American Journal of Applied Mathematics and Statistics, 2025, Vol. 13, No. 2, 24-29
Available online at http://pubs.sciepub.com/ajams/13/2/1
Published by Science and Education Publishing
DOI:10.12691/ajams-13-2-1

Variable Selection for Sparse Logistic Regression Model with Errors in Covariates

Zanhua Yin*, Zhichao Wang

School of Mathematical and Computer Sciences, Gannan Normal University, Ganzhou, People's Republic of China
*Corresponding author: yinzh226@163.com

Received March 01, 2025; Revised April 01, 2025; Accepted April 08, 2025

Abstract This paper addresses variable selection problems in sparse logistic regression model with errors-in-covariates. We propose a corrected score Lasso method, which combines the weighted score Lasso approach with a projected gradient descent algorithm, to handle the challenges posed by measurement errors. The weighted score Lasso introduces a correction-amenable score function, enabling direct extension to measurement error scenarios through subsequent score correction. Our method bridges the gap between rigorous measurement error correction and practical high-dimensional implementation, establishing a framework extensible to other generalized linear models with exponential family structure. Numerical studies demonstrate the superior performance of the corrected score Lasso in error correction scenarios, highlighting its potential as a robust tool for high-dimensional data analysis with measurement error.

Keywords: *Corrected score,Lasso, logistic regression model,measurement error, sparse*

Cite This Article: Zanhua Yin, and Zhichao Wang, "Variable Selection for Sparse Logistic Regression Model with Errors in Covariates." *American Journal of Applied Mathematics and Statistics*, vol. 13, no. 2 (2025): 24-29. doi: 10.12691/ajams-13-2-1.

1. Introduction

Within the modern framework of statistical inference, the logistic regression model serves as the theoretical cornerstone for binary classification analysis, with its canonical form defined as:

$$P(Y_i=1)=F(x_i^T\beta^0)=\frac{\exp(x_i^T\beta^0)}{1+\exp(x_i^T\beta^0)} \quad (1)$$

where the independent Bernoulli response variable $Y_i\in\{0,1\}, i=1,\cdots,n$, associates with deterministic covariates $x_i\in \mathrm{R}^p, i=1,\ldots,n$, through an unknown s - sparse parameter vector β^0 (i.e., β^0 has only s non-zero components). The focus of the research is on statistical inference in high-dimensional settings ($p>>n$), which essentially constitutes a variable selection (or model selection) problem.

To address the issue of variable selection, a variety of well-established regularization methods have been developed. The evolution of regularization methods, which originated from the Lasso framework [1], has given rise to advancements such as non - convex penalties in SCAD [2] , the elastic net [3], group Lasso [4], and the Dantzig selector [5]. These methodologies are extensively detailed in the authoritative monograph [6].

Moreover, adaptations of these methodologies to sparse logistic regression models have garnered considerable interest. Notably, Zou [7] introduced the adaptive Lasso, while Loh and Wainwright [8] explored regularized M-estimators incorporating non-convex penalties like SCAD, MCP, and capped ℓ_1 norms. Yin [9] presented the weighted score Lasso method, and Zhong et al. [10] proposed a penalized weighted score function approach to tackle group sparsity in data. In the context of robust estimation, Cornilly et al. [11] put forward a method based on the elastic net framework, and Basu et al. [12] developed a robust estimation technique for adaptive Lasso.

However, an implicit assumption in all the work mentioned so far is that the covariates are directly observable without errors. This assumption, however, is seldom true in many practical studies. Common examples include sensor network data [13], gene expression microarray data [14] and high-throughput sequencing data [15]. In such cases, the unobserved covariate x_i is related to the observed covariate W_i through an additive measurement error model

$$W_i=x_i+U_i \quad (2)$$

where U_i represents the measurement error. In classical regression models, simply replacing x_i with W_i leads to a naive estimator, which is well-known to be inconsistent and biased (see [16] for a comprehensive review). Consequently, the variable selection methods discussed earlier are no longer valid in the presence of measurement error.

