Mathematical Biosciences and Engineering

MBE, 20 (3): 5352–5378.
DOI: 10.3934/mbe.2023248
Received: 04 August 2021
Revised: 09 December 2021
Accepted: 18 January 2022
Published: 11 January 2023

http://www.aimspress.com/journal/MBE

Research article

Identification of influential observations in high-dimensional survival data through robust penalized Cox regression based on trimming

Hongwei Sun[1,2,*], Qian Gao[2], Guiming Zhu[1], Chunlei Han[1], Haosen Yan[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:** Email: hwsun2000@163.com, tongwang@sxmu.edu.cn;
Tel: +86-535-6913408; +86-351-4135397.

Abstract: Penalized Cox regression can efficiently be used for the determination of biomarkers in high-dimensional genomic data related to disease prognosis. However, results of Penalized Cox regression is influenced by the heterogeneity of the samples who have different dependent structure between survival time and covariates from most individuals. These observations are called influential observations or outliers. A robust penalized Cox model (Reweighted Elastic Net-type maximum trimmed partial likelihood estimator, Rwt MTPL-EN) is proposed to improve the prediction accuracy and identify influential observations. A new algorithm AR-Cstep to solve Rwt MTPL-EN model is also proposed. This method has been validated by simulation study and application to glioma microarray expression data. When there were no outliers, the results of Rwt MTPL-EN were close to the Elastic Net (EN). When outliers existed, the results of EN were impacted by outliers. And whenever the censored rate was large or low, the robust Rwt MTPL-EN performed better than EN. and could resist the outliers in both predictors and response. In terms of outliers detection accuracy, Rwt MTPL-EN was much higher than EN. The outliers who "lived too long" made EN perform worse, but were accurately detected by Rwt MTPL-EN. Through the analysis of glioma gene expression data, most of the outliers identified by EN were those "failed too early", but most of them were not obvious outliers according to risk estimated from omics data or clinical variables. Most of the outliers identified by Rwt MTPL-EN were those who "lived too long", and most of them were obvious outliers according to risk estimated from omics data or clinical variables. Rwt MTPL-EN can be adopted to detect influential observations in high-dimensional survival data.

Keywords: influential observations; penalized Cox regression; high-dimensional survival data; trimming; heterogeneity; robust; omics data; outliers

1. Introduction

The determination of biomarkers in high-dimensional genomic data related to disease prognosis can be used to understand the pathogenesis of disease prognosis, so as to find new therapeutic drugs targeted to improve the prognosis of patients. Biomarkers can also be used to predict the patient's prognosis, and to provide individualized treatment for patients. So the discovery of biomarkers associated with prognosis has become an active research area. The study has two challenges. One is that the prognostic data tends to contain censored survival time; the other is the high-dimensional data, in which the number of variables is much higher than the sample size. The Cox proportional hazard model is widely used to model censored survival time, to screen for associate factors and to establish a prognostic prediction model, but it is not suitable for high-dimensional data. Penalized Cox regression can solve the problem of prognostic factors screening and prediction model establishment in high-dimensional data. For example, Liu, Z., M. Li, Q. Hua, Y. Li and G. Wang [1] used L_1-penalized (i.e., LASSO-type) Cox regression to identify non-coding RNAs related to breast cancer prognosis. Patients were stratified based on risk scores. Shen, X. Y., X. P. Liu, C. K. Song, Y. J. Wang, S. Li and W. D. Hu [2] also used the LASSO-type penalized Cox proportional hazard regression model to finally identify two genes related to lung adenocarcinoma survival. Penalized Cox regression has also recently been used to identify driver genes for bladder cancer prognosis [3].

However, the prediction accuracy of these methods is often influenced by the heterogeneity of samples from cancer patients [4,5]. The main known source of the heterogeneity is genomic instability [6]. For example, genomic instability is a prominent source of genetic diversity within tumours, generating a diverse cell population that can be subject to selection in a given micro-environmental or therapeutic context. Some individuals have different dependent structures between survival time and covariates, which means that these patients may show different mechanisms from most individuals. Individuals with poor survival prediction by fitting Cox regression, “died too early” or “lived too long” as compared to the estimated survival probabilities for their covariate pattern composed of selected associated factors revealed by most individuals [7]. These observations are called influential observations or outliers. These outliers, especially long-term survivors have a great impact on Cox regression [8]. On the other hand, it is very important to detect outliers in survival data, because the analysis of individuals with long or short survival will lead to the identification of new prognostic factors [7,9]. Peng, S., H. Dhruv, B. Armstrong, B. Salhia, C. Legendre, J. Kiefer, J. Parks, S. Virk, A. E. Sloan and Q. T. Ostrom [10] compared the integrated genomics glioma “outliers” (patients with long-term survival and short-term survival to discover the molecular markers with different prognoses after standard treatment. Therefore, individualized treatment can avoid treatment failure caused by wrong treatment.

So it is very important to identify outliers in survival data. On the one hand, a robust model can be obtained to improve the prediction accuracy of the model by removing the influence of outliers. On the other hand, outliers that are identified may reveal hidden information on the covariate and probably be worth studying further.

Because outliers can affect parameter estimation, residual analysis cannot be directly used for

outlier identification, and there is a very high probability of masking, that is, the lack of identification of true outlier recognition [11]. Therefore, robust estimates are a prerequisite for distance-based outlier detection procedures [12]. For a low dimensional data, a robust estimation method, least trimmed square (LTS), was proposed by Rousseeuw, P. J. [11]. 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. Farcomeni, A. and S. Viviani [12] proposed a robust Cox regression model based on trimming to analyze survival data with outliers. They fitted Cox model by trimming the individuals with small contribution to partial likelihood function, to obtain a robust estimation which is not affected by outliers.

However, there are few studies on robustness in high dimensional survival analysis dealing with omics data. Carrasquinha, E., A. Veríssimo, M. B. Lopes and S. Vinga [9] studied the outliers in high dimensional survival analysis. Elastic net (EN) and LASSO were adopted to screen variables from high-dimensional data to low-dimensional data, and two low-dimensional robust Cox models were used to identify outliers. Then, rank product test and ensemble were used to combine the outliers identified by the two methods. The disadvantage of this method is that screening variables from high-dimensional data may be affected by outliers because EN and LASSO are not robust.

For high dimensional survival analysis, a robust penalized Cox model based on trimming is proposed in this study. Compared with Carrasquinha, E., A. Veríssimo, M. B. Lopes and S. Vinga [9], we considered robustness in high-dimensional data directly, to avoid the influence of outliers on dimensionality reduction.

In this article, a robust EN-type penalized Cox model based on trimming is proposed and the algorithm to find the solution of the model is described in section 2. In Section 3, the results of simulation studies and the analysis of glioma gene expression data are described. We conclude with a discussion in section 4 and a conclusion in section 5.

2. Materials and methods

2.1. Robust penalized Cox regression model based on trimming

Assume there are n observations in the follow-up study. δ_i represents the outcome of object i, where $\delta_i = 1$ represents event, and $\delta_i = 0$ represents censorship. Times of n objects when they died or censored are denoted as $t_1 < t_2 < \cdots < t_n$. Let $R(t_i)$ be the number of people alive at time t_i, that is, the number of people at risk.

Let us consider a penalized Cox observation model with outliers. Let I be a pure set without outliers,

$$\begin{cases} h(t, \boldsymbol{X}_i) = h_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}) & i \in I \\ h(t, \boldsymbol{X}_i) = \lambda_i(t) & i \notin I \end{cases} \tag{1}$$

That means individuals in pure set I are subject to Cox proportional risk model, and the outliers outside of I obey an unknown, unspecified risk function $\lambda_i(t)$. So the outlier cannot provide useful information for the estimation of β.

We proposed a penalized Cox model based on trimming (Maximum trimmed partial likelihood estimator, MTPL-EN). Assuming the trimmed ratio is $1 - \eta$ ($0 < \eta < 1$). The number of retained

observations is $h = \lfloor n\eta \rfloor$ accordingly, where $\lfloor \cdot \rfloor$ means round down. Then the maximum partial likelihood function of the MTPL-EN model is:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{MTPL-EN} &= argmax_{\boldsymbol{\beta}}(lnL - h\lambda \sum_{j=1}^{p} P_{\alpha}(\beta_j)) \\ &= argmax_{\boldsymbol{\beta}} \left(\sum_{i=1}^{h} \delta_i (\boldsymbol{x}'_{i_l}\boldsymbol{\beta} - ln \sum_{i_l \in R\left(t_{i_l}\right)} \exp\left(\boldsymbol{x}'_{i_l}\boldsymbol{\beta}\right)) - h\lambda \sum_{j=1}^{p} P_{\alpha}(\beta_j) \right), \\ P_{\alpha}(\boldsymbol{\beta}) &= (1-\alpha)\tfrac{1}{2}\left|\left|\boldsymbol{\beta}\right|\right|_2 + \alpha\left|\left|\boldsymbol{\beta}\right|\right|_1, \; i_l \in \{1,2,\cdots,n\}, \end{aligned} \tag{2}$$

where $\lambda \geq 0$ is a penalty parameter, and α, which is the mixing proportion of the ridge and LASSO penalties, takes a value in [0, 1]. The EN tends to select groups of correlated variables. The robust penalized Cox regression based on trimming is to find the subset whose sample size is *h*, which corresponding regularized partial likelihood function is the maximum. The corresponding regularized partial likelihood estimation of the subset is denoted as $\hat{\boldsymbol{\beta}}^{\boldsymbol{MTPL\text{-}EN}}$.

Generally, the proportion of outliers is unknown in practice, and the selection of $1-\eta$ is higher than the expected proportion of outliers. However, too high trimmed proportion usually leads to estimated asymptotic variance inflate and the reduction of the estimation efficiency. We adopted the trimmed ratio $1-\eta = 0.25$ in this article. In this study, after $\hat{\boldsymbol{\beta}}^{\boldsymbol{MTPL\text{-}EN}}$ was estimated, reweighted step was considered to detect outliers in the data. And then estimation on dataset in which outliers were removed again to further improve efficiency.

2.2. Algorithm

Maximizing (2) is equivalent to find an optimal subset with regularized maximum partial likelihood function in all subsets with a sample size $h = \lfloor n(1-\eta) \rfloor$. It is impossible to search exhaustively because the number of all subsets $\binom{n}{h}$ is too large and $\binom{n}{h}$ increases rapidly with the increase of the sample size. The computational burden of finding the optimal subset by exhaustive method in such a huge subset is considerable.

