Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 50 (5) (2021), 1572 – 1582
DOI : 10.15672/hujms.810383

Research Article

Robust variable selection in the logistic regression model

Yunlu Jiang[1], Jiantao Zhang[1], Yingqiang Huang[2], Hang Zou[1], Meilan Huang*[1], Fanhong Chen[3]

[1] *Department of Statistics, College of Economics, Jinan University, Guangzhou, 510632, China*
[2] *Huashang College, Guangzhou, 511300, China*
[3] *Institute of Intellectural Property, Jinan University, Guangzhou, 510632, China*

Abstract

In this paper, we proposed an adaptive robust variable selection procedure for the logistic regression model. The proposed method is robust to outliers and considers the goodness-of-fit of the regression model. Furthermore, we apply an MM algorithm to solve the proposed optimization problem. Monte Carlo studies are evaluated the finite-sample performance of the proposed method. The results show that when there are outliers in the dataset or the distribution of covariate variable deviates from the normal distribution, the finite-sample performance of the proposed method is better than that of other existing methods. Finally, the proposed methodology is applied to the data analysis of Parkinson's disease.

Mathematics Subject Classification (2020). 62G35, 62H12

Keywords. Logistic regression, variable selection, robustness, MM algorithm

1. Introduction

Since the logistic regression model can be used to classify samples, it has been widely applied in biomedicine and other fields, and has almost become the most commonly used analysis tool in epidemiology and medicine [17, 23, 30]. For a logistic regression model, the most commonly used method is maximum likelihood estimation (MLE). However, the MLE method is very sensitive to outliers [1,6,7,12,15,25,31], therefore, it will be seriously affected, and lead to a large deviation in the prediction of classification probability. In practical applications, many covariates are introduced in the initial stage of modeling. However, models that include all the covariates are difficult to interpret since irrelevant variables may increase variance. Therefore, the selection of significant covariates is one of the most important statistical problems.

For the logistic regression model, Vinterbo and Ohno-Machado [29] proposed a variable selection method via genetic algorithm. Zellner et al. [33] proposed a bootstrap method to select variables from a complete logistic regression model. This method simplifies the

*Corresponding Author.
Email addresses: tjiangyl@jnu.edu.cn (Y. Jiang), cheungjt@163.com (J. Zhang), acc2009@163.com (Y. Huang), zouhang19980304@163.com (H. Zou), 1726664935@qq.com (M. Huang), chenzhongxun@foxmail.com (F. Chen)
Received: 14.10.2020; Accepted: 03.07.2021

regression by using resampling process, and shows better performance when there is correlation between the predictor variables. Bursac et al. [3] proposed a purposeful and automatic logistic variable selection method, which started from a single variable. During the iteration of variable selection, the variables that have no significant influence are deleted from the model. Meier et al. [22] extended the group lasso method of linear model to logistic model based on the theory of block co-ordinate gradient descent minimization, and proposed a variable selection algorithm which is particularly suitable for high-dimensional case. Zhang et al. [34] proposed a new variable selection method via the variational Bayesian model. This method can make the corresponding logistic model adaptively determine the optimal value of the super parameters, thus can achieve effective sparsity. However, it is important to note that the above proposed methods are very sensitive to outliers in the dataset.

In many practical applications, the dataset contains a nonnormally distributed response variable and/or many covariates that potentially exist multiple high leverage points. This often causes some serious problems for the classical variable selection when the dataset exist outliers. In fact, there also exist many robust variable selection methods for the logistic regression model in the literature. For example, Stefanski et al. [28] put forward to use bounded influence function and leverage process to achieve the effect of robust variable selection in the presence of outliers, so as to fit out a reasonable and effective logistic regression model. By studying the prior distribution of logistic regression model, Chen et al. [4] proposed a new variable selection method which was very robust under various prior parameters, and this method can show higher superior performance in the logistic regression model with moderate dimensions. Kinney and Dunson [18] proposed a new variable selection method of fixed and random effect in binary response model. Compared with the early fixed models, it has higher effectiveness and stronger robustness. Guns and Vanacker [11] proposed a logistic regression method based on rare events with repetition, and overcame some limitations of traditional methods through robust variable selection. Li and Liu [19] proposed a robust variable selection via an adaptive forward backward SODA method, which did not need the joint normal hypothesis of predictors.

The above proposed robust variable selection methods only consider the treatment of outliers, and ignore the goodness-of-fit of the regression model. In this paper, we propose an adaptive variable selection procedure that is robust and considers the goodness-of-fit of the regression model. In addition, the minorization-maximization (MM) algorithm [14] is used to solve the proposed optimization problem. The Monte Carlo studies results illustrate that the finite sample performance of proposed method is better than that of some existing methods when there are outliers in the dataset.

