. As shown in subsection 3.3, not all genes are equally reliable and important but unreliable genes should not contribute equally to the establishment of models with important and reliable genes when noise is involved. Hence, we argue that gene reliability evaluated by noise information should be introduced into the model besides the gene significance. Combining gene significance evaluation criterion (7) and gene reliability criterion (8), we propose the following robust evaluation criterion for the tth gene in lth group:

$$C_t^{(l)} = \widehat{r}_t^{(l)} * \widehat{s}_t^{(l)}. \tag{9}$$

Compared with the gene significance evaluation criterion in the literature [30], the robust evaluation criterion (9) is the product of gene significance and reliability. The latter emphasizes not only the gene significance but also the gene reliability. Only important and reliable genes are considered to have an important impact on classification (cancer diagnosis). The gene significance evaluation criterion (7) is a special case of robust evaluation criterion (9). Especially, the robust evaluation criterion (9) will degenerate into the (7) when there is no noise in the data. According to (9), we construct the weight vector of lth group:

$$w^{(l)} = \left(\frac{1}{C_1^{(l)}}, \frac{1}{C_2^{(l)}}, \ldots, \frac{1}{C_{p_l}^{(l)}}\right). \tag{10}$$

and then construct a weight matrix:

$$\begin{aligned} W \qquad &= diag\{w^{(1)}, \ldots, w^{(m)}\} \\ &= diag\{w_1^{(1)}, \ldots, w_{p_1}^{(1)}, \ldots, w_1^{(m)} \cdots w_{p_m}^{(m)}\}. \end{aligned} \tag{11}$$

where p_l represents the number of genes in lth group, $l = 1, 2, \ldots, m$.

Introducing the adaptive weight (10) into sparse group lasso penalty, we propose the following logistic regression with adaptive sparse group lasso penalty (LR-ASGL):

$$\min_\beta L(\beta) + (1-\alpha)\lambda \sum\nolimits_{l=1}^{m} \sqrt{p_l} \parallel \beta^{(l)} \parallel_2 + \alpha\lambda \sum\nolimits_{l=1}^{m} \parallel W^{(l)}\beta^{(l)} \parallel_1, \tag{12}$$

where $L(\beta) = \frac{1}{n}[(\sum_{i=1}^{n}\sum_{l=1}^{m}\log(1 + e^{x_i^{(l)T}\beta^{(l)}}) - y_i x_i^{(l)T}\beta^{(l)})]$ is the negative log-likelihood loss function, $0 \le \alpha \le 1$ and $\lambda \ge 0$ are regularization parameters, coefficient vector $\beta^{(l)}$ corresponds to the lth subvector of β, p_l represents the number of genes in lth group, $x_i^{(l)}$ is subvector of x_i corresponding to the lth group, $W^{(l)}$ represents the lth block-diagonal submatrix of W, respectively. There are two regularization parameters α and λ in statistical model (12). It is difficult to find the optimal parameter pair in two-dimensional space. Hence, we fix the parameter α in advance as 0.1, 0.2, 0.3, …, 0.8, 0.9. For each value of α, the optimal λ is selected via cross-validation. In particular, when the data does not contain noise, we default $\widehat{r}_t^{(l)} = 1$. Then LR-ASGL degenerates to ASGL-CMI.

Remark 3. The robust evaluation criterion (9) is the product of gene importance and reliability rather than the sum of them. Generally, the addition rule can also take into account the significance and reliability of genes. However, there are significant differences in the orders of magnitude between the significance based on conditional mutual information and the reliability of genes constructed by L_1 norm. Generally, the order of magnitude for the latter is larger than that for the former. If the principle of addition is adopted, then gene reliability is overemphasized and the gene significance is hardly effective (due to significant differences of the orders of magnitude). The multiplication rule used in the robust evaluation criterion effectively solves the problem of significant differences of the orders of magnitude.

3.5. Algorithm

Following the idea in literature [17], we character the optimal solution $\widehat{\beta}$ of (13) by subgradient equations. The coefficient $\beta^{(l)}$ of lth group satisfies the following equation:

$$\frac{1}{n}x_i^{(l)} \sum\nolimits_{i=1}^{n} \sum\nolimits_{l=1}^{m} \left(y_i - \frac{e^{x_i^{(l)T}}\beta^{(l)}}{1 + e^{x_i^{(l)T}}\beta^{(l)}}\right) = \sqrt{p_l}(1-\alpha)\lambda u + \alpha\lambda W^{(l)} v, \tag{13}$$

where u and v are subgradient of $\|\beta^{(l)}\|_2$ and $\|\beta^{(l)}\|_1$, respectively. Let $R_{(-l)} = \sum_{i=1}^{n}\sum_{k \neq l}(\log(1 + e^{x_i^{(k)T}\beta^{(k)}}) - y_i x_i^{(k)T}\beta^{(k)})$ denote the partial residual of y. The optimal problem (13) can be transformed as:

$$\begin{aligned} &\min_{\beta^{(l)}} \frac{1}{n}\left(R_{(-l)} - \sum\nolimits_{i=1}^{n}(\log(1 + e^{x_i^{(l)T}}\beta^{(l)}) - y_i x_i^{(l)T})\right) + \sqrt{p_l}(1-\alpha)\lambda\|\beta^{(l)}\|_2 \\ &\quad + \alpha\lambda\|W^{(l)}\beta^{(l)}\|_1. \end{aligned} \tag{14}$$