This kind of optimization is very common in robust statistics, and the method usually used is a repeated concentration-steps algorithm, which also known as the C-step algorithm [13]. In the C-step algorithm, it is necessary to separate the individual's contribution to the objective function at each step of the iteration. But the objective function of Cox regression is the partial likelihood function, the contribution of the partial likelihood function corresponding to each observation is difficult to be decomposed from the partial likelihood function of the complete set. Because the risk set of an observation is related to the survival time order of other individuals. If an observation is included or excluded, and corresponding risk set changes accordingly. Especially for the censored individuals, its contribution to the likelihood function is reflected in the denominator of the partial likelihood function, and its contribution is also difficult to be separated. So $\hat{\boldsymbol{\beta}}^{\boldsymbol{MTPL\text{-}EN}}$ cannot be solved directly

by C-step algorithm. Another reason why the C-step algorithm cannot be used is that, in the penalized regression, the regulator λ of each step in C-step algorithm needs to be re-determined. The objective function does not decrease with iteration, so it does not necessarily converge on a subset of the optimal objective function. Farcomeni, A. and S. Viviani [12] applied the acceptance-rejection algorithm (proposed by Chakraborty, B. and P. Chaudhuri [14]) to solve the robust Cox model based on trimming in low-dimensional cases. In the acceptance-rejection algorithm, at each iteration, the candidate samples are randomly taken from the remaining samples. So, the direction of the iteration is random, and it leads to the slower convergence of the algorithm. In this study, the residuals of each individual are used to replace each individual's contribution to the partial likelihood function. The algorithm used to solve $\hat{\boldsymbol{\beta}}^{\boldsymbol{MTPL\text{-}EN}}$ is a combination of the acceptance-rejection algorithm and the C-step algorithm, which is called as C-step algorithm based on acceptance-rejection step (AR-Cstep).

2.2.1. AR-Cstep algorithm

2.2.1.1. Deviance residual

Therneau, T. M., P. M. Grambsch and T. R. Fleming [15] proposed martingale residuals for Cox model without time-dependent covariates based on counting process.

$$\widehat{M}_i = \delta_i - \widehat{H}_0(t_i)\exp\left(\hat{\beta}^T X\right) \tag{3}$$

where δ_i is the censoring indicator, and $\widehat{H}_0(t_i)$ is the baseline cumulative risk function. Martingale residuals can be interpreted as the difference between the number of observed events of an individual and the expected number of events under the Cox model, that is, the part that is not predicted by the model and exceeds the estimated number of events.

Martingale residuals can reveal that, compared to other observations with the same covariate, observations that do not fit the model well, that is, those who live too long ($\widehat{M}_i$ is a large negative value) and fail too early ($\widehat{M}_i$ is close to 1).

The martingale residual is asymmetric, and its range is $(-\infty, 1]$. The martingale residual is transformed to approximate its normal distribution to obtain the deviance residual:

$$\hat{d}_i = sign\left(\widehat{M}_i\right)\left\{-2\left[\widehat{M}_i + \delta_i\, log\left(\delta_i - \widehat{M}_i\right)\right]\right\}^{1/2} \tag{4}$$

Deviance residuals are more symmetrical than martingale residuals. A large negative residual indicates that the number of observed events is less than the model's predicted number, that is, an outlier that "lived too long". And a large positive residual indicates the number of observed events is larger than the model's predicted number, that is, an outlier that "died too early", relative to other observations with the same covariate. Some simulation studies have shown that for "lived too long" type outliers, can be detected both by the deviance residuals and martingale residuals. While for "failed too early" type outliers, can be identified only by the deviance residuals [15,16].

2.2.1.2. Find regression estimates from subsamples, and get the residuals of all samples.

In C-step algorithm, it is necessary to perform regression estimation on the subset, and then substitute the estimated coefficients into the complete set to obtain the contribution of each

individual to the criterion function. In the penalized Cox model, the baseline risk needs to be estimated before the residuals of all individuals are obtained based on the estimated coefficient $\hat{\boldsymbol{\beta}}$. However, we found that the baseline risk is greatly affected by the outliers through simulation experiments, so the baseline risk needs to be estimated through subset without outliers. Then it can be extended to that of all individuals.

At the k-step iteration, H_k denotes the current subset containing h observations. $\widehat{\boldsymbol{\beta}}_{H_k}$ is the solution of the penalty Cox regression based on H_k, and the corresponding partial likelihood function is recorded as $log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$. The baseline risk is estimated by

$$H_0(t) = \sum_{i:t_i<t} \frac{d_i}{\sum_{k\in R(t_i)} exp\,(\boldsymbol{x}_k'\widehat{\boldsymbol{\beta}}_{H_k})} \tag{5}$$

Under $\widehat{\boldsymbol{\beta}}_{H_k}$, for each sample in the subset H_k, the baseline risk $\hat{H}_0(t_i)$ is estimated according to the formula (5), $i \in H_k$. For the samples of the remaining set $C(H_k)$ of the subset H_k, the baseline risk is estimated according to the following method. We assumed that the baseline risk follows the Weibull distribution, that is the baseline risk $H_0(t)$ and survival time t are linear after transformed by logarithmic transformation,

$$ln\big(H_0(t)\big) = \beta_0 + \beta_1 ln\,(t). \tag{6}$$

Using the samples of the subset H_k, $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated by linear regression (6). Then the survival time t_i of the sample in $C(H_k)$ is substituted to (6) to obtain the corresponding baseline risk $\hat{H}_0(t_i)$, $i \in C(H_k)$. In order to make the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ robust, median regression is used here. The objective function of the usual least squares estimation is to minimizes the sum of the squared residuals, while the objective function of the median regression minimizes the median of the squared residuals, which is a robust estimate with high breakdown point [11].

According to the baseline risk of each individual, the deviance residual $\hat{d}_i$ of each individual is obtained according to formulas (3) and (4), $i = 1, 2, \cdots, n$. The h individuals with the smallest absolute value of residuals are selected to form a new subset H_{cand}. In order to keep the censoring rate of the subset the same as that of the full set, H_{cand} is composed of h_1 individuals with the smallest residuals among individuals who have an outcome, and h_0 individuals with the smallest residuals among the censored individuals. Let $n = n_1 + n_2$ where n_1 individuals have an outcome and survival time of n_2 individuals are censored. Let $h_1 = \lfloor (n_1 + 1)\,\eta \rfloor$ and $h_0 = h - h_1$, where $\lfloor . \rfloor$ means round down, and $1-\eta$ is the trimmed rate.

Then, the penalized Cox regression is performed on H_{cand} to get the estimate $\widehat{\boldsymbol{\beta}}_{H_{cand}}$. And then the partial likelihood function $log\ell(\widehat{\boldsymbol{\beta}}_{H_{cand}}, H_{cand})$ corresponding to H_{cand} under the estimated $\widehat{\boldsymbol{\beta}}_{H_{cand}}$ is obtained. If $log\ell(\widehat{\boldsymbol{\beta}}_{H_{cand}}, H_{cand}) \geq log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$, then $H_{k+1} = H_{cand}$, and then the above process is continued on H_{k+1} and keep iterating. If $log\ell(\widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}, H_{cand}) < log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$, to avoid falling into the local optimum, the idea of accept-reject algorithm is adopted in our study. U is a random number which obey Bernoulli distribution with $p = e^{\tau_k(log\ell(\widehat{\boldsymbol{\beta}}_{H_{\text{cand}}}, H_{cand}) - log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k))}$. If $U = 1$ then $H_{k+1} = H_{cand}$. And if $U = 0$, then $H_{k+1} = H_k$.

In order to make the algorithm reach the global optimal value, multiple initial subsets can be

taken. In order to ensure that the initial subset does not contain outliers, the sample size should be smaller. But too small a sample size will cause inaccurate estimation, especially too few samples with outcomes (i.e., not censored) in the subset will make it impossible to estimate the coefficient of the penalized Cox regression. In this study, the sample size of the initial subset was 20, and a total of 500 subsets were selected randomly. A two-step AR-Cstep were executed on these 500 subsets. Ten subsets with the largest partial likelihood function from the 500 subsets after iteration were selected. Then the AR-Cstep algorithm was ran on 10 subsets until convergence. In the 10 convergent subsets, the subset with the smallest partial likelihood function is selected as the optimal subset, which is represented by H_{opt}. And a penalized Cox regression was ran on H_{opt} to obtain the solution $\widehat{\boldsymbol{\beta}}_{opt}$.

Table 1. AR-Cstep Algorithm.

k indicates the number of iterations, and r indicates that the current maximum likelihood value has not changed after r iterations.

While (k<=kmax & r<=2)

do

$$\widehat{\boldsymbol{\beta}}_{H_k} \coloneqq arg\,\max_{\beta}\{\log\ell(\beta, H_k) - n\lambda\sum_{j=0}^{n} P_\alpha(\beta_j)\}$$

For $i \in H_k$, $\widehat{H}_0(t_i)$ is estimated based on $\widehat{\boldsymbol{\beta}}_{H_k}$.

According to $\ln\big(H_0(t)\big) = \beta_0 + \beta_1 \ln(t)$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained through median regression.

For $i \in C(H_k)$, $\ln\left(\widehat{H}_0(t_i)\right) = \hat{\beta}_0 + \hat{\beta}_1 \ln(t_i)$.

For $i \in \{1, 2, \cdots, n\}$, deviance residual $\hat{d}_i$ is obtained through $\widehat{H}_0(t_i)$ and $\exp(\hat{\beta}^T X)$.

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

$|\hat{\mathrm{d}}_{i_1}| \le |\hat{\mathrm{d}}_{i_2}| \le \cdots \le |\hat{\mathrm{d}}_{i_{n_1}}|$, i_k is the index of the individual who has an outcome.

$|\hat{\mathrm{d}}_{j_1}| \le |\hat{\mathrm{d}}_{j_2}| \le \cdots \le |\hat{\mathrm{d}}_{j_{n_0}}|$, j_k is the index of the censored individual.

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

$$\widehat{\boldsymbol{\beta}}_{H_{cand}} \coloneqq arg\,\max_{\beta}\{\log\ell(\beta, H_{cand}) - n\lambda\sum_{i=0}^{n} P_\alpha(\beta_j)\}$$

$log\ell(\widehat{\boldsymbol{\beta}}_{H_{cand}}, H_{cand})$ is obtained based on $\widehat{\boldsymbol{\beta}}_{H_{cand}}$.

If $log\ell(\widehat{\boldsymbol{\beta}}_{H_{cand}}, H_{cand}) \ge log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$ then

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

If $log\ell(\widehat{\boldsymbol{\beta}}_{H_{cand}}, H_{cand}) < log\ell(\widehat{\boldsymbol{\beta}}_{H_k}, H_k)$ then

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

U is a random number which obey Bernoulli distribution with p.

if U=1 then

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

else

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

end

end

end

2.2.2. Reweighted step

In this article, we choose the subset of size $h=\lfloor n\eta \rfloor$ where $\eta=0.75$. So, the trimmed rate 1-η 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 $\widehat{\boldsymbol{\beta}}_{opt}$. Then these outliers are excluded and a new subset $H_{reweighted}$ is obtained. Then EN-type penalized Cox regression is applied to $H_{reweighted}$ to get the solution $\widehat{\boldsymbol{\beta}}_{reweighted}$. Usually, the size of $H_{reweighted}$ is larger than h, such that more samples can improve the performance of $\widehat{\boldsymbol{\beta}}_{reweighted}$ compared to $\widehat{\boldsymbol{\beta}}_{opt}$.

In order to make the estimation of baseline risk $H_0(t)$ in deviance residuals more accurate and not affected by outliers, $H_0(t)$ was obtained on H_{opt}. For samples other than H_{opt}, the baseline risk is also estimated by Eq (6).