The rest of this paper is organized as follows. In Section 2, we review some classical methods, and introduce the proposed robust variable selection procedure and its corresponding algorithm. In Section 3, numerical simulations are conducted to evaluate the finite-sample performance of the proposed method. In Section 4, the proposed method is compared with the existing methods through the analysis of a real dataset. A discussion is given in Section 5.

2. Methodology and main results

2.1. Review some classical methods

Assume that $\{(y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$ is a random sample from some population and satisfies a following logistic regression model

$$\begin{aligned} P(Y_i = 1|\mathbf{x}_i) &= \pi(\mathbf{x}_i), \quad P(Y_i = 0|\mathbf{x}_i) = 1 - \pi(\mathbf{x}_i), \\ \pi(\mathbf{x}_i) &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}, \end{aligned}$$

where y_i is a random response variable with either 0 or 1, and $\mathbf{x}_i$ is a p-dimensional predictor variable, $\boldsymbol{\beta}$ is a p-dimensional unknown parameter vector, Y_i is a random variable and follows a following Bernoulli distribution

$$f(y_i|\mathbf{x}_i,\boldsymbol{\beta}) = [\pi(\mathbf{x}_i)]^{y_i}[1-\pi(\mathbf{x}_i)]^{1-y_i},\ y_i = 0,1 \tag{2.1}$$

According to Equation (2.1), the log-likelihood function for $\boldsymbol{\beta}$ is given as follows:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n}[y_i \log(\pi(\mathbf{x}_i)) + (1-y_i)\log(1-\pi(\mathbf{x}_i))] \\ &= \sum_{i=1}^{n}[y_i\mathbf{x}_i^T\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))]. \end{aligned}$$

In order to simultaneously achieve variable selection and parameter estimation, $\boldsymbol{\beta}$ can be estimated by maximizing the following penalized log-likelihood function $l_p(\boldsymbol{\beta})$:

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i\mathbf{x}_i^T\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))] - n\sum_{j=1}^{p}p_{\lambda_j}(|\beta_j|), \tag{2.2}$$

where $p_{\lambda_j}(\cdot)$ is a penalty function and λ_j is a penalized parameter for the j-th parameter component. We can find from Equation (2.2) that $l_p(\boldsymbol{\beta})$ is based on the maximum likelihood method, which is very sensitive to outliers. To obtain robust variable selection, Park and Konishi [24] proposed a weighted penalized log-likelihood function

$$l_p^R(\boldsymbol{\beta}) = \sum_{i=1}^{n}R_i^p[y_i\mathbf{x}_i^T\boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T\boldsymbol{\beta}))] - \sum_{j=1}^{p}\lambda\left(\frac{1}{2}\alpha\beta_j^2 + (1-\alpha)\,|\beta_j|\right), \tag{2.3}$$

where $0 \le \alpha \le 1$, $\lambda > 0$, and

$$R_i^p = \frac{\min\{\sqrt{k/\mathrm{R.MD}_i^{pc}}, 1\}}{\sum_{i=1}^{n}\min\{\sqrt{k/\mathrm{R.MD}_i^{pc}}, 1\}}.$$

Here, k is the 95% quantile of χ^2 distribution with degrees of freedom p, and

$$\mathrm{R.MD}_i^{pc} = \sqrt{(\mathbf{x}_i - \mathbf{T}^{pc})^T(\mathbf{C}^{pc})^{-1}(\mathbf{x}_i - \mathbf{T}^{pc})}$$

is a robust Mahalanobis distance, where mean $\mathbf{T}^{pc}$ and covariance matrix $\mathbf{C}^{pc}$ are given by minimum volume ellipsoid(MVE) [27].

2.2. Adaptive robust variable selection procedure

Davies [8] pointed out that the MVE has a slow $n^{-1/3}$ rate of convergence and it is very difficult to compute in high dimensions. Meanwhile, the above robust variable selection procedure is assigned a weight in advance, which can correctly reflect the outlying information among all covariates. However, according to [32], the robustness of the method based on the pre-assigned weights will be also deteriorated significantly due to a high percentage of outliers. In addition, this method does not consider the goodness-of-fit of the regression model.