Similar to the literature [17], the optimal solution $\widehat{\beta}^{(l)}$ of (15) can be approximately solved by the following iterative equation

$$\begin{aligned} \widetilde{\beta}^{(l)} = &\left(1 - \frac{t\sqrt{p_l}(1-\alpha)\lambda}{\left\|S(\beta_0^{(l)} - t\nabla l(R_{(-l)}, \beta_0^{(l)}), t\alpha\lambda W^{(l)} e_l)\right\|_2}\right)_+ \\ &* S\left(\beta_0^{(l)} - t\nabla l(R_{(-l)}, \beta_0^{(l)}), t\alpha\lambda W^{(l)} e_l\right), \end{aligned} \tag{15}$$

where $t > 0$ is the stepsize of gradient step, e_l is a p_l-dimensional vector with each element 1, $S(\bullet)$ is the coordinate-wise soft thresholding operator [17]. Fixing the values of other coefficient vectors, $\widetilde{\beta}^{(l)}$ in (15) will converge to the optimal solution $\widehat{\beta}^{(l)}$. Cyclically iterating through the groups, overall optimal solution $\widehat{\beta}$ of (13) can be obtained.

Based on the above derivation, we develop LR-ASGL algorithm for solving the proposed optimal problem (12). The brief steps of the algorithm for solving LR-ASGL and predicting the new sample label are as follows:

step 1 Input X_{train}, X_{test}, y_{train}, y_{test};

step 2 Decompose X_{train} into clean matrix $\widehat{D}$ and noise matrix $\widehat{E}$ via R package *rpca*;

step 3 Divide genes into groups on matrix $\widehat{D}$ by R package *WGCNA* (see, subsection 3.2) and build the group index vector V;

step 4 Calculate gene significance $\widehat{s}_t^{(l)}$, gene reliability $\widehat{r}_t^{(l)}$ and robust evaluation criterion $C_t^{(l)}$ according to (7), (8) and (9), respectively.

step 5 Fit LR-ASGL on training set (X_{train}, y_{train}):

(1) Input X_{train}, y_{train}, group index vector V, robust evaluation value $C_t^{(l)}$;

(2) Determine the optimal parameter pair (α_o, λ_o) by Algorithm 1;

(3) Fit LR-ASGL via R function *asgl* and obtain the coefficient vector $\widehat{\beta}$;

(4) Extract the nonzero coefficient vector $\widehat{\beta}_{nz}$.

step 6 Determine decision function $D(x)$ in (1) by using coefficient vector $\widehat{\beta}$;

step 7 Predict labels of the new samples on X_{test};

step 8 Determine the related genes and their corresponding group according to $\widehat{\beta}_{nz}$ and V;

It should note that the optimal parameter pair (α_o, λ_o) can be selected through independent tests on the validation set if the number of samples is sufficient. Note that cancer gene expression profile data is usually only dozens of samples. Hence, dividing the data into three parts is not appropriate: training set, validation set, and test set. Note also that the cost of any fold of cross-validation is acceptable. Hence, we assign a list of α in advance and determine the optimal parameter λ via the Leave-one-out method.

Algorithm 1. Pseudocode for the optimal parameter pair (α_o, λ_o)

Algorithm 1 Pseudocode for the optimal parameter pair (α_o, λ_o)

```
Input: $X$: feature matrix for training data,
       $y$: response vector for training data
       $n_1$: the number of samples
       $n_2$, $n_3$: positive integer
       $C_t^{(l)}$: robust evaluation criterion,
       $\alpha = (0.1 : 1/n_3 : 0.9)$
       $\lambda = (\lambda_{min} : (\lambda_{min} - \lambda_{max})/n_2 : \lambda_{max})$
Output: Optimal parameter pair $(\alpha_o, \lambda_o)$
for i=1:$n_3$ do
    for j=1:$n_1$ do
        for k=1:$n_2$ do
            Fit LR-ASGL on $X[-j,]$ via R function $asgl$ for parameter pair ($\alpha_i$, $\lambda_k$).
            Predict the label of the $j$th sample: $\hat{y}_j(\alpha_i, \lambda_k)$
        end for
    end for
    $\lambda_o(\alpha_i)$=arg $\min\limits_{k} \sum_{j=1}^{n_1} \frac{1}{n_1}(\hat{y}_j(\alpha_i, \lambda_k) - y_j)^2$
end for
($\alpha_o$, $\lambda_o$)=arg $\min\limits_{i} \lambda_o(\alpha_i)$
Return($\alpha_o, \lambda_o$)
```

4. Simulation study

We tested the performance of the proposed LR-ASGL with least absolute shrinkage and selection operator (lasso) [14], elastic net (EN) [15], group lasso (GL) [16] and sparse group lasso (SGL) [17], and in three settings: Gaussian noise, uniformly distributed noise and mixed noise. Two datasets of different sizes were used for each setting.