After obtaining the deviance residuals $\hat{d}_i$ of each observation, define a binary weight for the i-th observation as follows:

$$w_i = \begin{cases} 1 & if\ |\hat{d}_i| \le \Phi^{-1}(1-\delta) \\ 0 & if\ |\hat{d}_i| > \Phi^{-1}(1-\delta) \end{cases}, \tag{3-5}$$

where $\Phi(x)$ is the distribution function of a standard normal distribution. We set $\delta = 0.005$ and $\Phi^{-1}(1-\delta) = 2.57$. That means observations with residuals beyond 2.57 are regarded as outliers. $H_{\text{reweighted}}$ is composed of observations that are not flagged as outliers.

The reweighted estimator $\widehat{\boldsymbol{\beta}}_{reweighted}$ is the solution of the penalized Cox regression based on $H_{reweighted}$, It is called Reweighted MTPL-EN (Rwt MTPL-EN). To distinguish them, the unweighted $\widehat{\boldsymbol{\beta}}_{opt}$ is called Raw MTPL-EN.

2.2.3. Choice of the regulator parameter and standardization of predictors

We select λ over a grid of values in the interval $(0,\ \lambda_{\max}]$ as discussed by Breheny and Huang [14].

$$\hat{\lambda}_{max} = \max_{j\in\{1,2,\cdots,p\}} n^{-1}X_j' t$$

t is the survival time. In iteration step of AR-Cstep, we take a grid with steps of size $0.05\hat{\lambda}_{\max}$ and $\alpha = 0.5$ to reduce the computational burden. In the reweighted step, we take a grid with steps of size $0.01\hat{\lambda}_{\max}$ of λ to derive the solution $\widehat{\boldsymbol{\beta}}_{opt}$ and $\widehat{\boldsymbol{\beta}}_{reweighted}$. 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 Cox regression. Standardization mainly aims to eliminate the influence of dimension and quantity of a predictor. However, mean and standard deviation computed from all sample are not robust with outliers. In the algorithm described above, penalized Cox regression is applied to subsample 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 Cox regression.

2.3. Simulation settings

In this section, we will compare the accuracy of elastic net, Raw MTPL-EN and Rwt MTPL-EN in variable selection, outlier identification and prediction by simulating whether there are outliers in survival data, symmetrical and asymmetric outliers, different censoring ratios, and outliers in the response and predictors.

2.3.1. Simulation data generation

The simulation data in this study is based on Bender, R., T. Augustin and M. Blettner [17]. Assuming that the baseline risk $H_0(t)$ obeys the exponential distribution of λ = 1, the survival time T can be generated by the following formula,

$$T = \frac{-ln\,(U)}{exp\,(\boldsymbol{X^T\beta})}$$

where U obeys the uniform distribution on [0,1]. The censoring time is generated by an exponential distribution with mean $Vexp\,(\boldsymbol{X^T\beta})$, where V obeys the uniform distribution of $[CL, CU]$. And different values of CL and CU give rise to different censoring ratios.

2.3.2. Simulation scenario setting

Considering that the omics data is usually high-dimensional data, and the high group correlation caused by gene interaction, the following settings are made. We set the sample size n = 300, the number of independent variables p = 1000, where the independent variables X follows a p-dimensional multivariate normal distribution $N(0, \Sigma_p)$. The correlation structure of the independent variables is assumed to be block correlation. A block includes 50 independent variables. The correlation structure within the block is $corr(x_i, x_j) = \rho^{|i-j|}, i \neq j$, $\rho = 0.9$. There is no correlation between blocks. The related structure within blocks is as follows:

$$\Sigma = \begin{pmatrix} \Sigma_\rho & 0 & \cdots & 0 \\ 0 & \Sigma_\rho & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \Sigma_\rho \end{pmatrix}$$

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \cdots & \rho^{49} \\ \rho & 1 & \cdots & \rho^{48} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{49} & \rho^{48} & \cdots & 1 \end{pmatrix}_{50\times 50}$$

Considering that the effect of a single gene on prognosis is often low, the absolute value of the effect size is set between 0.3 and 0.8. The non-zero regression coefficients are set in the first 12 block groups, and the regression coefficient of each block group is set as:

$$\boldsymbol{\beta}_{1-50}^T = (\underbrace{0.3, \cdots, .0.3}_{5}, \underbrace{0, \cdots, 0}_{45}), \quad \boldsymbol{\beta}_{51-100}^T = (\underbrace{-0.4, \cdots, .-0.4}_{5}, \underbrace{0, \cdots, 0}_{45}),$$

$\boldsymbol{\beta}^T_{101-150}$=($\underbrace{0.5,\cdots,.0.5}_{5},\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{151-200}$=($\underbrace{-0.6,\cdots,.-0.6}_{5}$, $\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{201-250}$=($\underbrace{0.7,\cdots,.0.7}_{5}$, $\underbrace{0,\cdots,0}_{45}$),

$\boldsymbol{\beta}^T_{251-300}$=($\underbrace{-0.8,\cdots,.-0.8}_{5}$, $\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{301-350}$=($\underbrace{0.8,\cdots,.0.8}_{5}$,$\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{351-400}$=($\underbrace{-0.7,\cdots,.-0.7}_{5}$,$\underbrace{0,\cdots,0}_{45}$),

$\boldsymbol{\beta}^T_{401-450}$=($\underbrace{0.6,\cdots,.0.6}_{5}$, $\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{451-500}$=($\underbrace{-0.5,\cdots,.-0.5}_{5}$, $\underbrace{0,\cdots,0}_{45}$), $\boldsymbol{\beta}^T_{501-550}$=($\underbrace{0.4,\cdots,.0.4}_{5}$, $\underbrace{0,\cdots,0}_{45}$),

$\boldsymbol{\beta}^T_{551-600}$=(-0.3,-0.3,-0.3,-0.3,-0.3, $\underbrace{0,\cdots,0}_{45}$), and $\boldsymbol{\beta}^T_{601-1000}$ is set to the zero vector.

Considering that the censoring rate of survival data is often large, 35% censoring rate was set, which was also close to 37.58% censoring rate in glioma data analysis. In addition, censoring rates of 15%, 25% and 45% were also set to see the effect of censoring rate on the results.

Referring to the setting of the outliers by Farcomeni, A. and S. Viviani [12], the maximum and minimum values of $exp\,(\boldsymbol{X}^T\boldsymbol{\beta})$ were recorded as HR_{high} and HR_{low} respectively. Then, the observations with the proportion of ε were randomly selected from the data set, and their corresponding $exp\,(\boldsymbol{X}^T\boldsymbol{\beta})$ were changed to $u_iHR_{low}+(1-u_i)HR_{high}$. u_i is a random number obeying the Bernoulli distribution with parameter p. Two cases of symmetric outliers (p = 0.5) and asymmetric outliers (p = 0.9 and p = 0.1) are set. Figure 1 shows the setting of outliers in the simulation data.

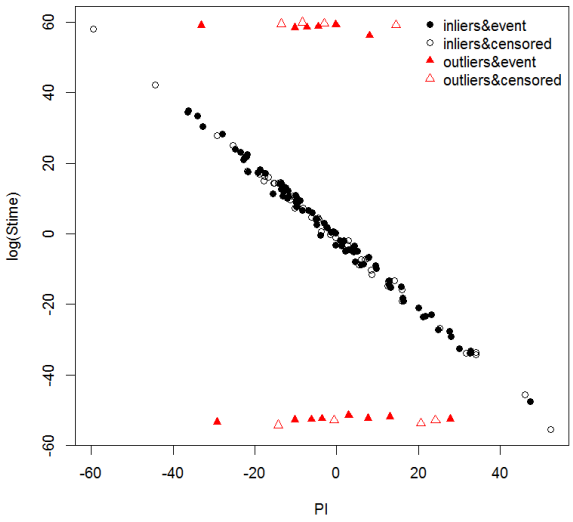

Figure 1. Graphical representation of outlier settings in simulated data (scatter plot of logarithmic survival time and prognostic index *PI*). (Notes: $PI = exp\,(\boldsymbol{X}^T\boldsymbol{\beta})$. Black solid dots: normal points with outcomes; black hollow dots: censored normal points; red solid triangles: outliers with outcomes; red hollow triangles: censored outliers.)

For elastic net, the R package "glmnetUtils" is used, which is an extension of the R package glmnet. The parameter α can also be cross-validated. Among them, the choice of λ adopts default choice in the package. That is, 100 steps of equal step size on a log scale from $\lambda_{\min}$ to $\lambda_{\max}$ are generated, where $\lambda_{\min} = 0.01$ and $\lambda_{\max}$ is the minimum value of λ that makes all regression coefficients 0. The range of α is [0.1,1], and 10 alpha values are generated with a step size of 0.1. For MTPL-EN, the trimmed rate is set to 25%.

Simulation scenario 1: n = 300, p = 1000, ε = 10%, censoring ratio 35%, Four cases are set:

(1) No outliers.

(2) Symmetric outliers, that is, outliers that "lived too long" relative to the prognosis index and outliers that "failed too early" relative to the prognosis index each account for 50%.

(3) Asymmetric outliers, that is, outliers that "lived too long" account for 90%.

(4) Asymmetric outliers, that is, outliers that "failed too early" account for 90%.

Simulation scenario 2: n = 300, p = 1000, ε = 10%, symmetrical outliers. Four cases are set:

(1) Censored rate was 15%.

(2) Censored rate was 25%.

(3) Censored rate was 35%.

(4) Censored rate was 45%.

Simulation Scenario 3: It was mainly to see the impact on the performance of the elastic net and MTPL-EN when outliers deviate from the main data in the response, or when the deviation also occurs in the predictors. n = 300, p = 1000, ε = 10%, symmetrical outliers, censoring ratio 35%.

Case A: 10% individuals are randomly selected, with a 50% probability of min (h (t)), which is the minimum value of the risk function, and a 50% probability of max (h (t)), which is the maximum value of the risk function.

Case B: Others are the same as case A, but there is a 50% probability that the outliers are min (h (t)) / exp (15) and a 50% probability is max (h (t)) * exp (15).

Case C: Others are the same as A, and the independent variables of the outliers are set to follow the independent N (3,1) distribution.

Case D: Others are the same as B, and the independent variables of the outliers are set to follow independent N (3,1) distributions.

In case B compared with case A, we can see that the survival time of the outlier that "lived too long" is longer, and that of the outlier that "died too early" is shorter. Compared with case A, in case C, the outliers are shifted to the right. Compared to case A, in case D, outliers are shifted to the right and deviates also farther from the main data in the response. Graphical representation is shown in Figure S1 in the supplementary file.

Training data and test data were generated according to the above sampling schemes. Training data were generated to fit the model and evaluate the accuracy of variables selection and outlier detection. And test data were generated to evaluate the prediction of the model. The test data were generated without outliers. For each setting, we calculated the average of the performance measures over 100 simulation replicates implemented in *R* software. Codes is available on Github (https://github.com/hwsun2000/MTPL-EN).

3. **Results**

3.1. Simulation results

3.1.1. Results of scenario 1

Figure 2 shows the performance of EN, Raw MTPL-EN, and Rwt MTPL-EN when there were no outliers, and when there were 10% outliers in the data.