Next, we propose a novel adaptive robust variable selection procedure, which is robust to outliers and considers the goodness-of-fit of the regression model. Let $\tilde{\boldsymbol{\beta}}_n$ be an initial robust estimator for $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\mu}}$, and $\tilde{\Sigma}$ be an initial robust location and scatter estimators of covariates $(\mathbf{x}_1^T,\ldots,\mathbf{x}_n^T)^T$ based on the minimum covariance determinant (MCD) estimators [26], respectively. Then, the square of robust Mahalanobis distance for $\mathbf{x}_i$ is defined as

$$m_i^2 = (\mathbf{x}_i - \tilde{\boldsymbol{\mu}})^T\left(\tilde{\Sigma}\right)^{-1}(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}),\ i = 1,\ldots,n.$$

Let F_n be the empirical distribution function for $\{m_1^2, \ldots, m_n^2\}$. When the covariate $\mathbf{X}$ follows a multivariate normal distribtion, F_n should converge to the chi-square distribution with degrees of freedom p. Let $m_{(1)}^2 \leqslant \ldots \leqslant m_{(n)}^2$ denote the order statistics for $\{m_1^2, \ldots, m_n^2\}$. According to [9], the statistic of the goodness-of-fit of the data can be defined as

$$\tilde{\eta} = \sup_{s \geqslant F_{\chi_p^2}^{-1}(1-\delta)} \left\{ F_{\chi_p^2}(s) - F_n(s) \right\}_+ = \max_{i \geqslant i_0} \left\{ F_{\chi_p^2}(m_{(i)}^2) - \frac{i-1}{n} \right\}_+,$$

where $\{\cdot\}_+$ denotes the positive function, $i_0 = \min\left\{ i : m_{(i)}^2 \geqslant F_{\chi_p^2}^{-1}(1-\delta) \right\}$, and δ is a constant that determines the length of the tails [9, 10]. According to [1, 5, 6], we set $\delta = 0.025$. Advocated by [16], the adaptive weighted function for observation data is given as follows:

$$w(\mathbf{x}_i, y_i) = \varphi\left(\tilde{\eta} \left\| \left(y_i - l^{(1)}\left(\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_n \right) \right) \mathbf{x}_i \right\| \right), \tag{2.4}$$

where $l^{(1)}(\cdot)$ is the first order differential of $l(\cdot)$, $\varphi(\cdot)$ is a non-increasing function [16]. According to [9], $\varphi(\cdot)$ has the following three forms: Hard-rejection: $\varphi(x) = I\{0 < x \leqslant 1\}$; Huber: $\varphi(x) = \min\{1, 1/x\}$; Gaussian: $\varphi(x) = \exp(0.5 - 0.5x)$. In this paper, we take Hard-rejection function. Based on above adaptive weighted function, we propose a following adaptive robust variable selection procedure

$$\hat{\boldsymbol{\beta}}_n = arg\max_{\boldsymbol{\beta}} [l_p^w(\boldsymbol{\beta})], \tag{2.5}$$

where

$$l_p^w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w(\mathbf{x}_i, y_i) \left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})) \right] - n \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|).$$

Remark 2.1. In Equation (2.4), $\tilde{\eta} > 0$ can indicate the quality of data. When $\tilde{\eta}$ is very close to 0, it illustrates that there are almost no outliers in the sample data, and the data quality is good. On the contrary, if $\tilde{\eta}$ is too large, it means that there are contaminated data, and the quality of the sample is poor. Since the MLE satisfies the estimating equation $\sum \left(y_i - l^{(1)}\left(\mathbf{x}_i^T \boldsymbol{\beta} \right) \right) \mathbf{x}_i = 0$, the leverage points can have any great influence on the MLE. Therefore, the MLE can result in the non-robustness of the regression model. The statistic $\tilde{\eta}$ can improve the robustness of regression estimator by decreasing the weights of leverage points and larger $\left\| \left(y_i - l^{(1)}\left(\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_n \right) \right) \mathbf{x}_i \right\|$ at the same time [16].

Furthermore, according to Equation (2.4), the non-increasing function $\varphi(\cdot)$ is used to measure the deviation of $(\mathbf{x}_i, y_i)$ in the model. It controls the influence of the potential outliers on the regression effect by giving a weight to $(\mathbf{x}_i, y_i)$. With the help of non-increasing weight function, for all $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, the larger the deviation from the normal value is, the smaller the corresponding adaptive weight is, so as to effectively reduce the potential impact of outliers on regression estimator and improve the robustness of the proposed method.

To implement Equation (2.5), we need an initial robust estimator $\tilde{\boldsymbol{\beta}}_n$ and penalty function $p_{\lambda_j}(\beta_j)$. In this paper, $\tilde{\boldsymbol{\beta}}_n$ is obtained from the estimation method proposed by [2]. The penalty function $p_{\lambda_j}(\beta_j)$ take an adaptive lasso, e.g., $p_{\lambda_j}(\beta_j) = \lambda \frac{|\beta_j|}{|\tilde{\beta}_{nj}|}$, where λ is a tuning parameter, and $\tilde{\beta}_{nj}$ is the j-th component of $\tilde{\boldsymbol{\beta}}_n$.