To mitigate the impact of measurement error, several

correction methods for variable selection have been proposed in the literature. These include the Matrix Uncertainty (MU) selector [17], the improved MU selector [18], the regularized M-estimators [19], the orthogonal matching pursuit algorithm [20], the Convex Conditioned Lasso (CoCoLasso) [21] and the Block coordinate Descent Convex Conditioned Lasso (BDCoCoLasso) algorithm [22] have been studied in the literature. However, most of these methods focus on linear models with measurement error. For high-dimensional generalized linear models, Sørensen et al. [23] and Sørensen et al. [24] proposed the generalized matrix uncertainly selector and the conditional score Lasso respectively. Chen [25] developed BOOME, a Boosting algorithm for measurement error in logistic regression and probit models. In the latter part of this paper, we propose a corrected score Lasso method to study sparse logistic regression with errors in covariates, leveraging the weighted score Lasso proposed by Yin [9] and the projected gradient descent algorithm suggested by Loh and Wainwright [19].

The remainder of this paper is organized as follows. In Section 2, we review the weighted score Lasso for sparse logistic regression without errors in covariates. Subsequently, in Section 3, we extend this approach to sparse logistic regression with errors in covariates and introduce our corrected score Lasso method. Section 4 presents numerical results obtained through simulations.

Notation: Write $X=(x_{ij})=(x_1,\cdots,x_n)^T\in \mathrm{R}^{n\times p}$, $W=(W_{ij})=(W_1,\cdots,W_n)^T\in \mathrm{R}^{n\times p}$ and $Y=(Y_1,\cdots,Y_n)^T$. Denote by $I=\{j:\beta_j^0\neq 0\}$ the non-zero coordinate of β^0 and let $s=card\{I\}$ be the number of non-zero elements of β^0. For a vector $\mathrm{V}=(v_1,\cdots,v_p)^T\in \mathrm{R}^p$, we define $\|\mathrm{V}\|_1=\sum_{j=1}^p|v_j|$ and $\|\mathrm{V}\|_\infty=\max_{1\le j\le p}|v_j|$. For function $f(\mathrm{V})\in \mathrm{R}$, we denote by $\nabla f(\mathrm{V})\in \mathrm{R}^p$ its gradient at $\mathrm{V}\in \mathrm{R}^p$.

2. Weighted Score Lasso

We now review weighted score Lasso proposed by Yin [9]. The population loss corresponding to the negative log-likelihood in model (1) is formulated as

$$\ell(\beta;X,Y)=-\frac{1}{n}\sum_{i=1}^n\{(x_i^T\beta)Y_i-\log(1+\exp(x_i^T\beta))\}$$

The sparse logistic regression model can be obtained through the ℓ_1-penalized population loss minimization:

$$\min_{\beta\in \mathrm{R}^p}\{\ell(\beta;X,Y)+\lambda\|\beta\|_1\} \qquad (3)$$

where $\lambda>0$ serves as a tuning parameter. The solution to this Lasso problem satisfies the Karush-Kuhn-Tucker (KKT) conditions:

$$\frac{1}{n}\sum_{i=1}^n[F(x_i^T\beta)-Y_i]x_i+\lambda\vec{v}=0$$

where $\vec{v}_j$ represents the subgradient of $|\beta_j|$ which is the sign of β_j if $\beta_j\neq 0$ and can be any value belonging to [0, 1] when $\beta_j=0$. A critical challenge arises in parameter selection: the optimal λ must satisfy $\|\nabla\ell(\beta^0;X,Y)\|_\infty\le c\lambda$ for some constant $0<c<1$. However, the score function valued at $\beta=\beta^0$:

$$\nabla\ell(\beta^0;X,Y)=\frac{1}{n}\sum_{i=1}^n[F(x_i^T\beta^0)-Y_i]x_i,$$

contains a random component $F(x_i^T\beta^0)-Y_i$ with variance $F(x_i^T\beta^0)(1-F(x_i^T\beta^0))$, introducing dependence on the true value β^0 (unknown) in λ selection.