We randomly generated noiseless covariate matrix $\widehat{X}^1 \in R^{50\times100}$ according to the normal distribution. Let x_i^T be the ith row of $\widehat{X}^1$. Label of instance x_i was defined as

$$y_i = \begin{cases} 1, & \textbf{if } x_i^T c_1 \geq \dfrac{e^T \widehat{X}^1 c_1}{n}, \\ 0, & \textbf{otherwise}, \end{cases} \tag{16}$$

where $e = (1,1,..,1)^T$ is the one-value vector, $c_1 = (-2,-1.79,-1.58,\ldots,1.79,2,0,0,0,\ldots,0)^T$ is coefficient vector in which the first 20 elements are nonzero. The first noiseless data was constructed as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_{50}, y_{50})\}$. Similarly, we generated a noiseless covariate matrix $\widehat{X}^2 \in R^{100\times200}$, defined the label for each instance, and constructed the second noiseless dataset with 100 samples and 200 features. For the second case, the coefficient vector was selected as $c_2 = (-3, -2.88, -2.76, \ldots, 2.88, 3, 0, 0, 0, \ldots, 0)$ (the first 50 elements are nonzero). Gaussian noise, uniformly distributed noise and mixed noise were added to the covariate matrix $\widehat{X}^1$ and $\widehat{X}^2$ respectively to form two datasets for each setting. Note that the above datasets were randomly generated. Hence, there was no group effects among features. For this case, we considered each feature as a individual group and set $\widehat{s}_t^{(l)} = 1$. In particular, when $\widehat{s}_t^{(l)} = 1$, the model ASGL-CMI is equivalent to SGL.

LR-ASGL is implemented by the algorithm in subsection 3.5, SGL and GL are implemented by *SGL* package, EN and lasso are implemented by *glmnet* package. For LR-ASGL, SGL and EN, we specified α = {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9} in advance and selected the optimal parameter pair (α_o, λ_o) by using cross-validation. GL is implemented by specifying α as 0 in the *SGL* package, and the lasso is implemented by specifying α as 1 in the *glmnet* package. Each model uses the logistic loss function in the experiment of this article. It should also be noted that only LR-ASGL is fitted and tested on clean data, and other methods are fitted and tested on noisy data. The prediction accuracy (ACC) and the number of the selected features (NUM) were adopted to evaluate the performance of these methods.

4.1. *Gaussian noise setting*

In this setting, we added Gaussian noise to the covariate matrices $\widehat{X}^1$ and $\widehat{X}^2$. Gaussian noise matrices X_{50}^N and X_{100}^N were randomly generated according to standard normal distribution. Let $X^1 = \widehat{X}^1 + 0.1{\cdot}X_{50}^N$ and $X^2 = \widehat{X}^2 + 0.1{\cdot}X_{100}^N$ be the covariate matrices with additive Gaussian noise. Two-thirds of the samples were randomly selected as training samples. In order to ensure the balance, the samples labeled 0 and 1 were respectively selected according to the above proportion. The remaining one-third of the samples were used as test samples. We fitted LR-ASGL, SGL, GL, EN, and lasso on training samples and predicted the labels of test samples. Simulation results show that LR-ASGL achieves

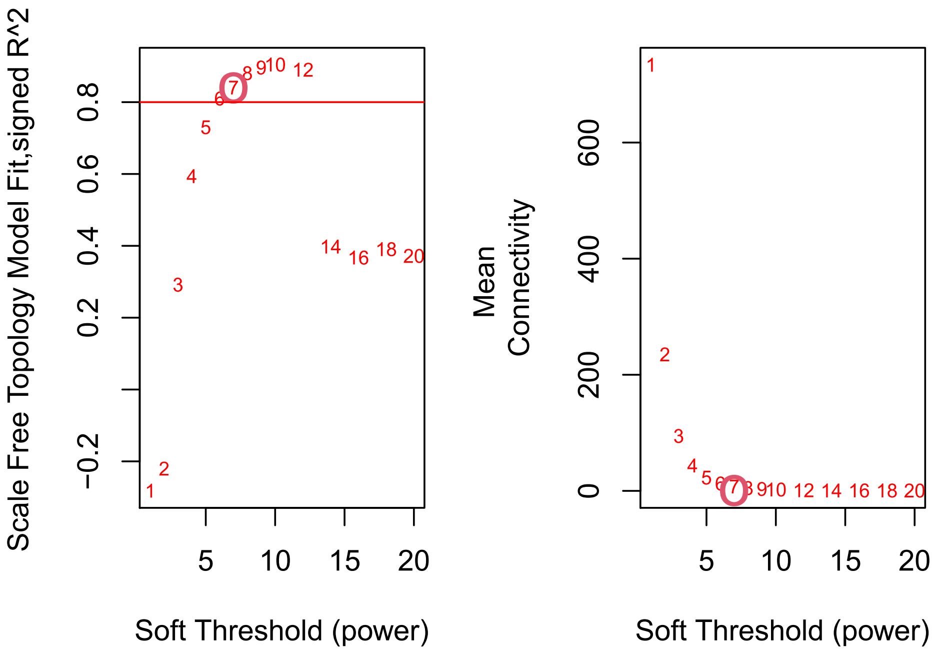

Fig. 1. Scatter plots of R^2-τ and connectivity-τ.

the highest prediction accuracy among the five methods for both cases. In the case 1 (n = 50, p = 100), the prediction accuracy of LR-ASGL is 87.87%, which is 6.06%, 6.06%, 19.12% and 19.12% higher than SGL, GL, EN and lasso, respectively. The number of features selected by LR-ASGL is 7, which is the least among the five methods. In the case 2 (n = 100, p = 200), the prediction accuracy of LR-ASGL is 75.75%, which is 6.06%, 3.03%, 6.06% and 12.12% higher than SGL, GL, EN and lasso, respectively. When the feature dimension becomes larger (p = 200), lasso selects the least number of features, but its prediction accuracy is much lower than LR-ASGL. In addition, there is no significant difference in performance between SGL and GL. The possible reason is that each feature is regarded as an individual group in the simulation experiment.