2.3. Algorithm

In this subsection, we apply an MM algorithm proposed by [14] to solve Equation (2.5). Hunter and Lange [14] pointed out that MM algorithm possessed a descent property,

which lended its remarkable numerical stability. Computer code for implementing MM algorithm is available from the authors upon request. First, according to [13], we construct a following surrogate function for the penalty function

$$Q_2(\beta_j|\hat{\beta}_{nj}^{(k)}) = \lambda \frac{\beta_j^2}{2|\tilde{\beta}_{nj}|(|\hat{\beta}_{nj}^{(k)}|)} + \frac{\lambda}{2|\tilde{\beta}_{nj}|}(|\hat{\beta}_{nj}^{(k)}|), \tag{2.6}$$

where $\hat{\beta}_{nj}^{(k)}$ is the k-th approximation of $\hat{\boldsymbol{\beta}}_n$. Denote

$$\begin{aligned} h(\boldsymbol{\beta}) &= \sum_{i=1}^{n} w\left(\mathbf{x}_i, y_i\right)\left[y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log(1+\exp(\mathbf{x}_i^T \boldsymbol{\beta}))\right], \\ \mathbf{X} &= \left(\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T\right)^T, \\ \mathbf{Y} &= \left(y_1, y_2, \ldots, y_n\right)^T, \\ \mathbf{W} &= \operatorname{diag}\left\{w\left(\mathbf{x}_1, y_1\right), \ldots, w\left(\mathbf{x}_n, y_n\right)\right\}, \\ \mathbf{M} &= -\frac{1}{4}\mathbf{X}^T \mathbf{W} \mathbf{X}. \end{aligned}$$

By using [14], we construct a following surrogate function for $h(\boldsymbol{\beta})$

$$Q_1(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)}) = h\left(\hat{\boldsymbol{\beta}}_n^{(k)}\right) + \left[\nabla h\left(\hat{\boldsymbol{\beta}}_n^{(k)}\right)\right]^T \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^{(k)}\right) + \tfrac{1}{2}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^{(k)}\right)^T \mathbf{M}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n^{(k)}\right), \tag{2.7}$$

where $\nabla h\left(\boldsymbol{\beta}\right)$ denotes the gradient function of $h\left(\boldsymbol{\beta}\right)$. Therefore, by Equation (2.6) and Equation (2.7), we obtain a surrogate function for $l_p^w(\boldsymbol{\beta})$ as follows:

$$Q\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)}\right) = Q_1(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)}) - n\sum_{j=1}^{p} Q_2(\beta_j|\hat{\beta}_{nj}^{(k)}) \tag{2.8}$$

It is easy to show that $Q(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)})$ minorizes $l_p^w(\boldsymbol{\beta})$ in the sense that

$$Q\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)}\right) \leq l_p^w(\boldsymbol{\beta}),\ \forall\, \boldsymbol{\beta} \in \mathbb{R}^p \quad \text{and} \quad Q\left(\hat{\boldsymbol{\beta}}_n^{(k)}|\hat{\boldsymbol{\beta}}_n^{(k)}\right) = l_p^w(\hat{\boldsymbol{\beta}}_n^{(k)}).$$

Given the k-th approximation $\hat{\boldsymbol{\beta}}_n^{(k)}$, the MM iteration updates

$$\hat{\boldsymbol{\beta}}_n^{(k+1)} = arg\max_{\boldsymbol{\beta}} Q\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_n^{(k)}\right).$$

Denote

$$\mathbf{P}^{(k)} = \operatorname{diag}\left\{\frac{n\lambda}{2|\tilde{\beta}_{n1}|(|\hat{\beta}_{n1}^{(k)}|)}, \ldots, \frac{n\lambda}{2|\tilde{\beta}_{np}|(|\hat{\beta}_{np}^{(k)}|)}\right\}.$$

Then, we have

$$\hat{\boldsymbol{\beta}}_n^{(k+1)} = \left(\mathbf{M} - 2\mathbf{P}^{(k)}\right)^{-1}\left[\mathbf{M}\hat{\boldsymbol{\beta}}_n^{(k)} - \nabla h\left(\hat{\boldsymbol{\beta}}_n^{(k)}\right)\right].$$

2.4. Tuning parameter selection

In order to put the above proposed algorithm into effect, an appropriate tuning parameter λ should be selected in the process of calculation. The selection of tuning parameter plays a significant part in the variable selection procedure. Normally, there are lots of methods to select tuning parameters, such as cross-validation (CV), generalized cross-validation (GCV), Akaike information criterion(AIC), and Bayesian information criterion

(BIC). Since BIC can obtain the consistent model selection, we apply this criterion to select the tuning parameter λ.

$$\mathrm{BIC}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i, y_i)\left[y_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n - \log(1+\exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_n))\right] + \frac{\log(n)}{n} df, \tag{2.9}$$

where df denotes the number of non-zero components for $\hat{\boldsymbol{\beta}}_n$. 150 tuning parameters λ are generated by using the default setting of R package *glmnet* and then we obtain the optimal tuning parameter by minimizing $\mathrm{BIC}(\lambda)$.