To address this fundamental difficulty, Yin [9] proposed a weighted score Lasso approach through innovative weighting of the score function components. The weighted score Lasso solves:

$$\frac{1}{n}\sum_{i=1}^n\omega(x_i^T\beta)[F(x_i^T\beta)-Y_i]x_i+\lambda\vec{v}=0,$$

where $\omega(\cdot)$ is a predefined positive weight function. The specific weight function:

$$\omega(x_i^T\beta)=\frac{1}{2}\exp\left(\frac{x_i^T\beta}{2}\right)+\frac{1}{2}\exp\left(-\frac{x_i^T\beta}{2}\right) \qquad (4)$$

yields the weighted score function:

$$\nabla\ell_\omega(\beta;X,Y)=\frac{1}{2n}\sum_{i=1}^n\left\{(1-Y_i)\exp\left(\frac{x_i^T\beta}{2}\right)-Y_i\exp\left(-\frac{x_i^T\beta}{2}\right)\right\}x_i \qquad (5)$$

This formulation leads to the weighted score Lasso estimator, defined as the solution to

$$\min_{\beta\in \mathrm{R}^p}\{\ell_\omega(\beta;X,Y)+\lambda\|\beta\|_1\} \qquad (6)$$

where the weighted loss function takes the form:

$$\ell_\omega(\beta;X,Y)=\frac{1}{n}\sum_{i=1}^n\left\{(1-Y_i)\exp\left(\frac{x_i^T\beta}{2}\right)+Y_i\exp\left(-\frac{x_i^T\beta}{2}\right)\right\}$$

A key advantage of this weighted approach lies in its correction-amenable score function, enabling direct extension to measurement error scenarios through subsequent score correction. The weighted formulation maintains convexity while decoupling the λ selection from dependence on β^0, resolving the technical challenge present in standard logistic regression Lasso.

3. Corrected Score Lasso

Building upon the weighted score framework, we now address the critical challenge of covariate measurement error through innovative methodological extensions. Consider the logistic regression model (1) augmented with an additive measurement error structure:

$$\begin{cases} \mathrm{P}(Y_i = 1) = F(x_i^T \beta^0), \\ W_i = x_i + U_i, \end{cases} \quad (7)$$

where W_i represents error-contaminated covariates, U_i follows Normal $(0, \Sigma_{uu})$. Direct substitution of x_i with W_i in standard Lasso implementations produces biased naive estimators—a well-documented phenomenon in measurement error literature [16]. This bias persists in high-dimensional settings, necessitating specialized correction mechanisms.

The corrected score approach circumvents this limitation by constructing an unbiased estimating function $\nabla \ell^*(\beta; W, Y)$ that satisfies

$$\mathrm{E}\left(\nabla \ell^*(\beta; W, Y) \middle| (Y, X)\right) = \nabla \ell(\beta; X, Y).$$

Since the score is unbiased, it follows that the corrected score, if it exists, is also unbiased. Therefore, corrected scores yield consistent estimators. However, Stefanski [26] established the non-existence of conventional corrected scores for logistic regression, our weighted score formulation enables correction through strategic exploitation of the measurement error structure.

Given a corrected score function $\nabla \ell_\omega^*(\beta; W, Y)$, it is natural to consider the following optimization problem:

$$\min_{\beta \in \mathrm{R}^p} \left\{ \ell_\omega^*(\beta; W, Y) + \lambda \|\beta\|_1 \right\} \quad (8)$$

Although the above solution seems very natural, the optimization program (8) is fundamentally different from the optimization program (6). When the dataset is corrupted by measurement errors, $\ell_\omega^*(\beta; W, Y)$ is not usually convex. Hence the difference is, the optimization program (8) is nonconvex, the optimization program (6) is convex. To overcome this difficulty, we extend the algorithm proposed by Loh and Wainwright [19] for linear models with measurement error, to model (8).