Here we used two indicators *Sn* (sensitivity) and *FPR* (false positive rate) in the screening

test[18]. The outliers to be identified is regarded as patients to be detected in the screening test. *Sn* represents the proportion of truly outliers among outliers that identified. *FPR* represents the proportion of normal samples that are determined to be outliers. A detailed description of the indicators is provided in the supplementary file. *PSR* (Positive Selection Rate) indicates the proportion of real disease-related biomarkers that are screened out, and *FDR* (False Discovery Rate) indicates the proportion of biomarkers screened out that are not related to the disease. We used a comprehensive indicator *GM* [19,20], which is the geometric mean of (*PSR* and (1−*FDR*)) to measure the accuracy of variable selection. High *PSR*s and low *FDR*s will give high *GM*s, which indicates high accuracy of variable selection. A detailed description of the indicators is provided in the supplementary file.

When there were no outliers, EN performed best. From both the accuracy of variable selection and the log-likelihood function, the results of Rwt MTPL-EN were close to EN. It showed that Rwt MTPL-EN didn't lose much efficiency for datasets without outliers.

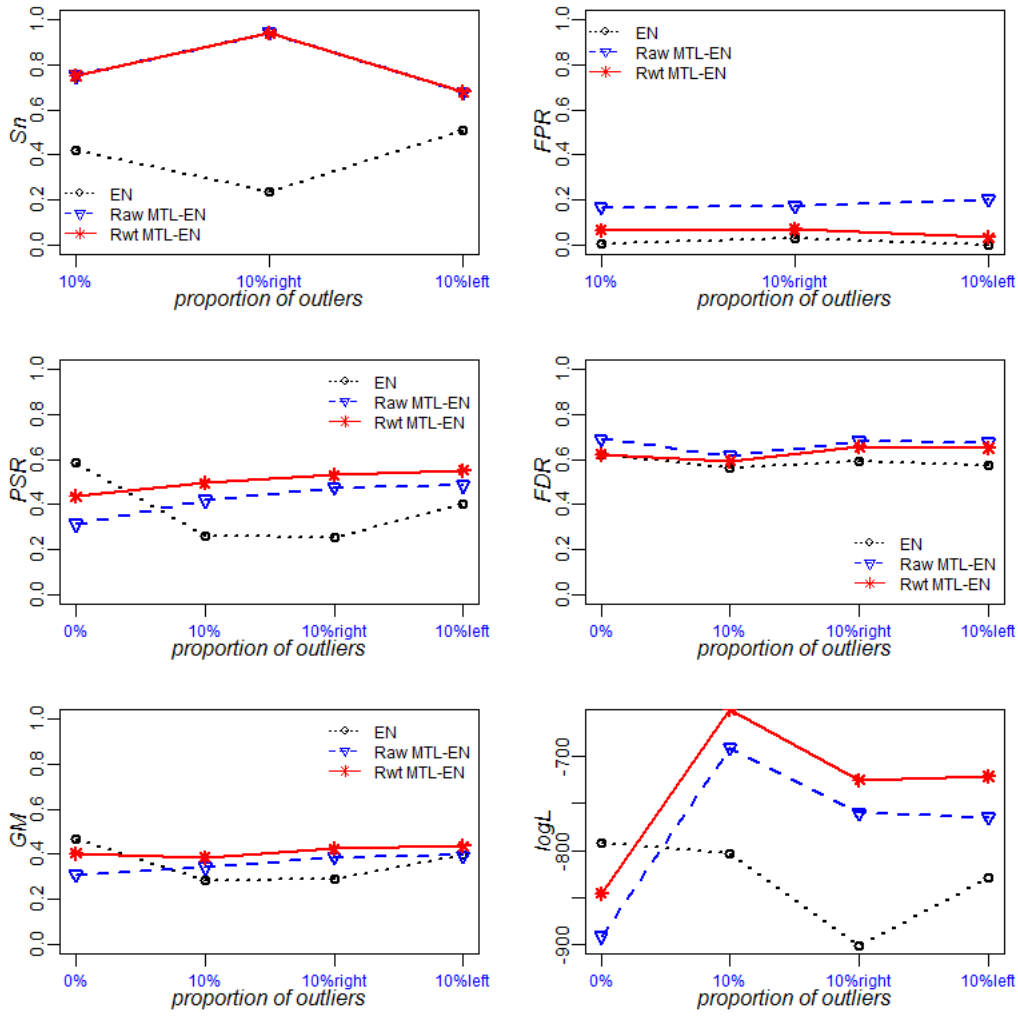

Figure 2. Comparison of results between EN and MTPL-EN under scenario 1 (n = 300, p = 1000).

When there are outliers, the estimation of EN is affected by outliers. Compared with the absence of outliers, the *FDR* changes little, but its *PSR* is reduced by about 60%, indicating that EN will miss a significant percentage of non-zero variables. The performance of MTPL-EN is better than EN. Compared with the case the absence of outliers, the *FDR* of Rwt MTPL-EN remains basically unchanged, and the *PSR* exceeds 50%. From the perspective of the comprehensive indicator *GM*, it shows that the accuracy of Rwt MTPL-EN in variables selection remained stable.

As far as the accuracy of outlier identification is concerned, 25% of samples were considered as outliers by Raw MTPL-EN, which is higher than the percentage of outliers actually set in the simulation experiment. So, its *FPR* was high with 18%. For Rwt MTPL-EN, outliers were further identified the through the reweighted step, so that the number of outliers was less than that in Raw MTPL-EN. And its FPR was less than 7%. The sensitivity is higher than EN (*Sn*, 0.75 vs 0.42), see Figure 2. Taken together, detected outliers was the least by EN and has the smallest *FPR*, but its sensitivity is also the lowest. The outliers identified by Rwt MTPL-EN has its *FPR* less than 7%, and its sensitivity reached more than 70%.

According to the log likelihood function, when there were outliers, the log-likelihood function of MTPL-EN were much larger than that of EN. Log-likelihood function of of Rwt MTPL-EN were higher than that of Raw MTPL-EN and EN, indicating that Rwt MTPL-EN had the highest prediction accuracy when there were outliers.

As can be seen from Figure 2, compared with the case of symmetric outliers, En behaved differently under asymmetric outliers. When 90% of the outliers were outliers that "lived too long" relative to their prognosis index, the accuracy of the variable selection of EN was worse (*GM* 0.28 vs 0.29), and the ability to identify outliers becomes worse (*Sn* 0.24 vs 0.42, *FPR* 0.03 vs 0.003). However, the accuracy of variables selection of Rwt MTPL-EN was improved (*GM*, 0.43 vs 0.39) and the ability of outliers identification was also improved (*Sn* 0.94 vs 0.75, *FPR* 0.07 vs 0.07).

When 90% of outliers were samples that "died too early" relative to their prognosis index, the impact on EN was smaller than that of symmetrical outliers. The accuracy of variable selection was higher (*GM* 0.40 vs 0.28) and the ability of outliers identification remains almost unchanged (*Sn* 0.51 vs 0.42, *FPR* 0.001 vs 0.003). The effect on the robust Rwt MTPL-EN was similar to that of the symmetrical outliers. The accuracy of variable selection remains was improved (*GM* 0.44 vs 0.39), and the ability to identify outliers decreases (*Sn* 0.66 vs 0.75, *FPR* 0.03 vs 0.07).

Compared to the outliers that "failed too early", the outliers that "lived too long" made EN perform worse. It was easy for Rwt MTPL EN to identify outliers that "lived too long", and the performance of variable selection is not affected by outliers.

In short, when there were no outliers, the results of MTPL-EN were close to EN. When there were 10% outliers, the accuracy of variable selection and prediction of MTPL-EN were higher than those of EN. More outliers were identified by Rwt MTPL-EN and its *FPR*s were within 7%. Compared to outliers "failed too early", outliers that "lived too long" made EN perform worse. And it was easier for Rwt MTPL-EN to identify outliers that "lived too long". In any case, the performance of Rwt MTPL-EN remained stable under outliers.

3.1.2. Simulation Scenario 2

As can be seen from Figure 3, when the censored ratio was 15% and 25%, the accuracy of variable selection of Rwt MTPN-EN is higher than that when the censored ratio is 35%. And the

accuracy of outlier identification is also improved.

When the censored ratio is 45%, the accuracy of variable selection and outlier identification decreases. But EN was lower than Rwt MTPN-EN on the accuracy of variable selection, outlier identification or prediction. There is a large gap in the ability to identify outliers between EN and Rwt MTPN-EN. When the censoring ratio is 45%, the sensitivity of outliers detection of EN is 0.370, and the *FPR* is 0.007. The sensitivity of Rwt MTL-EN is 0.748, and the *FPR* is 0.088. This showed that whenever the censored rate was large or low, using Rwt MTL-EN could identify outliers accurately.

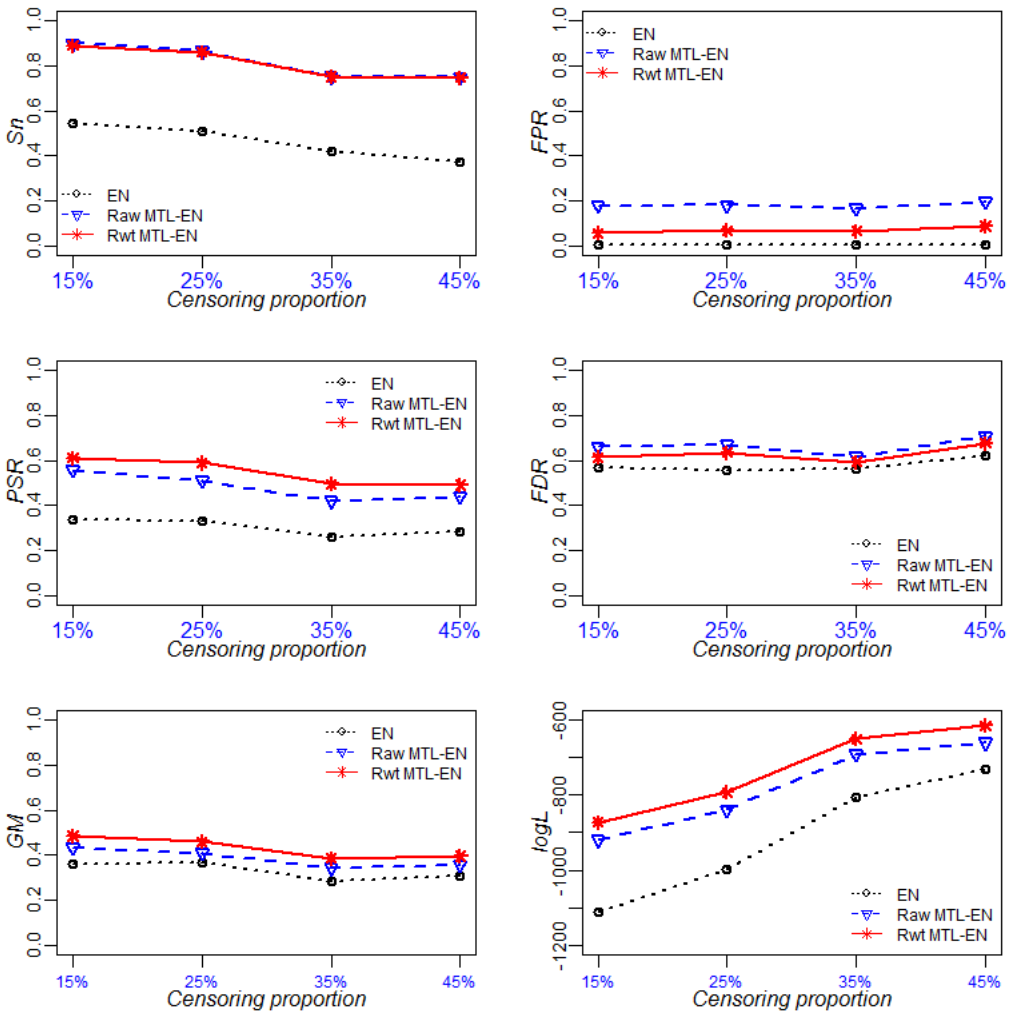

Figure 3. Comparison of results between EN and MTPL-EN under different censoring proportions (n = 300, p = 1000).