3. Simulation

In this section, we illustrate the performance of the proposed method via numerical simulations. We simulate 200 data sets from the logistic regression model with sample sizes of $n = 300, 500, 1000$. In this simulation, we select $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{20})^T$, where $\beta_1 = \beta_4 = 1.5$, $\beta_2 = 0.5$, $\beta_3 = \beta_5 = 1$, and $\beta_j = 0$, $j \in \{6, 7, \ldots, 20\}$. The uncontaminated covariates $\mathbf{X}$ follows a 20-dimensional standard normal distribution (denoted as **D0**), the response vector $\mathbf{Y}$ follows the Bernoulli distribution.

In order to study robustness of the proposed method, we consider the following two contaminated distributions, where α denotes the contaminated rate:

D1: Using Maronna's contamination distribution [21] as

$$\tilde{\mathbf{x}}_i = \boldsymbol{\eta}_i + \frac{k_{lev}}{\sqrt{\mathbf{a}^T \Sigma^{-1} \mathbf{a}}}\mathbf{a}, \ i = 1, \ldots, m,$$

where $m = \lfloor \alpha n \rfloor$, $\boldsymbol{\eta}_i \sim N_p\left(\mathbf{0}_{20}, 0.1^2 I_{20\times 20}\right)$, $\mathbf{a} = \tilde{\mathbf{a}} - \frac{1}{20}\tilde{\mathbf{a}}^T \mathbf{1}_{20}$, and $\tilde{a}_j$ in $\tilde{\mathbf{a}}$ follows $U(-1, 1)$, $j = 1, \ldots, 20$. The parameter k_{lev} denotes the control of the distance in the direction that has the most influence for the estimator. We set $k_{lev} = 2000$ in this simulation.

D2: Referring to the contamination distribution proposed by [6], we define

$$x_{ij} \sim \begin{cases} S \cdot N(5, 0.2), & j = 1, 2, 3, 4, 5 \\ S \cdot N(0, 0.2), & j \neq 1, 2, 3, 4, 5 \end{cases}, \ i = 1, \ldots, m,$$

where $m = \lfloor \alpha n \rfloor$, $P(S = 1) = P(S = -1) = 0.5$. We set $\alpha = 0.1$ and $\alpha = 0.2$ to compare our proposed method (ARVSP) with the method (WRVSP) proposed by [24] and the penalized maximum likelihood (PML) method.

To examine the finite sample performance of the proposed method, we calculate the following three measures:

$$\mathrm{NCZ} = \#\left\{j : \beta_j = 0 \wedge \hat{\beta}_{nj} = 0\right\}, \ \mathrm{NIZ} = \#\left\{j : \beta_j \neq 0 \wedge \hat{\beta}_{nj} = 0\right\}, \text{ and } \mathrm{MSE} = \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n\|_2^2,$$

where $\#\{A\}$ denotes the number of elements within A. NCZ (number of correct zeros) indicates the number of the zero component in the true value of the parameter to be still correctly estimated as zero. NIZ (number of incorrect zeros) indicates the number of the non-zero component in the true value of the parameter to be incorrectly estimated as zero. MSE denotes the mean square error. Clearly, MSE and NIZ should be as small as possible while NCZ should be as large as possible. The corresponding simulation results are shown in Table 1.

From Table 1, we have the following findings:

(1) Under **D0**, three methods have similar performance. Meanwhile, the finite sample performance of proposed method is the same as those of PML method.

(2) Under **D1**, and **D2**, our proposed ARVSP method has the lowest NIZ and NCZ is close to the true value 15. In addition, MSE of proposed method is smaller than that of WRVSP method and PWL method. Therefore, the finite-sample performance of the proposed method is better than that of other existing methods

when there are outliers in the dataset, although its performance is slightly worse as the contamination ratio increases.

(3) MSE of all three methods decreases as sample size n increase.

Table 1. Simulation results for ARVSP, WRVSP and PML.