Under Gaussian measurement errors, we establish the pivotal moment relationships:

$$\mathrm{E}\left\{ \exp\left(\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) \middle| x_i \right\} = \exp\left(\frac{x_i^T \beta}{2} \right)$$

$$\mathrm{E}\left\{ \exp\left(\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) \left(W_i - \frac{\Sigma_{uu} \beta}{2} \right) \middle| x_i \right\} = \exp\left(\frac{x_i^T \beta}{2} \right) x_i$$

Due to the conditional independence of W_i and Y_i given x_i, we can obtain a correction score function based on (5), which has the form

$$\begin{aligned} &\nabla \ell_\omega^*(\beta; W, Y) \\ &= \frac{1}{2n} \sum_{i=1}^n \left\{ (1 - Y_i) \exp\left(\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) \left(W_i - \frac{\Sigma_{uu} \beta}{2} \right) \right. \\ &\quad \left. - Y_i \exp\left(-\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) \left(W_i + \frac{\Sigma_{uu} \beta}{2} \right) \right\} \end{aligned} \quad (9)$$

The corresponding corrected loss function becomes:

$$\ell_\omega^*(\beta; W, Y) = \frac{1}{n} \sum_{i=1}^n \left\{ (1 - Y_i) \exp\left(\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) + Y_i \exp\left(-\frac{W_i^T \beta}{2} - \frac{\beta^T \Sigma_{uu} \beta}{8} \right) \right\}$$

This formulation preserves the essential unbiasedness property $\mathrm{E}\left(\ell_\omega^*(\beta; W, Y) \middle| (Y, X)\right) = \ell_\omega(\beta; X, Y)$ under perfect covariate observation. When $n > p$, the estimator for β^0 can be obtained by minimizing $\ell_\omega^*(\beta; W, Y)$ using the standard gradient descent method. However, when $n < p$, without addition regularization, optimizing $\ell_\omega^*(\beta; W, Y)$ is an ill-posed mathematical problem because it does not have a unique solution.

To address the dual challenges of non-convex optimization and high-dimensionality ($p >> n$), we adapt the projected gradient descent framework [19] to our corrected score context. The proposed corrected score Lasso estimator solves:

$$\tilde{\beta} \in \underset{\|\beta\|_1 \le \kappa}{\operatorname{argmin}} \left\{ \ell_\omega^*(\beta; W, Y) + \lambda \|\beta\|_1 \right\} \quad (10)$$

where $\kappa > 0$ constrains the parameter space to ensure feasibility of β^0. The projection step onto the ℓ_1-ball $\{\beta : \|\beta\|_1 \le \kappa\}$ combines naturally with ℓ_1-regularization to enforce sparsity while maintaining computational tractability. The projected gradient descent algorithm generates a sequence $\{\tilde{\beta}_t, t = 0, 1, 2, \ldots\}$ of iterates by the recursion:

$$\tilde{\beta}^{t+1} = \Pi\left\{ \tilde{\beta}^t - \eta \nabla \ell_\omega^*\left(\tilde{\beta}^t; W, Y \right) \right\},$$

where Π denotes projection onto the ℓ_1-ball of radius κ, and η is the step size parameter. As Loh and Wainwright [19] showed, if κ is properly chosen, the above iterates converge with high probability to a vector extremely close to any global minimizers of the program (10).

4. Numerical Studies

In this section we use simulated datasets to investigate the finite sample performances of proposed procedures. The performance of estimator β^{est} is assessed by the Mean Absolute Error (MAE)

$$\mathrm{MAE}(\beta^{\text{est}}) = \|\beta^{\text{est}} - \beta^0\|_1$$

To assess the ability of variable selection methods for recovering the varying sparsity, we record “TP” and “FP” that denote the number of true positives and the number of false positives, respectively.