3.1.3. Simulation scenario 3

As can be seen from Figure 4 that, in case B, when outliers deviate in the response farther, outlier detection results of Rwt MTPL-EN were better than that of Case A (*Sn* 0.83 vs 0.75, *FPR* 0.07 vs 0.07). The *PSR* and *FDR* of variable selection of Rwt MTL-EN and EN changed little. The

prediction accuracy of EN was lower than that in case A (*Log-likelihood*, -876 vs -804). Rwt MTL-EN performed better than EN in terms of outlier detection, variables selection and prediction.

In case C, when outliers also deviated in the predictors compared to case A, outliers detection accuracy of EN decreased (*Sn* 0.344 vs 0.417, *FPR* 0.001 vs 0.003). Variables selection accuracy of EN changed (*PSR* 0.578 vs 0.261, *FDR* 0.806 vs 0.563). That is due to the number of variables selected by EN increased from 47.9 to 185.6 on average, which is much larger than 60, the number of true non-zero variables. The number of variables selected by Rwt MTPL-EN changed little, and the outlier identification and log-likelihood functions also remained stable.

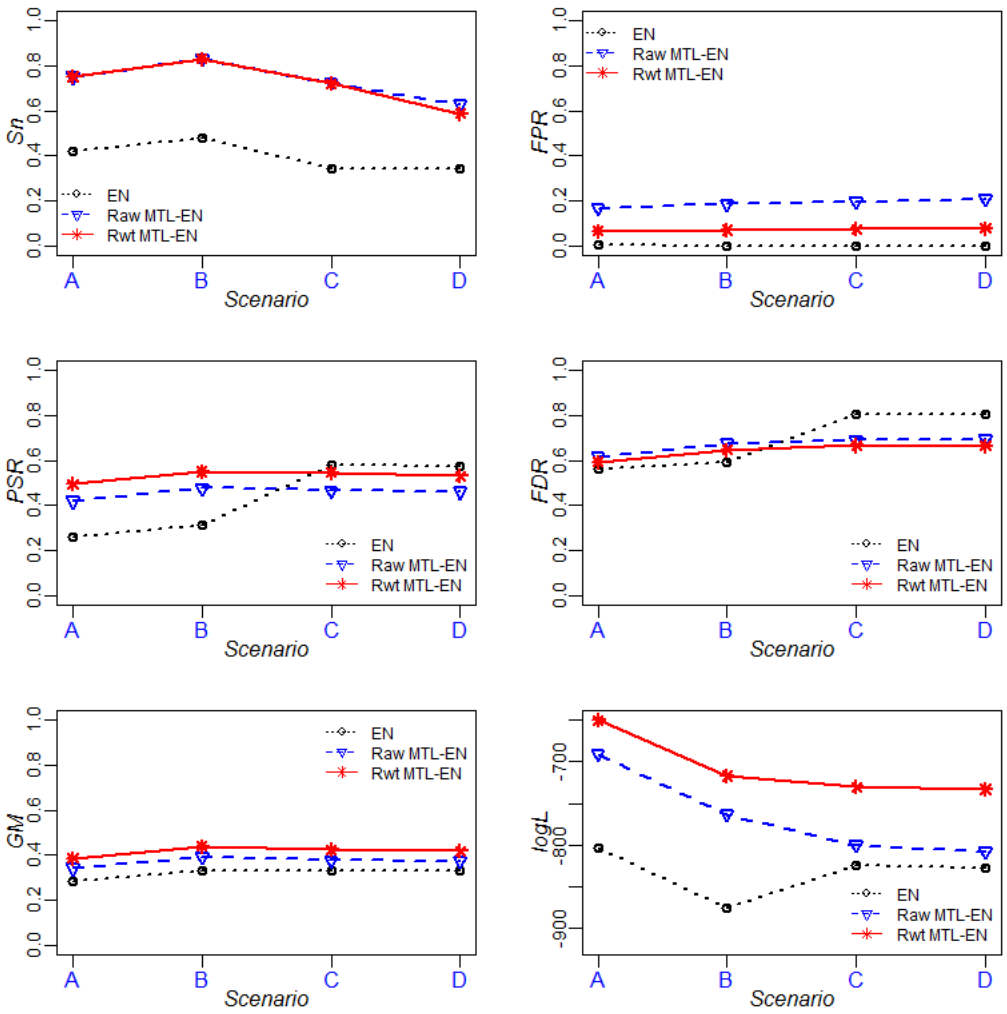

Figure 4. Comparison of the results between EN and MTPL-EN when there were outliers in the response or (and) predictors (n = 300, p = 1000).

In case D, when outliers deviate in the response farther than case C, the results of EN, Raw MTL-EN and Rwt MTL-EN were similar to that in case C.

As can be seen from above simulation experiments, the results of EN were impacted by outliers. However, outlier detection accuracy of Rwt MTPL-EN was much higher than that of EN. The prediction and variables selection accuracy were higher than EN, which showed that Rwt MTPL-EN

could resist the outliers in both predictors and response.

In addition, we simulated the situations when the trimmed ratio 1-η was 5%, 15%, 25%, and 35%, respectively. The results are shown in Figure S2 of the supplementary materials. As can be seen from Figure S2, since the proportion of outliers is 10%, Rwt MTPL-EN had little difference in the accuracy of outliers detection and variable selection when the trimmed ratio 1-η is 15%, 25% and 35%. But when the trimmed ratio was 35%, the prediction accuracy was slightly lower. When the trimmed ratio was 5%, the sensitivity of outliers detection decreased. So Rwt MTPL-EN remained stable when the trimmed ratio was higher than the outlier ratio. However, when the proportion of trimmed samples was much higher than that of outliers, the accuracy of outliers detection and variable selection remained stable, but the accuracy of prediction decreased slightly. When the trimmed ratio was lower than the outlier ratio, the accuracy of outlier detection was affected. In practice, the percentage of outliers is usually less than 25%, so we recommend a trimmed ratio of 25% to make it larger than the percentage of outliers, so that the result is not affected by outliers. If, based on practical experience, the percentage of outliers in the data is likely to exceed 25%, the trimmed ratio should be increased.

We also simulated situations of different sample sizes n = 300 and 500, and different dimensions p = 600 and 1000, as shown in Figure S3 of the supplementary material. The accuracy of variable selection and outlier detection was the best when n = 500 and p = 600, and the accuracy decreased when the sample size n decreased or the dimension p increased.

3.2. Results of the analysis on a TNBC dataset

3.2.1. Glioma gene expression dataset

Example data were obtained from gene microarray expression data of 301 patients with glioma in China (CGGA, http://www.cgga.org.cn/), and 298 patients were analyzed after removing 3 patients with missing survival time. Rwt MTPL-EN and EN were used to screen the genes that affect the prognosis of gliomas, and to detect the possible outliers. The results of the two methods were compared. The parameter setting of two methods were the same as that of simulation evaluation.

The median survival time of the data set was 38.3 months, and the range was 0.7~158.7 months, and the censored rate was 37.58%. There were 116 cases of glioma WHO II, 66 cases of III, 126 cases of IV, and 3 cases of missing WHO classification. The clinical variables include TCGA subtype, PRS type, histological type, grade, gender, age, radiation status, chemotherapy status, IDH mutation status, 1p19q co deletion status and so on. Gene expression microarray expression profiles were analyzed on an Agilent Whole Human Genome Array. There were 19,416 genes in the gene expression data.

3.2.2. Application of EN to glioma dataset

In this study, EN and Rwt MTPL-EN were applied to glioma gene expression dataset (n = 298, p = 19416). The parameter setting was the same as that set in the simulation study.

Eighty-seven genes were identified by EN, which are listed in supplementary file.

As can be seen from Table 2 and Figures 5 and 6 that, there were 12 outliers identified by EN, and only one outlier was that “lived too long” relative to the prognosis index estimated by EN. The

remaining 11 were all outliers that "died too early" relative to the prognostic index estimated by EN. As shown in Figure 5, 11 outliers were also the dead individuals with the shortest survival time in all samples.

Table 2. outliers identified by EN and their corresponding values on clinical variables.

ID	Time(day)	Status	Residual	PI**	WHO grade	Histological type	Age	IDH mutation
CGGA_444	225	1	2.02	-0.33	IV	GBM	70	Wildtype
CGGA_640*	3922	0	-2.66	0.64	IV	GBM	55	Wildtype
CGGA_649	147	1	1.99	0.64	IV	GBM	58	Wildtype
CGGA_713	27	1	2.22	2.06	IV	GBM	35	Wildtype
CGGA_764	67	1	2.36	1.33	IV	GBM	50	Wildtype
CGGA_346	104	1	2.21	0.81	IV	GBM	45	Wildtype
CGGA_1011	109	1	2.18	0.73	IV	GBM	46	Wildtype
CGGA_662	284	1	2.29	-1.39	II	O	59	Mutant
CGGA_1059	21	1	2.71	1.51	II	rA	54	Mutant
CGGA_406	90	1	2.06	1.48	III	rAA	24	Mutant
CGGA_719	101	1	2.44	0.41	IV	sGBM	51	Wildtype
CGGA_1068	68	1	2.15	1.51	IV	sGBM	29	Mutant

*：Outliers that "lived too long" relative to prognosis index that estimated by EN. **：$PI = \boldsymbol{x}_i'\boldsymbol{\beta}^{EN}$.