Distribution	α	n	ARVSP			PML			WRVSP		
			NCZ	NIZ	MSE	NCZ	NIZ	MSE	NCZ	NIZ	MSE
D0	0	300	14.8005	0.3300	0.6138	14.8005	0.3300	0.6138	14.7900	0.3350	0.6156
		500	14.8650	0.1700	0.3336	14.8650	0.1700	0.3336	14.8695	0.1800	0.3367
		1000	14.9145	0.0100	0.1262	14.9145	0.0100	0.1262	14.9205	0.0150	0.1277
D1	0.1	300	14.7255	0.3800	0.6844	14.3100	0.8100	1.7613	14.4555	0.5750	1.3354
		500	14.8095	0.1650	0.3613	14.0400	0.4900	1.2652	14.2155	0.3850	0.9968
		1000	14.9100	0.0300	0.1568	13.5300	0.3050	1.0643	13.5195	0.1450	0.7822
	0.2	300	14.6805	0.5500	0.9537	14.2200	0.9650	1.9749	14.3205	1.0100	2.0541
		500	14.7705	0.2250	0.4210	14.1105	0.5650	1.4114	14.1345	0.5600	1.3900
		1000	14.8650	0.0400	0.2249	13.4250	0.4100	1.1438	13.4895	0.4150	1.1256
D2	0.1	300	14.7555	0.4400	0.9049	14.9850	2.2850	5.0819	14.9550	1.3100	3.5730
		500	14.8995	0.3800	0.5928	15.0000	1.6000	4.8031	14.9805	0.6400	3.0193
		1000	14.9505	0.1950	0.3320	15.0000	0.5450	4.3149	15.0000	0.4000	2.6675
	0.2	300	14.0445	0.8400	1.6127	15.0000	4.5100	6.5886	15.0000	4.4500	6.4830
		500	14.2200	0.4250	0.8371	15.0000	4.3950	6.5441	14.9955	4.0950	6.3489
		1000	14.6700	0.2700	0.4063	14.9955	3.1950	6.2721	14.9850	2.4200	5.8598

4. Real data analysis

In this section, the proposed method will be applied to analyze a Parkinsons dataset [20], which can be downloaded from

`https://archive.ics.uci.edu/ml/datasets/Parkinsons.`

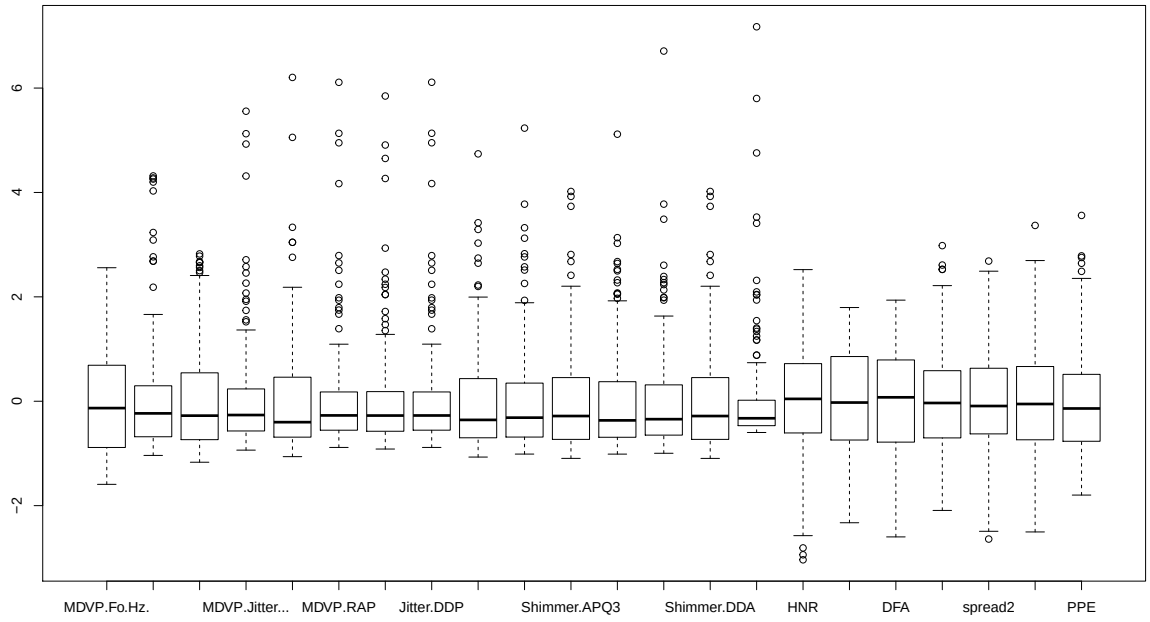

Figure 1. Boxplot of 22 attributes of a Parkinsons dataset.

This dataset consists of a series of biomedical voice measurements from 31 people, 23 of whom have Parkinson's disease (PD). The dataset contains 195 voice records and 23 attributes, one of which is the health status of the subjects, with "0" for health and "1" for Parkinson's disease patient. The remaining 22 attributes are numerical, including: Average vocal fundamental frequency (MDVP:Fo(Hz)), maximum vocal fundamental frequency (MDVP:Fhi(Hz)), minimum vocal fundamental frequency (MDVP:Flo(Hz)), five measures of variation in fundamental frequency (MDVP: Jitter(%), MDVP: Jitter(Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP), six measures of variation in amplitude (MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, Shimmer: DDA), two measures of ratio of noise to tonal components in the voice (NHR, HNR), two nonlinear dynamical complexity measures(RPDE, D2), signal fractal scaling exponent (DFA), three nonlinear measures of fundamental frequency variation (spread1, spread2, PPE). In this section, the predictors are scaled to have mean zero and unit variance with Z-score, and then the boxplot is drawn. The results are shown in Figure 1. From Figure 1, we can find that there are outliers in the dataset.