We generate 100 datasets from model (1), each consisting of n=200 observations. We set $s = 10$ and $\beta^0 = (1, \cdots, 1, 0, \cdots, 0)^T$, for $p = 500$ or 1000, respectively. The matrix X had i.i.d. rows $x_i \sim N(0, \Sigma_x)$, for $i = 1, \cdots, n$.

We consider two model for covariance matrix Σ_x, component independent ($\Sigma_{x,jk} = I\{j=k\}$) and autoregressive ($\Sigma_{x,jk} = 0.5^{|j-k|}$). The response Y_i is binomially distributed with mean $F(x_i^T\beta^0)$. A measurement matrix $W = X + U$ is generated with the rows of U i.i.d. distributed Normal $(0, \sigma_u^2 \mathrm{I})$ where $\sigma_u^2 = 0.2$ or $\sigma_u^2 = 0.5$.

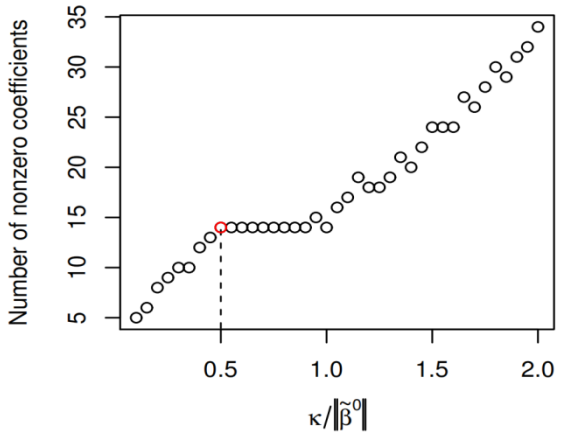

Figure 1. The elbow rule for corrected score Lasso, when covariance matrix Σ_x has entries $\Sigma_{x,jk} = I\{j=k\}$, $j,k = 1,\ldots,p$ and $\sigma_u^2 = 0.2$ with $n = 200$ and $p = 500$

We use R package "glmnet" to solve program (3), where two tuning parameters are chosen via 10-fold cross-validation: one (denote $\hat{\lambda}_{\min}$) corresponding to the minimum deviance, and one (denote $\hat{\lambda}_{se}$) corresponding to the "one-standard-error" rule [27]. We apply weighted score Lasso estimate (denoted as "WsLasso") to solve programm (6), where the settings of $\omega(\cdot)$, λ and ε refer to the paper by [9]. For comparison, we consider "glmnet" and "WsLasso" estimators based on corrupted data (Y,W). Our corrected score Lasso estimator (denoted as "CsLasso") can be computed using an efficient projection algorithm proposed by [28]. CsLasso requires an initial estimator $\tilde{\beta}^0$ which was given by glmnet (with $\hat{\lambda}_{\min}$) estimate based on (Y,W). CsLasso also requires knowledge of $\|\beta^0\|_1$ or a loss function for choosing the constraint parameter κ. Since, β^0 is impossible to know beforehand and CsLasso lacks a well-defined loss function, the elbow rule [17] is used to select the constraint parameter κ from 39 equally spaced values in $\left[0.1\times\|\tilde{\beta}^0\|_1, 2\times\|\tilde{\beta}^0\|_1\right]$. Figure 1 shows the number of nonzero coefficients plotted against the value of $\kappa/\|\beta^0\|_1$, and the eblow rule now amounts to selecting κ where the curve begins to flatten.