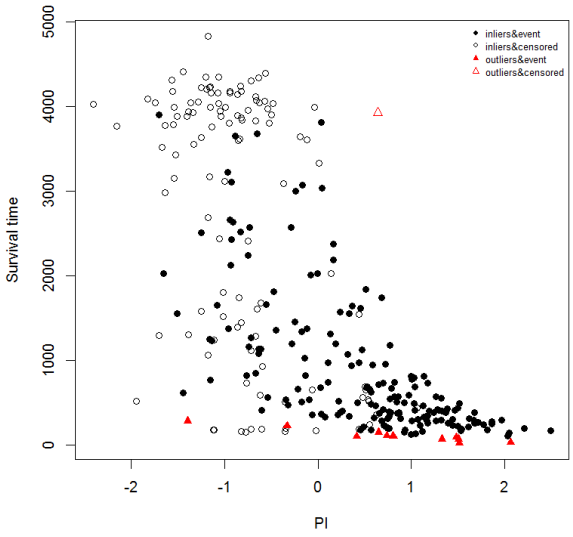

Figure 5. Scatter plot of the prognostic index PI estimated by EN and survival time. (Note: black solid dots: normal points with outcomes; black hollow dots: censored normal points; red solid triangles: outliers with outcomes; red hollows: censored outliers, $PI = \boldsymbol{x}_i'\boldsymbol{\beta}^{EN}$)

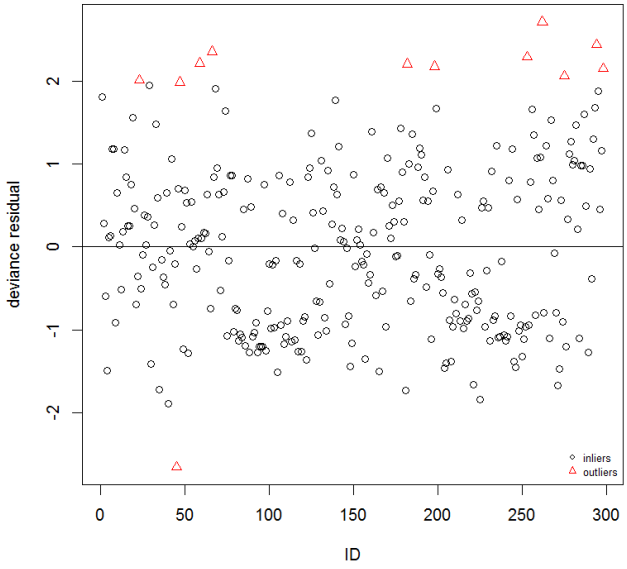

Figure 6. Deviance residuals corresponding to the model estimated by EN. (Note: black hollow dots: normal points; red hollow triangles: outliers)

However, from Figure 5, most of the outliers who “failed too early” had PI greater than 0. And even 5 outliers whose PI were greater than 1, which showed a relatively high risk of death estimated from gene expression data, and their survival times were relatively short. So, it is not appropriate to consider them as “outliers” estimated from gene expression data. From Table 2, according to the clinical characteristics of the outliers, 8 of the 11 individuals were classified as grade IV by WHO. In terms of age, six individuals were over 50 years old, and eight were over 40 years old. From the perspective of histological types, 8 are glioblastomas (GBM or sGBM) with poor prognosis. From the IDH mutation, 7 are wild-type with poor prognosis. So, for most of the outliers that "failed too early”, their illness was also serious, and it was also unreasonable to identify them as outliers from their clinical characteristic data.

3.2.3. Results of Rwt MTPL-EN applied to glioma dataset

As can be seen from Table 3, EN identified 87 genes and 12 outliers. Its log-likelihood functions was −833.4, and the C index was 0.842. The AUC corresponding to the median survival time was 1164 days was 0.793. Rwt MTPL-EN screened 56 genes and identified 18 abnormal points, and the corresponding prediction index was lower than EN. However, for a subset with 18 outliers removed, the log-likelihood, C-index, and AUC of Rwt MTPL-EN were all higher than those of EN.

As can be seen from Table 4, Figures 7 and 8, the number of outliers identified by Rwt MTPL-EN was 18, of which 3 were outliers that are “failed too early” relative to the prognosis index estimated by Rwt MTPL-EN. There are 15 outliers that “lived too long”. Of the 15 outliers that “lived too long”, 13 had prognostic indices greater than 0, 10 were greater than 1, and 5 were greater than 2. Among them, the absolute value of 8 residuals were greater than 4, indicating that most outliers who "lived too long" had a higher risk of death estimated from gene expression data. This means that there were different correlation patterns between the prognosis and covariate values in

these outliers from that in most individuals. According to the clinical characteristics, three of the 15 outliers that "lived too long" were classified as IV and 6 were on level III. In terms of IDH mutation types, there were 9 wild types with poor prognosis. This means that most of the 15 individuals who "lived too long" were obviously outliers with high risk but "lived too long".

Table 3. Number of genes, outliers, prediction estimated by EN and Rwt MTPL-EN.

Methods	Genes	Outliers	Log-likelihood	Log-likelihood (subset*)	C-index	C-index (subset*)	AUC#	AUC# (subset*)
EN	87	12	-833.4	-781.3	0.831	0.842	0.793	0.800
Rwt MTPL-EN	56	18	-850.2	-768.8	0.785	0.851	0.773	0.803

* Subset refers to the subset after removing 18 outliers identified by Rwt MTPL-EN; #: ROC curve corresponding to a median survival time of 1164 days.

Table 4. Outliers identified by Rwt MTPL-EN and corresponding values in clinical variables.

ID	Time(day)	status	devres	PI**	WHO grade	Histology	Age	IDH_mutation
CGGA_11*	155	1	2.86	0.98	IV	GBM	57	Wildtype
CGGA_225	1741	1	-4.48	2.33	IV	GBM	32	Wildtype
CGGA_640	3922	0	-14.20	3.10	IV	GBM	55	Wildtype
CGGA_365	3593	0	-3.05	0.15	II	A	32	Mutant
CGGA_331	1638	1	-3.38	1.99	III	AA	27	Wildtype
CGGA_352	4304	0	-3.69	0.26	III	AA	18	Wildtype
CGGA_393	2190	1	-3.43	1.58	III	AOA	65	Mutant
CGGA_438	686	1	-2.81	3.06	III	AOA	53	Mutant
CGGA_577	3901	0	-4.91	0.98	III	AA	38	Wildtype
CGGA_275	4387	0	-6.59	1.39	II	O	43	Wildtype
CGGA_323	4338	0	-4.99	0.86	II	O	48	Wildtype
CGGA_484	4116	0	-2.86	-0.18	II	O	32	Mutant
CGGA_868	3640	0	-14.75	3.29	II	O	37	Mutant
CGGA_523	4063	0	-5.13	1.01	II	OA	61	Mutant
CGGA_541	4047	0	-2.73	-0.25	II	A	36	Wildtype
CGGA_662*	284	1	2.79	-2.41	II	O	59	Mutant
CGGA_1059*	21	1	2.68	1.86	II	rA	54	Mutant
CGGA_474	2029	0	-6.70	2.60	III	AO	NA	Wildtype

*: Outliers that "died too early" relative to prognosis index that estimated by Rwt MTPL-EN. **: $PI = \boldsymbol{x}_i'\boldsymbol{\beta}^{MTPL-EN}$.

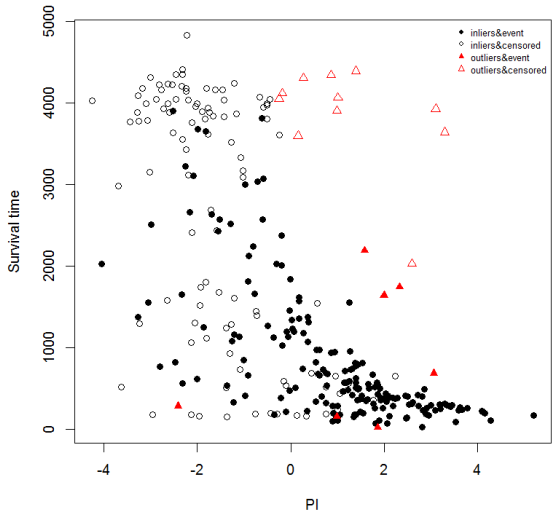

Figure 7. Scatter plot of the prognostic index PI estimated by Rwt MTPL-EN and survival time. (Note: black solid dots: normal points with outcomes; black hollow dots: censored normal points; red solid triangles: outliers with outcomes; red hollows: censored outliers, $PI = \boldsymbol{x}_i'\boldsymbol{\beta}^{MTPL-EN}$)

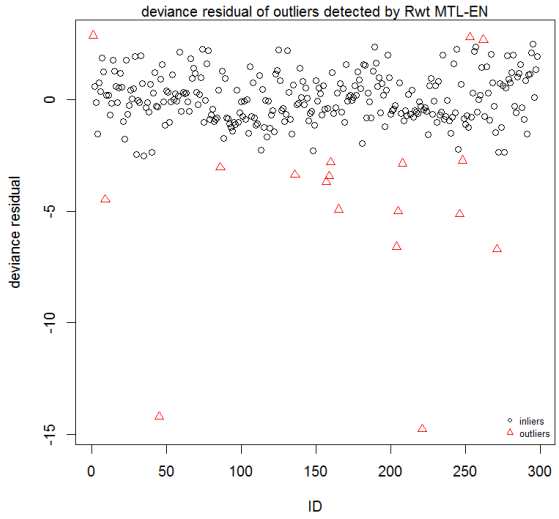

Figure 8. Deviance residuals corresponding to the model estimated by Rwt MTPL-EN. (Note: black hollow dots: normal points; red hollow triangles: outliers)

Among 3 outliers that "failed too early" identified by Rwt MTPL-EN, the survival time of CGGA was 21 days and their outcomes were death, but the prognosis index were very low, only −2.41. According to the covariate value of the individuals, they should have longer survival times. This showed that there were different correlation patterns between the survival time and covariates of these individuals. The other two outliers that "failed too early", CGGA_1059 and CGGA_11, had shorter survival time, which were 21 days and 155 days respectively. However, their prognostic indexes were higher, 0.98 and 1.86 respectively, indicating that they were not obvious outliers.

Except for CGGA_640, EN did not identify other outliers that "lived too long". As shown in the results of simulation experiments, outliers that "live too long" were hardly detected by EN but

accurately detected by Rwt MTPL-EN. In addition, the outliers that "lived too long" had a greater impact on the accuracy of variable selected by EN than outliers that "failed too early". Outliers that "lived too long" were easily identified by Rwt MTPL-EN, so that their influence on performance of EN could be removed.

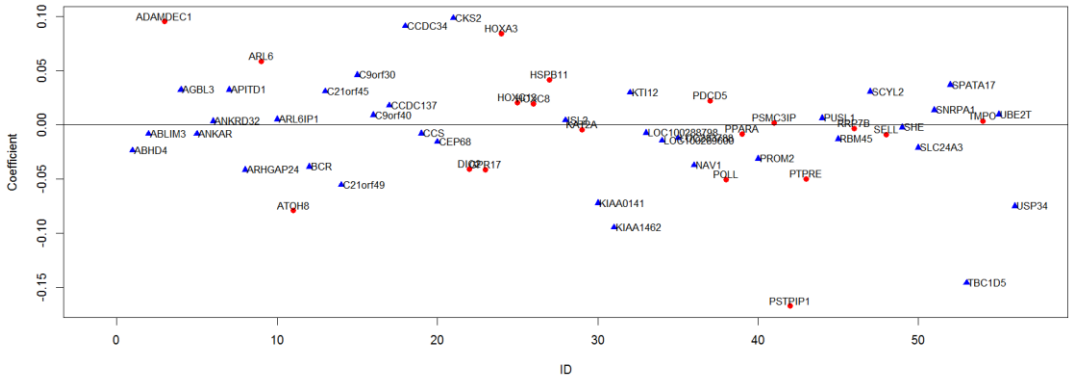

Figure 9. Fifty six Genes identified by Rwt MTPL-EN and their coefficients. (Red dots: genes related to gliomas have been reported in the literature)

As can be seen from Figure 9 and Table S5 in supplementary file, Rwt MTPL-EN identified 56 genes, 19 of which have been reported in the literature to be related to the prognosis or occurrence of glioma. For example, HSPB11 (Cheng, W., M. Li, Y. Jiang, C. Zhang, J. Cai, K. Wang and A. Wu [21]), DIO2 (Bunevicius, A., E. R. Laws, A. Saudargiene, A. Tamasauskas, G. Iervasi, V. Deltuva, T. R. Smith and R. Bunevicius [22]), PTPRE (Carvalho, D., A. Mackay, L. Bjerke, R. G. Grundy, C. Lopes, R. M. Reis and C. Jones [23]), HOXC8 (Sibin, M., S. Harshitha, K. Narasingarao, I. B. Dhananjaya, P. S. Dhaval and G. Chetan [24]), TMPO (Zhang, L., G. Wang, S. Chen, J. Ding, S. Ju, H. Cao and H. Tian [25]), ADAMDEC1 [26], ARL6 [27], ATOH8 [28], Gpr17 [29], HOXA3 [30], HOXC13 [31], KAT2A [32], PDCD5 [33], POLL [34], PPARA [35], PSMC3IP [36], PSTPIP1 [37], RRP7B [38], Sell [39].