Next, we apply ARVSP method, WRVSP method and PWL method to analyze this biological voice measurement dataset. The results are given in Table 2.

Table 2. Variable selection results of a Parkinsons dataset.

Variables	ARVSP	PML	WRVSP
Intercept	6.0751	2.5091	1.7067
MDVP:Fo(Hz)	0	0	0
MDVP:Fhi(Hz)	-0.6783	-0.0087	0
MDVP:Flo(Hz)	0	0	0
MDVP:Jitter(%)	-7.6771	-2.0984	0
MDVP:Jitter(Abs)	0	0	0
MDVP:RAP	0	0	0
MDVP:PPQ	0	0	0
Jitter:DDP	7.0520	1.9057	0
MDVP:Shimmer	0	0	0
MDVP:Shimmer(dB)	0	0	0
Shimmer:APQ3	0	0	0.2904
Shimmer:APQ5	-0.8083	0	0
MDVP:APQ	9.0565	2.9141	0.2596
Shimmer:DDA	-5.3202	-1.2894	0.0080
NHR	0	0	0
HNR	0	0	0
RPDE	-0.5139	0	0
DFA	0	0.0126	0
spread1	0	0	0
spread2	0.1921	0.2778	0
D2	0	0	0
PPE	5.4077	1.8878	1.5435

From Table 2, we find that the proposed ARVSP method, WRVSP method and PWL method selects 9 variables, 8 variables, and 4 variables, respectively. To further compare among the different methods, we calculate the confusion matrix representing the accuracy of classification results for the above methods and random forest. The corresponding results are shown in Tables 3-6.

From Tables 3-6, we find that the proposed ARVSP method possesses the smallest false positive and false negative percentages. Meanwhile, the correct classification rate for the proposed method is 91.79%, which is the highest for all four methods.

Table 3. Classification results of ARVSP method.

Actual status	Predicted results		
	0	1	error rate
0	35	13	0.2708
1	3	144	0.0204

Table 4. Classification results of PML method.

Actual status	Predicted results		
	0	1	error rate
0	34	14	0.2917
1	11	136	0.0748

Table 5. Classification results of WRVSP method.

Actual status	Predicted results		
	0	1	error rate
0	27	21	0.4375
1	11	136	0.0748

Table 6. Classification results of random forest.

Actual status	Predicted results		
	0	1	error rate
0	35	13	0.2708
1	5	142	0.0340

5. Discussion

In this paper, we proposed an adaptive robust variable selection procedure for the logistic regression model. Meanwhile, an MM algorithm is used to solve the proposed optimization problem. Furthermore, the merits of our proposed methodology were illustrated through some simulations and a real data analysis. Numerical simulation results show that when there are outliers in the dataset, the performance of our proposed method is better than that of some existing methods. By analyzing a Parkinsons dataset, our proposed method had the smallest false positive and false negative percentages and the highest correct classification rate. Finally, we will study the large sample properties of the proposed ARVSP method as future work.

Acknowledgements

Yunlu Jiang's research is partially supported by the Natural Science Foundation of Guangdong (No. 2018A030313171, No. 2019A1515011830).

References

[1] A. Bergesio and V.J. Yohai, *Projection estimators for generalized linear models*, J. Amer. Statist. Assoc. **106** (494), 661-671, 2011.

[2] A.M. Bianco and V.J. Yohai, *Robust Estimation in the Logistic Regression Model*, Robust Statistics, Data analysis, and Computer Intensive methods, Springer, 1996.

[3] Z. Bursac, C.H. Gaussh, D.K. Williams and D.W. Hosmer, *Purposeful selection of variables in logistic regression*, Source Code Biol. Med. **3** (1), 1-8, 2008.

[4] M.H. Chen, J.G. Ibrahim and C. Yiannoutsos, *Prior elicitation, variable selection and Bayesian computation for logistic regression models*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **61** (1), 223-242, 1999.