Table 1, Table 2, Table 3 and Table 4 offer a systematic quantification of the error correction superiority of CsLasso over both glmnet (with $\hat{\lambda}_{se}$ and $\hat{\lambda}_{\min}$) and WsLasso. The comparative analysis reveals three critical advantages:

1. Enhanced Selection Accuracy: CsLasso achieves higher true positives (TPs) than glmnet with $\hat{\lambda}_{se}$ while simultaneously reducing false positives (FPs). Although WsLasso demonstrates the lowest FPs, its substantially compromised TPs reveal excessive conservatism. In contrast, CsLasso maintains greater detection power without sacrificing specificity.

2. Measurement Error Resilience: Under escalating measurement error (σ_u^2 from 0.2 to 0.5), CsLasso exhibits superior stability with less TPs reduction compared to glmnet and WsLasso. While greater σ_u^2 leads to a smaller TP for all methods, CsLasso shows only a slight reduction in TP and an increase in FP. This robustness stems from CsLasso's integrated measurement error correction with its optimization framework.

Table 1. Results from the simulation study, when covariance matrix Σ_x has entries $\Sigma_{x,jk} = I\{j=k\}$, $j,k = 1,\ldots,p$, and $\Sigma_{uu} = \sigma_u^2\mathrm{I}$, $p = 500$. Reported numbers are the averages and standard errors (show in parentheses)

$\sigma_u^2 = 0.2$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	9.03(0.92)	35.65(12.65)	10.21(0.92)
glmnet($\hat{\lambda}_{se}$)	7.24(2.00)	11.64(13.64)	9.54(0.52)
WsLasso ($\varepsilon = 0.025$)	6.70(1.34)	4.19(2.32)	9.23(0.40)
CsLasso	9.38(1.44)	7.28(6.09)	9.12(0.49)
$\sigma_u^2 = 0.5$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	7.43(1.60)	24.56(13.60)	10.26(0.76)
glmnet($\hat{\lambda}_{se}$)	6.13(2.26)	15.60(21.37)	10.24(1.53)
WsLasso ($\varepsilon = 0.025$)	3.54(1.37)	1.36(1.21)	9.74(0.19)
CsLasso	8.60(1.60)	8.00(5.07)	9.49(0.65)

Table 2. Results from the simulation study, when covariance matrix Σ_x has entries $\Sigma_{x,jk} = I\{j=k\}$, $j,k = 1,\ldots,p$, and $\Sigma_{uu} = \sigma_u^2\mathrm{I}$, $p = 1000$. Reported numbers are the averages and standard errors (show in parentheses)

$\sigma_u^2 = 0.2$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	8.05(1.30)	34.40(17.51)	10.48(1.11)
glmnet($\hat{\lambda}_{se}$)	6.38(1.95)	13.51(16.88)	9.88(0.91)
WsLasso ($\varepsilon = 0.025$)	5.91(1.29)	5.41(2.46)	9.46(0.38)
CsLasso	9.26(1.64)	11.67(8.25)	9.44(0.75)
$\sigma_u^2 = 0.5$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	6.09(2.09)	24.42(16.72)	10.50(0.87)
glmnet($\hat{\lambda}_{se}$)	4.93(2.23)	16.14(22.13)	10.42(1.54)
WsLasso ($\varepsilon = 0.025$)	2.66(1.06)	1.52(1.18)	9.85(0.14)
CsLasso	8.24(1.83)	12.15(6.12)	9.87(0.83)

Table 3. Results from the simulation study, when covariance matrix Σ_x has entries $\Sigma_{x,jk} = 0.5^{|j-k|}$, $j,k = 1,\ldots,p$, and $\Sigma_{uu} = \sigma_u^2 \mathrm{I}$, $p = 500$. Reported numbers are the averages and standard errors (show in parentheses)