4.2. Uniformly distributed noise setting

In this setting, we added uniformly distributed noise to the covariate matrices $\widehat{X}^1$ and $\widehat{X}^2$. We randomly generated a 50 × 100 matrix X^{U_1} and a 100 × 200 matrix X^{U_2} according to uniform distribution U_1(0.1, 0.3) and U_2(0.05, 0.25), respectively. Let $X^3 = \widehat{X}^1 + X^{U_1}$ and $X^4 = \widehat{X}^2 + X^{U_2}$ be covariate matrices with uniformly distributed noise. According to the method in 4.1, we divided the data into training samples and test samples. Simulation results show that LR-ASGL achieves the highest prediction accuracy among the five methods for both cases. In the case 1 (n = 50, p = 100), the prediction accuracy of LR-ASGL is 90.90%, which is 3.03%, 6.06%, 15.90% and 22.15% higher than SGL, GL, EN and lasso, respectively. The number of features selected by LR-ASGL is 12, the least among the five methods. In the case 2 (n = 100, p = 200), the prediction accuracy of LR-ASGL is 78.78%, which is 6.06%, 6.06%, 3.03% and 15.15% higher than SGL, GL, EN and lasso, respectively. When the feature dimension becomes larger (p = 200), lasso and LR-ASGL respectively select the least and the second least features, which are far less than the other four methods.

4.3. Mixed noise setting

Note that the noise type is often unknown in practice. Hence, we simulated noise by using additive mixture of Gaussian noise and uniformly distributed noise. For the case 1, we let $X^{M1} = 0.1 \cdot X^N_{50} + X^{U_3}$ denote the mixed noise matrix, where X^{U_3} obeys the uniform distribution U_3(0.1, 0.3). For the case 2, we let $X^{M2} = 0.1 \cdot X^N_{100} + X^{U_4}$ denote mixed noise matrix, where X^{U_4} obeys uniform distribution U_4(0.05, 0.1). Let $X^5 = \widehat{X}^1 + X^{M1}$ and $X^6 = \widehat{X}^1 + X^{M2}$ be covariate matrices. In the case 1 (n = 50, p = 100), the prediction accuracy of LR-ASGL is 90.90%, which is 3.03%, 6.06%, 15.90% and 22.15% higher than SGL, GL, EN and lasso, respectively. LR-ASGL selects the least number of features among the five methods. In the case 2 (n = 100, p = 200), both LR-ASGL and EN have achieved 75.75% prediction accuracy, which is 3.03%, 3.03%, and 12.12% higher than SGL, GL and lasso, respectively. When the feature dimension becomes larger (p = 200), lasso and LR-ASGL select the least and the second least features.

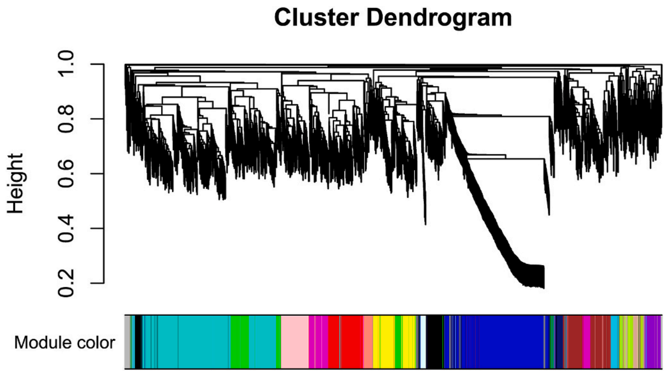

Fig. 2. Identified network modules.

5. Application in diagnosis of acute leukemia

We applied LR-ASGL to diagnose acute leukemia in this section. Acute leukemia gene expression profile data includes 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples [36]. For convenience, we labeled ALL samples as 1 and AML samples as 0. According to the preprocessing method in the literature [37], 3571 most significant genes of 72 samples were selected from 7129 genes. We compared the performance of LR-ASGL with the other five methods: (1) ASGL-CMI, (2) SGL, (3) GL, (4) EN, and (5) lasso,

Table 1
The prediction accuracy and the number of gene selection in one random partition of acute leukemia data.

Method	ACC	NUM
LR-ASGL	100%	145
ASGL-CMI	96%	124
SGL	88%	95
GL	88%	248
Elastic net	92%	70
lasso	96%	19

Table 2
Average prediction accuracy and average number of selected genes for six methods on acute leukemia data.

Method	Average-ACC	Average-NUM
LR-ASGL	97.2%(0.026 7)	188.9(47.7)
ASGL-CMI	96.4%(0.029 5)	102.8(26.7)
SGL	92.8%(0.036 8)	235.4(85.4)
GL	91.6%(0.044 1)	579.2(174.6)
Elastic net	94%(0.038 9)	67.4(8)
lasso	93.2%(0.050 1)	15(2.8)

respectively. The implementation process of LR-ASGL, SGL, GL, EN, and lasso is similar to that of the simulation study (see section 4). The implementation of ASGL-CMI is shown in the literature [30]. In order to evaluate the performance of the six methods, we used ACC, NUM for a single experiment, and Average-ACC, Average-NUM for multiple experiments.

5.1. Classical data partition setting

It is a classic method to divide acute leukemia data into 38 training set and 34 test set [36]. The training set contains 27 ALL samples and 11 AML samples, and the test set contains 20 ALL samples and 14 AML samples.