EN identified 87 genes, of which 22 were reported to be related to gliomas, see Table S4. Among them, 28 genes overlapped with those by Rwt MTPL-COX, which was shown in Table S5. Of these 28 genes, 9 genes have been reported in the literature to be related to gliomas. However, there are still 10 reported genes (HSPB11, PTPRE, PSMC3IP, GT198, PPARA, PDCD5, TMPO, ARL6, ATOH8, RRP7B) that were not screened by EN. Among them, we divided patients into two groups according to whether the expression value of ARL6 was greater than 0, and compared the survival curves of the two groups. In all samples, the p value of the survival curve comparison between the two groups is 3×10^{-5}. However, the p value of the survival curves of the two groups in the subset without outliers is 3×10^{-7} as shown in Figure S4 in the supplementary file. This suggests that the presence of outliers leads to the underestimation of the effect of ALR6 on the prognosis of gliomas.

Although other genes identified by Rwt MTPL-EN were not reported to be associated with gliomas, there are reports in the literature related to the occurrence or prognosis of brain diseases or other tumors. For example, TBC1D5 [40], APITD1(Han, S. J., K. Begum, C. E. Foulds, R. A.

Hamilton, S. Bailey, A. Malovannaya, D. Chan, J. Qin and B. W. O'Malley [41]), CCDC34 [42], CKS2 [43], KIAA0141 [44] and USP34 [45]. As potential genes related to the prognosis of glioma, they provide reference information for the next experimental verification.

Through the analysis of the glioma gene expression data, only 28 genes selected by Rwt MTPL-EN and EN were coincident, indicating that there were samples in this data that have a greater impact on the estimation of EN. The dependence structure of these patients' survival time and covariates is different from that of other patients, that is to say, they "died too early" or "lived too long" relative to the model's estimated risk of death. After removing the outliers, the prediction accuracy of Rwt MTPL-EN was higher than that of EN. In terms of identified outliers, most of outliers identified by EN were those "failed too early", but most of them were not obvious outliers according to their *PI* and clinical variables. And only one outlier that "lived too long" was identified, indicating that the sensitivity in identifying outliers that "lived too long" was low. Most of outliers that identified by Rwt MTPL-EN were those "lived too long", and most of them were obvious outliers from their clinical characteristics and prognostic index. And through simulation experiments, it can be known that outliers that "lived too long" had a greater impact on the accuracy of variable selection of EN. Rwt MTPL-EN had advantages in identifying outliers that "lived too long". So, their influence on the estimation of EN can be removed by Rwt MTPL-EN.

4. Discussion

It is a great challenge to find biomarkers related to prognosis from high-dimensional genomic data, and to be able to resist the influence of noise and heterogeneity of samples in the experimental process, to obtain robust estimation. In this article, a robust penalized Cox model based on maximum trimmed partial likelihood estimation was established, and an AR-Cstep algorithm combining Metropolis-type acceptance-rejection algorithm and C-step algorithm was proposed to solve the estimation of MPTL-EN. By simulating high-dimensional datasets with outliers, the robust MPTL-EN performed better than non-robust EN-type penalized Cox regression in variable selection, outlier detection, and prediction. Moreover, Rwt MTPL-EN is better than Raw MTPL-EN. When outliers in response deviated farther, the number of vairables selected by EN became less. When the outliers in predictors also occurs, the number of vairables selected by EN was far greater than the number of real non-zero variables. Both situations made the accuracy of variables selection of EN decrease. However, Rwt MTPL-EN remains stable under various conditions, which indicates that the Rwt MTPL-EN can resist outliers in the reponse and predictors. According to the analysis of glioma gene expression data, the variables selected by Rwt MTPL-EN were different from those of EN, and a higher proportion of genes related to glioma had been identified by Rwt MTPL-EN. After removing outliers, prediction accuracy of Rwt MTPL-EN was higher than that of EN, and more outliers that "lived too long" relative to the prognosis index were identified.

The robust penalized Cox model based on trimming directly selected variables for high-dimensional dataset. Compared with robust estimation after reduced dimensions from high dimensions, it avoided the influence of outliers on the accuracy of dimensionality reduction. Compared with the residual analysis, it avoided the influence of "masking" and "swamping" on the estimation. Compared with other robust methods such as Huber's loss function and Tukey's loss function, it could resist the outliers in the response and the predictors.

The AR-Cstep algorithm, which combined the acceptance-rejection and C-step algorithm, can

solve the problem that the C-step algorithm does not converge because the penalty parameter changes during the iteration. In order to achieve the maximum of the trimmed likelihood function, and to avoid falling into local optimum, the metropolis-type probabilistic acceptance rejection algorithm was combined. This improvement can make the AR-Cstep algorithm generalized to other robust penalized regression models, such as robust Adaptive LASSO, Group LASSO, SCAD, MCP and so on. The improved AR-Cstep algorithm based on residuals no longer relies on separating individual contributions from the model's likelihood function, but instead used residuals to measure the individual contributions. This idea can also be generalized to solve the robust problem of similar models. Such as robust Cox regression in low-dimensional situations, conditional logistic regression, and so on. In the likelihood function of these models, it is difficult to separate the individual's contribution to the objective function, so AR-Cstep algorithm based on residuals can be used.

In this paper, it is found that the outliers that “lived too long” and “failed too early” had different effects on EN, and outliers that “lived too long” had a greater impact on the estimation of EN. This is also the case in Cox regression. Valsecchi M et al. [8] explained why long-term survivors have a greater impact on Cox regression, which is also applicable to penalized Cox regression. First, long-term survivors are part of many risk sets (all individuals who fail before them). Secondly, the risk set of early failure individuals is usually very large, but the individuals who fail at the end of the study correspond to very small risk sets. The risk sets of the two groups with different exposure states were similar at the beginning. But over time, as the individual failure rate of the high-risk group is higher than that of the low-risk group, the comparison of the risk set size between the two groups will change accordingly. In the end, the risk sets of the two groups are highly unbalanced, and the risk set of the high-risk group may be only one or two individuals. So, removing or adding such an individual will have a great impact on the estimation of the hazard ratio.

Subsequent analysis methods of identified outliers need to be explored in combination with the application. Peng S et al. [10] compared the integrated genomics of long-term survival and short-term survival glioma patients to discover the molecular markers with different prognoses after standard treatment. Therefore, individualized treatment can avoid treatment failure caused by wrong treatment. According to Burrell, R. A., N. McGranahan, J. Bartek and C. Swanton [6], phenotypic heterogeneity is not determined solely through genetic distinctions between subclones, but also through stochastic events in gene expression and protein stability, epigenetic divergence and micro-environmental fluctuations. There is a crucial need to understand mechanisms driving genomic instability so that therapeutic approaches to limit cancer diversity, adaptation and drug resistance can be developed.

In practice, when the penalized Cox model is used to screen prognostic biomarkers, it is often impossible to know whether there are outliers in the data. So, both robust and non-robust models can be used to fit the data. If the results of the variable selection of the two models are similar, it means that the non-robust model is not affected by outliers, and there are no outliers in the data. In this case, a non-robust model can be used because the efficiency of the non-robust model is higher. However, If the results of the two models are quite different, it means that there are some individuals who are not suitable for the model and the penalized Cox regression model estimation is incorrect. At this time, the robust Cox model is more suitable for this data.

In this article, we assumed that the pure data without outliers satisfied the proportional risk assumption of Cox model. Time dependent covariates often needs to be specified according to practical experience, which requires a better understanding of the impact of associated biomarkers on

prognosis. If we are at the stage of extensive screening of related biomarkers from high-dimensional data, and the actual problems are not well understood, the proportional risk model can be as a preliminary choice. Whether the variable is related to the prognosis is estimated firstly, and then whether the influence of the variable changes with time is determined by a more sophisticated model.

5. Conclusions

The robust penalized Cox model based on trimming Rwt MTPL-EN established in this paper can select variables more accurately than the non-robust EN model when outliers exist. It can resist outliers both in response and predictors. Rwt MTPL-EN can identify outliers more accurately, especially in identifying outliers that "lived too long", while outliers that "lived too long" had a greater impact on the accuracy of variables selection in EN. The AR-Cstep algorithm established in this article solves the problem that the C-step algorithm does not converge due to the change of the penalty parameters of penalized regression. It no longer depends on separating the individual contributions from the likelihood function of the model. This improvement allows the AR-Cstep algorithm to be generalized to more models.