$\sigma_u^2 = 0.2$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	9.76(0.43)	34.95(15.02)	8.98(1.15)
glmnet($\hat{\lambda}_{se}$)	8.37(1.13)	2.77(8.70)	8.55(0.57)
WsLasso ($\varepsilon = 0.025$)	8.03(0.98)	0.13(0.39)	8.15(0.42)
CsLasso	9.68(0.85)	5.70(5.94)	7.26(0.77)
$\sigma_u^2 = 0.5$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	9.51(0.61)	30.78(14.68)	9.31(1.09)
glmnet($\hat{\lambda}_{se}$)	7.98(1.30)	2.96(6.95)	8.88(0.42)
WsLasso ($\varepsilon = 0.025$)	5.17(1.06)	0.03(0.17)	9.29(0.27)
CsLasso	8.91(1.11)	2.93(3.22)	7.86(0.70)

Table 4. Results from the simulation study, when covariance matrix Σ_x has entries $\Sigma_{x,jk} = 0.5^{|j-k|}$, $j,k = 1,\ldots,p$, and $\Sigma_{uu} = \sigma_u^2 \mathrm{I}$, $p = 1000$. Reported numbers are the averages and standard errors (show in parentheses)

$\sigma_u^2 = 0.2$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	9.54(0.56)	39.78(16.48)	9.45(1.28)
glmnet($\hat{\lambda}_{se}$)	8.12(1.21)	3.42(7.37)	8.61(0.41)
WsLasso ($\varepsilon = 0.025$)	7.46(0.99)	0.14(0.38)	8.40(0.40)
CsLasso	9.25(0.94)	10.05(7.60)	7.50(0.79)
$\sigma_u^2 = 0.5$	TP	FP	MAE
glmnet($\hat{\lambda}_{\min}$)	9.32(0.65)	35.50(17.41)	9.74(1.07)
glmnet($\hat{\lambda}_{se}$)	7.37(1.64)	3.27(6.12)	9.06(0.27)
WsLasso ($\varepsilon = 0.025$)	4.18(1.13)	0.00(0.00)	9.53(0.20)
CsLasso	8.57(1.17)	4.40(4.36)	8.16(0.62)

3. Estimation Precision Dominance: CsLasso has the lowest MAE in almost all cases. In high-error conditions ($\sigma_u^2 = 0.5$), this advantage becomes even more pronounced, with CsLasso's MAE being significantly lower than those of glmnet and WsLasso.

These findings collectively demonstrate the superior performance of CsLasso in error correction scenarios, highlighting its potential as a robust tool for high-dimensional data analysis with measurement error.

5. Conclusion

This paper focuses on variable selection in sparse logistic regression model when covariates are subject to measurement error. The key contribution of this study is the development of the corrected score Lasso (CsLasso) method, which effectively addresses the challenges posed by measurement errors in covariates.

The CsLasso method builds upon the weighted score Lasso framework, introducing a correction-amenable score function that allows for direct extension to measurement error scenarios through subsequent score correction. This approach maintains convexity while decoupling the regularization parameter selection from dependence on the true parameter value, thereby resolving a technical challenge present in standard logistic regression Lasso methods.

The numerical studies conducted in this paper demonstrate the superior performance of CsLasso in error correction scenarios. CsLasso achieves higher true positives (TPs) than glmnet with $\hat{\lambda}_{se}$ while simultaneously reducing false positives (FPs). It also exhibits superior stability under escalating measurement error, with less reduction in TPs compared to glmnet and WsLasso. Furthermore, CsLasso consistently achieves the lowest Mean Absolute Error (MAE) in almost all cases, with this advantage becoming even more pronounced in high-error conditions.

The methodology proposed in this paper bridges the gap between rigorous measurement error correction and practical high-dimensional implementation. It establishes a framework that is extensible to other generalized linear models with exponential family structure, highlighting its potential as a robust tool for high-dimensional data analysis with measurement error.

Future research could explore the application of the CsLasso method to other types of generalized linear models and further investigate its theoretical properties in ultra-high dimensional settings. Additionally, the development of more efficient algorithms for implementing CsLasso in large-scale data analysis could be a valuable direction for future work.

ACKNOWLEDGEMENTS

The authors' work was supported by the Educational Commission of Jiangxi Province of China (GJJ211403).