We firstly performed RPCA via R package *rpca* (see, subsection 3.1) on training set, obtained clean matrix $\widehat{D}$ and noise matrix $\widehat{E}$. The scatter plots of R^2-τ and connectivity-τ are shown in Fig. 1. It is shown that $R^2 > 0.8$ and the network connectivity is not significantly reduced when the soft thresholding power $\tau > 7$. Hence, we select $\tau = 7$. Then, we divided genes into groups on a clean matrix by using WGCNA with the soft thresholding power $\tau = 7$. According to the identified modules, 3571 most significant genes are divided into 17 groups. The division results are as follows: 216 genes in black module, 678 genes in blue module, 273 genes in brown module, 63 genes in cyan module, 232 genes in green module, 91 genes in green-yellow module, 46 genes in grey module, 44 genes in light cyan module, 183 genes in magenta module, 50 genes in midnight-blue module, 194 genes in pink module, 109 genes in purple module, 231 genes in red module, 63 genes in salmon module, 64 genes in tan module, 786 genes in turquoise module, 248 genes in yellow module. Fig. 2 shows 17 identified modules with colors according to the dynamic tree cut method.Experiment results show that LR-ASGL, EN, and lasso have the same prediction accuracy of 94.1%, which is 5.9%, 5.9%, and 11.8% higher than that of ASGL-CMI, SGL, GL, respectively. It seems that LR-ASGL not only selects more genes than EN and lasso but also has no obvious improvement in prediction accuracy. However, the prediction performance of the machine learning model is highly dependent on the data partition. We will illustrate the advantages of LR-ASGL in prediction and gene selection via random data partition.

Table 3
Forty-four genes selected by LR-ASGL on acute leukemia data.

Module color	Gene symbol
Magenta	VPREB1 CD79A CTSD LYZ AZU1 IFI16 MEF2A ATRX NME4 RRAGA DGKD CD19 CDK9 RFTN1 FAH CSRP1 SON SP3
Red	TCF3 DNTT CST3 CXCL8 CCND3 APLP2 LTC4S YWHAQ RBBP4 LEPROT TOP2B LMNA EDEM1 MYB ZAP70 CYFIP2 IL7R PLCB2 CDKN2D TIMP2 TRAC TGFB1 STOM PTPRCAP SPTAN1 CBFB

5.2. Random data partition setting

In order to reduce the interference of data partition on the prediction performance, we randomly partitioned the data ten times. To this end, we randomly selected two-thirds of the samples to train the model, the rest to test the model, and repeated the process ten times. For the sake of the class balance, 31 ALL samples and 16 AML samples were selected as training samples, and the rest 25 samples were selected as test samples. According to the biomedical viewpoint, gene pathways or gene groups of cancer should not change with different sample partitions. Therefore, we adopted the same gene modules as in subsection 5.1.

To compare with the results of the classical data partition, we showed the results of one experiment in ten random data partitions. The prediction accuracy and the number of the selected genes for six methods on acute leukemia data via one random partition are shown in

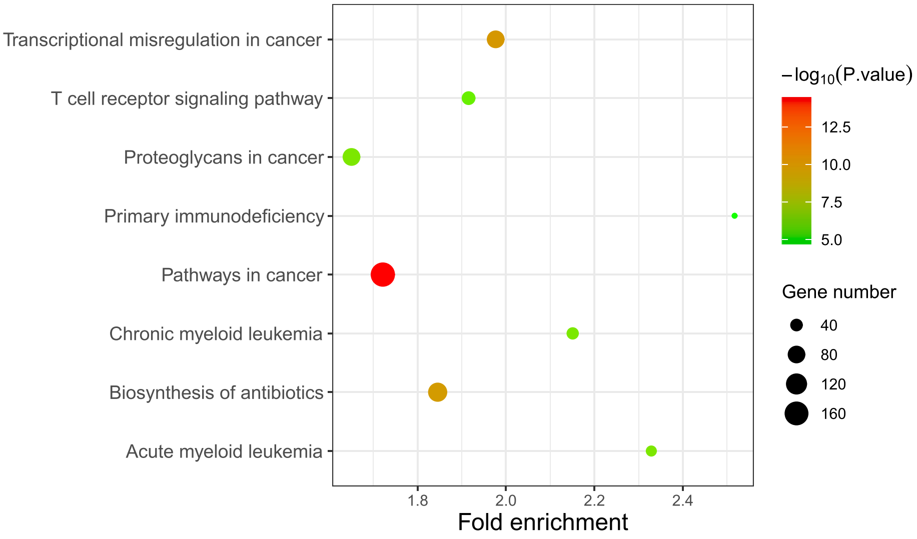

Fig. 3. Bubble plot of selected genes.