Acknowledgments

This research work was funded by National Natural Science Foundation for Young Scholars of China (Grant No.81502891) and the National Natural Science Foundation of China (Grant No. 81872715). The funders played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Z. Liu, M. Li, Q. Hua, Y. Li, G. Wang, Identification of an eight-lncrna prognostic model for breast cancer using wgcna network analysis and a cox-proportional hazards model based on l1-penalized estimation, *Int. J. Mol. Med.,* **44** (2019), 1333–1343. https://doi.org/10.3892/ijmm.2019.4303
2. X. Y. Shen, X. P. Liu, C. K. Song, Y. J. Wang, S. Li, W. D. Hu, Genome-wide analysis reveals alcohol dehydrogenase 1c and secreted phosphoprotein 1 for prognostic biomarkers in lung adenocarcinoma, *J. Cellular Physiol.,* **234** (2019), 22311–22320. https://doi.org/10.1002/jcp.28797
3. L. Wang, J. Shi, Y. Huang, S. Liu, J. Zhang, H. Ding, et al., A six-gene prognostic model predicts overall survival in bladder cancer patients, *Cancer Cell Int.,* **19** (2019), 229. https://doi.org/10.1186/s12935-019-0950-7
4. J. Choi, S. Park, Y. Yoon, J. Ahn, Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers, *Bioinformatics,* **33** (2017), 3619–3626. https://doi.org/10.1093/bioinformatics/btx487

5. K. Polyak, Heterogeneity in breast cancer, *J. Clin. Invest.,* **121** (2011), 3786–3788. https://doi.org/10.1172/JCI60534
6. R. A. Burrell, N. McGranahan, J. Bartek, C. Swanton, The causes and consequences of genetic heterogeneity in cancer evolution, *Nature,* **501** (2013), 338–345. https://doi.org/10.1038/nature12625
7. A. Nardi, M. Schemper, New residuals for cox regression and their application to outlier screening, *Biometrics,* **55** (1999), 523–529. https://doi.org/10.1111/j.0006-341X.1999.00523.x
8. M. Valsecchi, D. Silvestri, P. Sasieni, Evaluation of long-term survival: Use of diagnostics and robust estimators with cox's proportional hazards model, *Stat. Med.,* **15** (1996), 2763–2780. https://doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2763::AID-SIM319>3.0.CO;2-O
9. E. Carrasquinha, A. Veríssimo, M. B. Lopes, S. Vinga, Identification of influential observations in high-dimensional cancer survival data through the rank product test, *BioData mining,* **11** (2018), 1. https://doi.org/10.1186/s13040-018-0162-z
10. S. Peng, H. Dhruv, B. Armstrong, B. Salhia, C. Legendre, J. Kiefer, et al., Integrated genomic analysis of survival outliers in glioblastoma, *Neuro-oncol.,* **19** (2017), 833–844. https://doi.org/10.1093/neuonc/nox036.104
11. P. J. Rousseeuw, Least median of squares regression, *J. Am. Stat. Assoc.,* **79** (1984), 871–880. https://doi.org/10.1080/01621459.1984.10477105
12. A. Farcomeni, S. Viviani, Robust estimation for the cox regression model based on trimming, *Biometr. J.,* **53** (2011), 956–973. https://doi.org/10.1002/bimj.201100008
13. P. J. Rousseeuw, K. Van Driessen, Computing lts regression for large data sets, *Data mining and knowledge discovery,* **12** (2006), 29–45. https://doi.org/10.1007/s10618-005-0024-4
14. B. Chakraborty, P. Chaudhuri, On an optimization problem in robust statistics, *J. Comput. Graph. Stat.,* **17** (2008), 683–702. https://doi.org/10.1198/106186008X340751
15. T. M. Therneau, P. M. Grambsch, T. R. Fleming, Martingale-based residuals for survival models, *Biometrika,* **77** (1990), 147–160. https://doi.org/10.1093/biomet/77.1.147
16. J. Klein, M. Moeschberger, Survival analysis: Techniques for censored and truncated data springer, *New York* (1997),
17. R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate cox proportional hazards models, *Stat. Med.,* **24** (2005), 1713–1723. https://doi.org/10.1002/sim.2059
18. L. D. Maxim, R. Niebo, M. J. Utell, Screening tests: A review with examples, *Inhal Toxicol.,* **26** (2014), 811–828. https://doi.org/10.3109/08958378.2014.955932
19. N. Ternes, F. Rotolo, S. Michiels, Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional cox regression models, *Stat. Med.,* **35** (2016), 2561–2573. https://doi.org/10.1002/sim.6927
20. H. Uno, T. Cai, M. J. Pencina, On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data, *Stat. Med.,* 30 (2011), 1105–1117. https://doi.org/10.1002/sim.4154
21. W. Cheng, M. Li, Y. Jiang, C. Zhang, J. Cai, K. Wang, et al., Association between small heat shock protein b11 and the prognostic value of mgmt promoter methylation in patients with high-grade glioma, *J. Neurosurg.,* (2015), 1–10. https://doi.org/10.3171/2015.5.JNS142437
22. A. Bunevicius, E. R. Laws, A. Saudargiene, A. Tamasauskas, G. Iervasi, V. Deltuva, et al., Common genetic variations of deiodinase genes and prognosis of brain tumor patients, *Endocrine,* **66** (2019), 563–572. https://doi.org/10.1007/s12020-019-02016-6

23. D. Carvalho, A. Mackay, L. Bjerke, R. G. Grundy, C. Lopes, R. M. Reis, et al., The prognostic role of intragenic copy number breakpoints and identification of novel fusion genes in paediatric high grade glioma, *Acta Neuropathol. Commun.,* **2** (2014), 23. https://doi.org/10.1186/2051-5960-2-23
24. M. Sibin, S. Harshitha, K. Narasingarao, I. B. Dhananjaya, P. S. Dhaval, G. Chetan, Effect of rs11614913 polymorphism on mature mir196a2 expression and its target gene hoxc8 expression in human glioma, *J. Mol. Neurosci.,* **61** (2017), 144–151. https://doi.org/10.1007/s12031-016-0855-z
25. L. Zhang, G. Wang, S. Chen, J. Ding, S. Ju, H. Cao, et al., Depletion of thymopoietin inhibits proliferation and induces cell cycle arrest/apoptosis in glioblastoma cells, *World J. Surg. Oncol.,* **14** (2016), 267. https://doi.org/10.1186/s12957-016-1018-y
26. A. Jimenez-Pascual, J. D. Lathia, F. A. Siebzehnrubl, Adamdec1 and fgf2/fgfr1 signaling constitute a positive feedback loop to maintain gbm cancer stem cells, *Mol. Cell. Oncol.,* **7** (2020), 1684787. https://doi.org/10.1080/23723556.2019.1684787
27. S. H. Miao, H. B. Sun, Y. Ye, J. J. Yang, Y. W. Shi, M. Lu, et al., Astrocytic jwa expression is essential to dopaminergic neuron survival in the pathogenesis of parkinson's disease, *CNS Neurosci. Ther.,* **20** (2014), 754-762. https://doi.org/10.1111/cns.12249
28. F. Ducray, A. Idbaih, A. de Reyniès, I. Bièche, J. Thillet, K. Mokhtari, et al., Anaplastic oligodendrogliomas with 1p19q codeletion have a proneural gene expression profile, *Mol. Cancer,* **7** (2008), 41. https://doi.org/10.1186/1476-4598-7-41
29. J. D. Dougherty, E. I. Fomchenko, A. A. Akuffo, E. Schmidt, K. Y. Helmy, E. Bazzoli, et al., Candidate pathways for promoting differentiation or quiescence of oligodendrocyte progenitor-like cells in glioma, *Cancer Res.,* **72** (2012), 4856–4868. https://doi.org/10.1158/0008-5472.CAN-11-2632
30. A. Di Vinci, I. Casciano, E. Marasco, B. Banelli, G. L. Ravetti, L. Borzì, et al., Quantitative methylation analysis of hoxa3, 7, 9, and 10 genes in glioma: Association with tumor who grade and clinical outcome, *J. Cancer Res. Clin. Oncol.,* **138** (2012), 35–47. https://doi.org/10.1007/s00432-011-1070-5
31. N. Liu, Z. Wang, D. Liu, P. Xie, Hoxc13-as-mir-122-5p-satb1-c-myc feedback loop promotes migration, invasion and emt process in glioma, *Onco Targets Ther.,* **12** (2019), 7165–7173. https://doi.org/10.2147/OTT.S220027
32. K. Liu, Q. Zhang, H. Lan, L. Wang, P. Mou, W. Shao, et al., Gcn5 potentiates glioma proliferation and invasion via stat3 and akt signaling pathways, *Int. J. Mol. Sci.,* **16** (2015), 21897–21910. https://doi.org/10.3390/ijms160921897
33. C. Wang, J. K. Li, H. Z. Li, H. D. Gong, The importance of expressing pdcd4 and pdcd5 anti-oncogenes in glioma, *J. Biol. Regul. Homeost. Agents,* **32** (2018), 731–736.
34. H. Wang, W. Wu, H. W. Wang, S. Wang, Y. Chen, X. Zhang, et al., Analysis of specialized DNA polymerases expression in human gliomas: Association with prognostic significance, *Neuro-oncology,* **12** (2010), 679–686. https://doi.org/10.1093/neuonc/nop074
35. R. Luo, L.-Y. Su, G. Li, J. Yang, Q. Liu, L.-X. Yang, et al., Activation of ppara-mediated autophagy reduces alzheimer disease-like pathology and cognitive decline in a murine model, *Autophagy,* (2019), 1–18. https://doi.org/10.1080/15548627.2019.1596488
36. L. Zhang, Y. Wang, M. H. Rashid, M. Liu, K. Angara, N. F. Mivechi, et al., Malignant pericytes expressing gt198 give rise to tumor cells through angiogenesis, *Oncotarget,* **8** (2017), 51591–

51607. https://doi.org/10.18632/oncotarget.18196
37. G. Li, Z. Wang, C. Zhang, X. Liu, F. Yang, L. Sun, et al., Megf10, a glioma survival-associated molecular signature, predicts idh mutation status, *Dis. Markers,* **2018** (2018), 5975216. https://doi.org/10.1155/2018/5975216
38. P. Yang, W. Yan, W. Zhang, G. You, Z. Bao and T. Jiang, Whole-genome messenger rna profiling reveals genes involved in malignant progression of glioma, *Zhonghua yi xue za zhi,* **93** (2013), 5–7.
39. V. Haage, M. Semtner, R. O. Vidal, D. P. Hernandez, W. W. Pong, Z. Chen, et al., Comprehensive gene expression meta-analysis identifies signature genes that distinguish microglia from peripheral monocytes/macrophages in health and glioma, *Acta Neuropathol. Com.,* **7** (2019), 20. https://doi.org/10.1186/s40478-019-0665-y
40. M. N. J. Seaman, A. S. Mukadam, S. Y. Breusegem, Inhibition of tbc1d5 activates rab7a and can enhance the function of the retromer cargo-selective complex, *J. Cell. Sci.,* **131** (2018), jcs217398. https://doi.org/10.1242/jcs.217398
41. S. J. Han, K. Begum, C. E. Foulds, R. A. Hamilton, S. Bailey, A. Malovannaya, et al., The dual Receptor α inhibitory effects of the tissue-selective estrogen complex for endometrial and breast safety, *Mol. Pharmacol.,* **89** (2015), 14–26. 10.1124/mol.115.100925.
42. L. B. Liu, J. Huang, J. P. Zhong, G. L. Ye, L. Xue, M. H. Zhou, et al., High expression of ccdc34 is associated with poor survival in cervical cancer patients, *Med. Sci. Monit.,* **24** (2018), 8383–8390. https://doi.org/10.12659/MSM.913346
43. N. Huang, Z. Wu, H. Hong, X. Wang, F. Yang, H. Li, Overexpression of cks2 is associated with a poor prognosis and promotes cell proliferation and invasion in breast cancer, *Mol. Med. Rep.,* **19** (2019), 4761–4769. https://doi.org/10.3892/mmr.2019.10134
44. T. Harada, A. Iwai, T. Miyazaki, Identification of dele, a novel dap3-binding protein which is crucial for death receptor-mediated apoptosis induction, *Apoptosis,* **15** (2010), 1247–1255. https://doi.org/10.1007/s10495-010-0519-3
45. C. Li, L. Huang, H. Lu, W. Wang, G. Chen, Y. Gu, et al., Expression and clinical significance of ubiquitin-specific-processing protease 34 in diffuse large b-cell lymphoma, *Mol. Med. Rep.,* **18** (2018), 4543–4554. https://doi.org/10.3892/mmr.2018.9447

AIMS Press