References

[1] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288, 1996.

[2] Fan, J. and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association 96 (456), 1348-1360, 2001.

[3] Zou, H. and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2), 301-320, 2005.

[4] Yuan, M. and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49-67, 2006.

[5] Candes, E. and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2313-2351, 2007.

[6] Bühlmann, P. and S. Van De Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

[7] Zou, H. The adaptive lasso and its oracle properties. Journal of the American statistical association 101 (476), 1418-1429, 2006.

[8] Loh, P.-L. and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In Advances in Neural Information Processing Systems, pp. 476-484, 2013.

[9] Yin, Z. Variable selection for sparse logistic regression. Metrika 83 (7), 821-836, 2020.

[10] Zhong, M., Z. Yin, and Z. Wang. Variable selection for sparse logistic regression with grouped variables. Mathematics 11 (24), 4979, 2023.

[11] Cornilly, D., L. Tubex, S. Van Aelst, and T. Verdonck. Robust and sparse logistic regression. Advances in Data Analysis and Classification 18 (3), 663-679, 2024.

[12] Basu, A., A. Ghosh, M. Jaenada, and L. Pardo. Robust adaptive lasso in high-dimensional logistic regression. Statistical Methods & Applications 33 (5), 1217-1249, 2024.

[13] Feng, J., S. Megerian, and M. Potkonjak. Model-based calibration for sensor networks. In Sensors, Proceedings of IEEE, Volume 2, pp. 737-742, 2003.

[14] Purdom, E. and S. P. Holmes. Error distribution for gene expression data. Statistical applications in genetics and molecular biology 4 (1), 2005.

[15] Benjamini, Y. and T. P. Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. Nucleic acids research 40 (10), e72-e72, 2012.

[16] Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. Measurement error in nonlinear models: a modern perspective. CRC press, 2006.

[17] Rosenbaum, M., A. B. Tsybakov, et al. Sparse recovery under matrix uncertainty. The Annals of Statistics 38 (5), 2620-2651, 2010.

[18] Rosenbaum, M. and A. B. Tsybakov. Improved matrix uncertainty selector. 9, 276-291, 2013.

[19] Loh, P.-L. and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In Advances in Neural Information Processing Systems, pp. 2726-2734, 2011.

[20] Chen, Y. and C. Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high dimensional results. arXiv preprint arXiv:1206.0823, 2012.

[21] Datta, A. and H. Zou. Cocolasso for high-dimensional error-in-variables regression. Annals of Statistics 45 (6), 2400-2426, 2017.

[22] Escribe, C., T. Lu, J. Keller-Baruch, V. Forgetta, B. Xiao, J. B. Richards, S. Bhatnagar, K. Oualkacha, and C. M. Greenwood. Block coordinate descent algorithm improves variable selection and estimation in error-in-variables regression. Genetic Epidemiology 45 (8), 874-890, 2021.

[23] Sørensen, Ø., A. Frigessi, and M. Thoresen. Measurement error in lasso: Impact and likelihood bias correction. Statistica Sinica, 809-829, 2015.

[24] Sørensen, Ø., K. H. Hellton, A. Frigessi, and M. Thoresen. Covariate selection in high-dimensional generalized linear models with measurement error. Journal of Computational and Graphical Statistics 27 (4), 739-749, 2018.

[25] Chen, L.-P. Variable selection and estimation for misclassified binary responses and multivariate error-prone predictors. Journal of Computational and Graphical Statistics 33 (2), 407-420, 2024.

[26] Stefanski, L. A. Unbiased estimation of a nonlinear function a normal mean with application to measurement error of models. Communications in Statistics-Theory and Methods 18 (12), 4335-4358, 1989.

[27] Friedman, J., T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33 (1), 1-22, 2010.

[28] Duchi, J., S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In Proceedings of the 25th international conference on Machine learning, pp. 272-279, 2008. ACM.