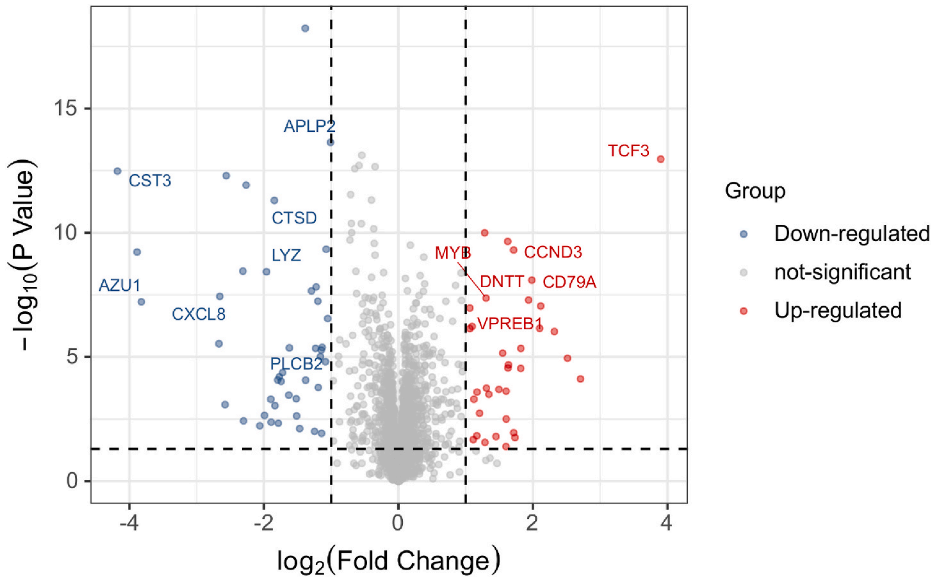

Fig. 4. Volcano plot of differential expression genes.

Table 1. It shows that LR-ASGL achieves the highest prediction accuracy of 100%, which is 4%, 12%, 12%, 8% and 4% higher than that of ASGL-CMI, SGL, GL, EN, and lasso, respectively. It should be noted that this is only the result of an experiment with random data division, which can not prove that the proposed method is superior to other methods in prediction accuracy. In fact, this does not ensure that 100% prediction accuracy can be obtained for each random division experiment. In order to avoid the randomness of a single experiment and prove the superiority of the proposed method in terms of prediction accuracy, we used the average prediction accuracy for all ten results to evaluate the performance of statistical learning models (see Table 2).

The average prediction accuracy and the average number of the selected genes for six methods on acute leukemia data via ten random partitions are shown in Table 2. It shows that the average prediction accuracy of LR-ASGL is 97.2%, which is 0.8%, 4.4%, 5.6%, 3.2% and 4% higher than ASGL-CMI, SGL, GL, EN, and lasso, respectively. Furthermore, the LR-ASGL method has the smallest standard deviation (0.026 7) with the highest average prediction accuracy. The main reason for the improved accuracy may come from the adoption of robust principal component analysis and the robust evaluation criterion. Compared with ASGL-CMI, EN, and lasso, LR-ASGL selects more genes and has no advantage in the number of the selected genes. In the next subsection, we will demonstrate that the genes selected by LR-ASGL are associated with acute leukemia and have biological interpretability.

5.3. *Gene screening*

Genes are usually regulated by each other and the co-regulation of genes plays an important role in the cause of leukemia. Here we will demonstrate the grouped gene selection performance of the LR-ASGL. 75 genes are repeatedly selected 9 or more times by LR-ASGL for ten experiments. 44 genes were selected 10 times, and 31 genes are selected 9 times. These genes are thought to be related to leukemia. As can be seen from Fig. 3, 75 genes are involved in multiple pathways: Chronic myeloid leukemia, Transcriptional misregulation in cancer, T cell receptor signaling pathway, and so on.We list 44 genes with the highest frequency of occurrence in Table 3. It shows that these genes come from the two modules (magenta and red). This indicates that LR-ASGL has the ability to select genes in a group.

Fig. 4 shows the volcano plot of differential expression genes by setting P Value ≤ 0.05, $\log_2$FoldChange ≥ -1 and $\log_2$FoldChange ≤ 1. Among the 44 genes shown in Table 3, VPREB1, TCF3, DNTT, CD79A, CCND3, and MYB are significantly up-regulated; CXCL8, AZU1, CST3, APLP2, PLCB2, CTSD, and LYZ are significantly down-regulated. It should be pointed out that there is no significant difference in the expression of 31 genes. But that doesn't mean they don't matter. They interact with differentially expressed genes to produce a grouping effect. The functional annotations of genes are obtained through literature retrieval in the NCBI database at https://www.ncbi.nlm.nih.gov/gene. By referring to the functional annotations, we found that the selected genes are highly related to the cause of leukemia. As a representative, three significantly up-regulated genes are listed below.

VPREB1 encodes the iota polypeptide chain that is associated with the Ig-mu chain to form a molecular complex that is thought to regulate Ig gene rearrangements in the early steps of B-cell differentiation. Deletions of the VPREB1 gene have been observed in childhood B-cell acute lymphoblastic leukemia (B-ALL). Mangum et al. found that VPREB1 focal deletions are common in B-ALL and occur independent of V(D)J light chain recombination [38].

CD79A encodes the Ig-alpha protein of the B-cell antigen component. Surface Ig non-covalently associates with two other proteins, Ig-alpha and Ig-beta, which are necessary for the expression and function of the B-cell antigen receptor. CD79A functions in and has a high degree of specificity for B-cell differentiation. Kozlov et al. found that CD79A represented the aberrant presence of a B-cell antigen in leukemias of distinct myeloid linage [39].

TCF3 encodes a member of the E protein (class I) family of helix-loop-helix transcription factors. Deletion of this gene or diminished activity of the encoded protein may play a role in lymphoid malignancies. This gene is also involved in several chromosomal translocations that are associated with lymphoid malignancies, including pre-B-cell acute lymphoblastic leukemia, childhood leukemia, and acute leukemia. Ma et al. have shown that TCF-3 gene silencing inhibits Eca-109 cell growth and proliferation suppresses cell cycle progression and promotes apoptosis [40].