[5] P. Čížek, *Trimmed likelihood-based estimation in binary regression models*, Austrian J. Stat. **35** (2&3), 223-232, 2006.
[6] P. Čížek, *Robust and efficient adaptive estimation of binary-choice regression models*, J. Amer. Statist. Assoc. **103** (482), 687-696, 2008.
[7] C. Croux, C. Flandre and G. Haesbroeck, *The breakdown behavior of the maximum likelihood estimator in the logistic regression model*, Statist. Probab. Lett. **60** (4), 377-386, 2002.
[8] L. Davies, *The asymptotics of Rousseeuw's minimum volume ellipsoid estimator*, Ann. Statist. **20** (4), 1828-1843, 1992.
[9] D. Gervini, *Robust adaptive estimators for binary regression models*, J. Statist. Plann. Inference **131** (2), 297-311, 2005.
[10] D. Gervini and V.J. Yohai, *A class of robust and fully efficient regression estimators*, Ann. Statist. **30** (2), 583-616, 2002.
[11] M. Guns and V. Vanacker, *Logistic regression applied to natural hazards: Rare event logistic regression with replications*, Nat. Hazard Earth Sys. **12** (6), 1937-1947, 2012.
[12] Y. Güney, Y. Tuac, S. Özdemir and O. Arslan, *Robust estimation and variable selection in heteroscedastic regression model using least favorable distribution*, Comput. Statist. **36** (2), 805-827, 2021.
[13] D.R. Hunter and K. Lange, *Quantile regression via an MM algorithm*, J. Comput. Graph. Statist. **9** (1), 60-77, 2000.
[14] D.R. Hunter and K. Lange, *A tutorial on MM algorithms*, Amer. Statist. **58** (1), 30-37, 2004.
[15] Y. Jiang, Y.G. Wang, L.Y. Fu and X. Wang, *Robust estimation using modified Huber's functions with new tails*, Technometrics **61** (1), 111-122, 2019.
[16] R.J. Karunamuni, L.L. Kong and W. Tu, *Efficient robust doubly adaptive regularized regression with applications*, Stat. Methods Med. Res. **28** (7), 2210-2226, 2019.
[17] R.L. Kennedy, A.M. Burton, H.S. Fraser, L.N. McStay and R.F. Harrison, *Early diagnosis of acute myocardial infarction using clinical and electrocardiographic data at presentation: Derivation and evaluation of logistic regression models*, Eur. Heart J. **17** (8), 1181-1191, 1996.
[18] S.K. Kinney and D.B. Dunson, *Fixed and random effects selection in linear and logistic models*, Biometrics **63** (3), 690-698, 2007.
[19] Y. Li and J.S. Liu, *Robust variable and interaction selection for logistic regression and general index models*, J. Amer. Statist. Assoc. **114** (525), 271-286, 2019.
[20] M.A. Little, P.E. McSharry, S.J. Roberts, D.A. Costello and I.M. Moroz, *Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection*, Biomed. Eng. Online **6** (1), 23, 2007.
[21] R.A. Maronna, *Robust ridge regression for high-dimensional data*, Technometrics **53** (1), 44-53, 2011.
[22] L. Meier, S.A. van de Geer and P. Bühlmann, *The group lasso for logistic regression*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **70** (1), 53-71, 2008.
[23] L. Ohno-Machado, *Modeling medical prognosis: Survival analysis techniques*, J. Biomed. Inform. **34** (6), 428-439, 2001.
[24] H. Park and S. Konishi, *Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection*, J. Stat. Comput. Simul. **86** (7), 1450-1461, 2016.
[25] D. Pregibon, *Logistic regression diagnostics*, Ann. Statist. **9** (4), 705-724, 1981.
[26] P.J. Rousseeuw, *Multivariate estimation with high breakdown point*, in: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (ed.) Mathematical Statistics and Applications, Reidel, 1985.
[27] P.J. Rousseeuw and B.C. van Zomeren, *Unmasking multivariate outliers and leverage points*, J. Amer. Statist. Assoc. **85** (411), 633-639, 1990.

[28] L.A. Stefanski, R.J. Carroll and D. Ruppert, *Optimally hounded score functions for generalized linear models with applications to logistic regression*, Biometrika **73** (2), 413-424, 1986.

[29] S. Vinterbo and L. Ohno-Machado, *A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction*, in: Proceedings of the AMIA Symposium, Washington, 984-988, 1999.

[30] S. Wang, X.Q. Jiang, Y. Wu, L.J. Cui, S. Cheng and L. Ohno-Machado, *Expectation Propagation Logistic Regression (EXPLORER): Distributed privacy-preserving online model learning*, J. Biomed. Inform. **46** (3), 480-496, 2013.

[31] X. Wang, Y. Jiang, M. Huang and H. Zhang, *Robust variable selection with exponential squared loss*, J. Amer. Statist. Assoc. **108** (502), 632-643, 2013.

[32] F. Xue and A. Qu, *Variable selection for highly correlated predictors*, arXiv: 1709.04840 [stat.ME].

[33] D. Zellner, F. Keller and G.E. Zellner, *Variable selection in logistic regression models*, Comm. Statist. Simulation Comput. **33** (3), 787-805, 2004.

[34] C.X. Zhang, S. Xu and J.S. Zhang, *A novel variational Bayesian method for variable selection in logistic regression models*, Comput. Statist. Data Anal. **133** (7), 1-19, 2019.