6. Conclusion

In this paper, the logistic regression with adaptive sparse group lasso penalty (LR-ASGL) is presented for dealing with the problems of noise, gene grouping, and adaptive gene selection on cancer gene expression profile data. The robust principal component analysis is used to decompose data into a clean matrix and a sparse noise matrix. Weighted gene co-expression network analysis is performed on the clean matrix for dividing genes into groups. Both noise information and structural information are used to construct the LR-ASGL model, and the solving algorithm is presented. Simulation studies in three noise settings, i.e., Gaussian noise, uniformly distributed noise, and mixed noise, are provided to test the effectiveness of LR-ASGL.The experimental results on acute leukemia data via random partition illustrate that the proposed robust evaluation criterion significantly improves the accuracy of cancer diagnosis and adaptively selects the related genes in groups. For ten times random partition, the average prediction accuracy of LR-ASGL is 97.2%, which is 0.8%, 4.4%, 5.6% 3.2% and 4% higher than ASGL-CMI, SGL, GL, EN, and lasso, respectively. The volcano and bubble plots of the selected genes are presented to verify the adaptive grouped gene selection ability of LR-ASGL. As a representative, three significantly upregulated gene functional annotations are provided in this paper.

Data availability statement

Acute leukemia data s available at: http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43.

Declaration of competing interest

None declared.

Acknowledgments

The solving algorithm of this paper refers to the following R packages: *rpca*, *WGCNA*, *asgl*. Sincerely thank Maciek Sykulski for providing the *rpca* package, Horvath et al. for providing the *WGCNA* package, Simon et al. for providing the *SGL* package, Hastie et al. for providing the *glmnet* package, Jeffrey Daniel for providing the *asgl* package on GitHub. This work was supported by the National Natural Science Foundation of China (61 203 293), the Scientific and Technological Project of Henan Province (212 102 210 140).

References

[1] L.A. Torre, F. Bray, R.L. Siegel, J. Ferlay, J. Lortet-Tieulent, A. Jemal, Global cancer statistics, 2012, CA-A Cancer Journal for Clinicians 65 (2) (2015) 87–108, https://doi.org/10.3322/caac.21262.

[2] W. Tang, Z.J. Liao, Q. Zou, Which statistical significance test best detects oncomirnas in cancer tissues? an exploratory analysis, Oncotarget 7 (51) (2016) 85613–85623, https://doi.org/10.18632/oncotarget.12828.

[3] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914, https://doi.org/10.1093/bioinformatics/16.10.906.

[4] R. Diaz-Uriarte, S.A. de Andres, Gene selection and classification of microarray data using random forest, BMC Bioinf. 7 (1) (2006) 3, https://doi.org/10.1186/1471-2105-7-3.

[5] J.T. Li, Y.D. Wang, Y.M. Cao, C.S. Xu, Weighted doubly regularized support vector machine and its application to microarray classification with noise, Neurocomputing 173 (2016) 595–605, https://doi.org/10.1016/j.neucom.2015.08.002.

[6] L. Chen, X. Pan, X.H. Hu, Y. Zhang, S.P. Wang, T. Huang, Y.D. Cai, Gene expression differences among different msi statuses in colorectal cancer, Int. J. Cancer 143 (7) (2018) 1731–1740, https://doi.org/10.1002/ijc.31554.

[7] J.T. Li, Y.Y. Wang, X.K. Song, H.M. Xiao, Adaptive multinomial regression with overlapping groups for multi-class classification of lung cancer, Comput. Biol. Med. 100 (2018) 1–9, https://doi.org/10.1016/j.compbiomed.2018.06.014.

[8] L. Chen, Z.D. Li, T. Zeng, Y.H. Zhang, Y.D. Cai, Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes, Front. Mol. Biosci. 7 (2020) 604794, https://doi.org/10.3389/fmolb.2020.604794.

[9] H. Vikalo, B. Hassibi, A. Hassibi, A statistical model for microarrays, optimal estimation algorithms, and limits of performance, IEEE Trans. Signal Process. 54 (6) (2006) 2444–2455, https://doi.org/10.1109/TSP.2006.873716.

[10] L. Klebanov, A. Yakovlev, How high is the level of technical noise in micarray data? Biol. Direct 2 (2007) 9, https://doi.org/10.1186/1745-6150-2-9.

[11] H.S. Wang, G.D. Li, G.H. Jiang, Robust regression shrinkage and consistent variable selection through the LAD-lasso, J. Bus. Econ. Stat. 25 (3) (2007) 347–355, https://doi.org/10.1198/073500106000000251.

[12] S. Lambert-Lacroix, L. Zwald, Robust Regression through the Huber's criterion and adaptive lasso penalty, Electron. J.Stat. 5 (2011) 1015–1053, https://doi.org/10.1214/11-EJS635.

[13] J. Pannu, N. Billor, Robust group-lasso for functional regression model, Commun. Stat. Simulat. Comput. 46 (5) (2015) 3356–3374, https://doi.org/10.1080/03610918.2015.1096375.

[14] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1) (1996) 267–288, https://doi.org/10.1111/j.1467-9868.2011.00771.x.

[15] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2) (2005) 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[16] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. Roy. Stat. Soc. B 68 (1) (2006) 49–67, https://doi.org/10.1111/j.1467-9868.2005.00532.x.

[17] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse group lasso, J. Comput. Graph Stat. 22 (2) (2013) 231–245, https://doi.org/10.1080/10618600.2012.681250.

[18] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. Roy. Stat. Soc. B 67 (1) (2010) 91–108, https://doi.org/10.1111/j.1467-9868.2005.00490.x.

[19] S.G. Ma, X. Song, J. Huang, Supervised group lasso with applications to microarray data analysis, BMC Bioinf. 8 (1) (2007) 60, https://doi.org/10.1186/1471-2105-8-60.

[20] L. Meier, S.A. van de Geer, P. Bühlmann, The group lasso for logistic regression, J. Roy. Stat. Soc. B 70 (1) (2008) 53–71, https://doi.org/10.1111/j.1467-9868.2007.00627.x.

[21] Y.M. Li, B. Nan, J. Zhu, Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure, Biometrics 71 (2) (2015) 354–363, https://doi.org/10.1111/biom.12292.

[22] B. Koch, D.M. Vock, J. Wolfson, Covariate selection with group lasso and doubly robust estimation of causal effects, Biometrics 74 (1) (2018) 8–17, https://doi.org/10.1111/biom.12736.

[23] Z.J. Liao, D.P. Li, X.R. Wang, L.S. Li, Q. Zou, Cancer diagnosis through isomir expression with machine learning method, Curr. Bioinf. 13 (1) (2018) 57–63, https://doi.org/10.2174/1574893611666160609081155.

[24] B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, Stat. Appl. Genet. Mol. Biol. 4 (2005) 17, https://doi.org/10.2202/1544-6115.1128.

[25] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, BMC Bioinf. 9 (1) (2008) 559, https://doi.org/10.1186/1471-2105-9-559.

[26] H. Zou, The adaptive lasso and its oracle properties, J. Am. Stat. Assoc. 101 (476) (2006) 1418–1429, https://doi.org/10.1198/016214506000000735.

[27] Q. Zheng, C. Gallagher, K.B. Kulasekera, Robust adaptive lasso for variable selection, Commun. Stat. Theor. Methods 46 (9) (2017) 4642–4659, https://doi.org/10.1080/03610926.2015.1019138.

[28] H.S. Wang, C.L. Leng, A note on adaptive group lasso, Comput. Stat. Data Anal. 52 (12) (2008) 5277–5286, https://doi.org/10.1016/j.csda.2008.05.006.

[29] K.N. Fang, X.Y. Wang, S.W. Zhang, J.P. Zhu, S.G. Ma, Bi-level variable selection via adaptive sparse group lasso, J. Stat. Comput. Simulat. 85 (13) (2015) 2750–2760, https://doi.org/10.1080/00949655.2014.938241.

[30] J.T. Li, W.P. Dong, D.Y. Meng, Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information, IEEE ACM Trans. Comput. Biol. Bioinf 15 (6) (2018) 2028–2038, https://doi.org/10.1109/TCBB.2017.2761871.

[31] M.A. van De Wiel, T.G. Lien, W. Verlaat, W.N. Van Wieringen, S.M. Wilting, Better prediction by use of co-data: adaptive group-regularized ridge regression, Stat. Med. 35 (3) (2016) 368–381, https://doi.org/10.1002/sim.6732.

[32] S. Chakraborty, C.B. Begg, R.L. Shen, Using the "Hidden" Genome to Improve Classification of Cancer Types, Biometrics, 2020, https://doi.org/10.1111/biom.13367 published online.

[33] H.D. Yi, Q. Zhang, C.J. Lin, S.G. Ma, Information-incorporated Gaussian Graphical Model for Gene Expression Data, Biometrics, 2021, https://doi.org/10.1111/biom.13428 published online.

[34] E.J. Candes, X.D. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM 58 (3) (2011) 1–37, https://doi.org/10.1145/1970392.1970395.

[35] J. Liu, Y. Xu, C. Zheng, H. Kong, Z. Lai, RPCA-Based tumor classification using gene expression data, IEEE ACM Trans. Comput. Biol. Bioinf 12 (4) (2015) 964–970, https://doi.org/10.1109/TCBB.2014.2383375.

[36] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537, https://doi.org/10.1126/science.286.5439.531.

[37] S. Dudoit, J. Fridlyand, T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (457) (2002) 77–87, https://doi.org/10.1198/016214502753479248.

[38] D.S. Mangum, J. Downie, C.C. Mason, M.S. Jahromi, D. Joshi, V. Rodic, M. Mueschen, N. Meeker, N. Trede, J.K. Frazer, Y. Zhou, C. Cheng, S. Jeha, C.-H. Pui, C.L. Willman, R.C. Harvey, S.P. Hunger, J.J. Yang, P. Barnette, C. G. Mullighan, R.R. Miles, J. Schiffman, VPREB1 deletions occur independent of

lambda light chain rearrangement in childhood acute lymphoblastic leukemia, Leukemia 28 (1) (2014) 216–220, 0.1038/leu.2013.223.

[39] I. Kozlov, K. Beason, C. Yu, M. Hughson, CD79a expression in acute myeloid leukemia t(8;21) and the importance of cytogenetics in the diagnosis of leukemias with immunophenotypic ambiguity, Cancer Genet. Cytogenet. 163 (1) (2005) 62–67, https://doi.org/10.1016/j.cancergencyto.2005.06.002.

[40] J. Ma, X.B. Wang, R. Li, H. Xuan, F. Wang, X.H. Li, Z.P. Zhang, L. Tan, L. Li, RNAi-mediated TCF-3 gene silencing inhibits proliferation of Eca-109 esophageal cancer cells by inducing apoptosis, Biosci. Rep. 37 (6) (2017), BSR20170799, https://doi.org/10.1042/BSR20170799.