mathematics

MDPI

Article

Variable Selection for Sparse Logistic Regression with Grouped Variables

Mingrui Zhong, Zanhua Yin * and Zhichao Wang

School of Mathematics and Computer Science, Gannan Normal University, Ganzhou 341000, China; zmrui_stat@gnnu.edu.cn (M.Z.); wangzhichao@gnnu.edu.cn (Z.W.)
* Correspondence: yinzanhua@gnnu.edu.cn or yinzh226@163.com

Abstract: We present a new penalized method for estimation in sparse logistic regression models with a group structure. Group sparsity implies that we should consider the Group Lasso penalty. In contrast to penalized log-likelihood estimation, our method can be viewed as a penalized weighted score function method. Under some mild conditions, we provide non-asymptotic oracle inequalities promoting the group sparsity of predictors. A modified block coordinate descent algorithm based on a weighted score function is also employed. The net advantage of our algorithm over existing Group Lasso-type procedures is that the tuning parameter can be pre-specified. The simulations show that this algorithm is considerably faster and more stable than competing methods. Finally, we illustrate our methodology with two real data sets.

Keywords: high-dimensional data; non-asymptotic inequality; logistic regression; variable selection; block coordinate descent algorithm

MSC: 62J12

Citation: Zhong, M.; Yin, Z.; Wang, Z. Variable Selection for Sparse Logistic Regression with Grouped Variables. *Mathematics* **2023**, *11*, 4979. https://doi.org/10.3390/math11244979

Academic Editor: Heng Lian

Received: 16 November 2023
Revised: 7 December 2023
Accepted: 11 December 2023
Published: 17 December 2023

1. Introduction

Logistic regression models are a powerful and popular technique for modeling the relationship between the predictors and a categorical response variable. Let $(x_1, y_1), \cdots, (x_n, y_n)$ be independent pairs of observed data which are realizations of a random vector (X, Y), with p-dimensional predictors $X \in \mathbb{R}^p$ and univariate binary response variable $Y \in \{0, 1\}$. (X, Y) is assumed to satisfy

$$\mathbb{P}(Y = 1|X = x) = G(x^T\beta^0) = \frac{\exp(x^T\beta^0)}{1 + \exp(x^T\beta^0)}, \tag{1}$$

where $\beta^0 \in \mathbb{R}^p$ is a regression vector to be estimated. We are especially concerned with a sparse logistic regression problem in which the dimension p is high and the sample size n might be small, i.e., the so-called "small n, large p" framework, which is a variable selection problem for high-dimensional data.

When dealing with high-dimensional data, there are usually two important considerations: model sparsity and prediction accuracy. The Lasso [1] was proposed to address these two objectives, since Lasso can determine submodels with a moderate number of parameters that still fit the data adequately. There are also other similar methods including SCAD [2], elastic net [3], Dantzig selector [4], MCP [5] and so on. In high-dimensional logistic regression models, Lasso study topics range from asymptotic results, including the consistency and asymptotic distribution of the estimator, e.g., Sur et al. [6], Ma et al. [7], Bianco et al. [8], to non-asymptotic results, including the non-asymptotic oracle inequalities of the estimation and prediction errors, e.g., Abramovich et al. [9], Huang et al. [10] and Yin [11].

In many applications, predictors can often be thought of as grouped. For example, in genome-wide association studies (GWASs), genes usually do not act individually, but are reflected in the covariation of several genes with each other. Additionally, in histologically normal epithelium (NlEpi) studies, we need to consider the non-linear effects of genes for microarray data. Similar to the Lasso, considering this grouped information in the modeling process should improve the interpretability and the accuracy of the model. Yuan and Lin [12] proposed an extension of the Lasso, called the Group Lasso, which imposes an L_2 penalty to individual groups of variables and then an L_1 penalty to the resulting block norms, rather than only an L_1 penalty to individual variables. Suppose x_i and β^0 in model (1) are divided into g known groups, where we consider a partition $\{G_1, \cdots, G_g\}$ of $\{1, \cdots, p\}$ into groups and denote the cardinality of a group G_l by $|G_l|$, $x_i = (x_{i(1)}^T, x_{i(2)}^T, \cdots, x_{i(g)}^T)^T$, $\beta^0 = ((\beta_{(1)}^0)^T, (\beta_{(2)}^0)^T, \cdots, (\beta_{(g)}^0)^T)^T$, $x_{i(l)} \in \mathbb{R}^{|G_l|}$, $\beta_{(l)}^0 \in \mathbb{R}^{|G_l|}$. We wish to achieve sparsity at the level of groups, i.e., to β^0 such that $\beta_{(l)}^0 = 0$ for some of the groups $l \in \{1, \cdots, g\}$. When using high-dimensional logistic regression models, Group Lasso provides an estimator for β^0:

$$\hat{\beta}^{GL} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \Big\{ \log\Big(1 + \exp(x_i^T \beta)\Big) - (x_i^T \beta) y_i \Big\} + \lambda \sum_{l=1}^{g} \omega_l \|\beta_{(l)}\|_2, \tag{2}$$

where $\lambda \geq 0$ is a tuning parameter which controls the amount of penalization, $\omega_l = \sqrt{|G_l|}$ is used to normalize across groups of different sizes and $\|\cdot\|_2$ denotes the L_2 norm of a vector. Meier et al. [13] established the asymptotic consistency theory of Group Lasso for logistic regression, Wang et al. [14] analyzed the rates of convergence, Blazere et al. [15] stated oracle inequalities and Kwemou [16] and Nowakowski [17] studied non-asymptotic oracle inequalities. Furthermore, Zhang et al. [18] studied the $L_{p,q}$ regularization penalty estimates for logistic regression. In terms of computational algorithms, Meier et al. [13] applied the block coordinate descent algorithm of Tseng [19] to Group Lasso for logistic regression, ans Breheny and Huang [20] proposed the Group descent algorithm. While the aforementioned methods have shown promising performance in practical settings (Abramovich [21], Chen [22], Tyan [23], Yang [24]), a pressing issue that remains unresolved is that these approaches are just computing the exact coefficients fast enough at those selected values of λ.

However, it is well known that for the Lasso (or the Group Lasso) in linear regression models, the respective optimal values of the tuning parameter λ depend on the unknown parameter σ^2, the homogeneous noise variance, and its accurate estimation is generally more difficult when $p \gg n$. To solve this problem, Belloni et al. [25] proposed square-root Lasso, which removed this unknown parameter by using a weighted score function (i.e., the square root of the empirical loss function). Bunea et al. [26] extended the ideas behind the square-root Lasso for group selection and developed the Group square-root Lasso. Inspired by Group square-root Lasso, we propose a new penalized weighted score function method, which alternatively replaces the original score function (i.e., the gradient of negative loglikelihood function) with a weighted score function (Huang and Wang [27]) to study sparse logistic regression with a Group Lasso penalty. We obtain convergence rates for the estimation error and provide a direct choice for the tuning parameter. Moreover, we propose a modified block coordinate descent algorithm based on the weighted score function, which greatly optimizes the computational complexity.

The framework of this paper is as follows. In Section 2, we apply this idea behind the Group square-root Lasso to sparse logistic models and develop our method, the penalized weighted score function method. In Section 3, we propose asymptotic bounds for our new estimator and a direct selection for the tuning parameter. In Section 4, we provide the weighted block coordinate descent algorithm. In Section 5, numerical simulations show the advantages of our algorithm in terms of selection effects and computational time. In Section 6, we present real data for genes and musk to support the simulations and theoretical results. Section 7 concludes our work. All proofs are given in Appendix A.

Notation: Throughout the paper, the non-zero coordinate of β^0 is denoted by $I = \{l : \|\beta^0_{(l)}\|_2 \neq 0\}$ and $s = \text{card}\{I\}$ is the number of non-zero elements of β^0. For all $\delta \in \mathbb{R}^p$ and subset I, δ_I has the same coordinates as δ on I and zero coordinates on the complement I^C of I. For a function $f(\beta) \in \mathbb{R}$, we denote by $\nabla f(\beta) \in \mathbb{R}^p$ its gradient and $\mathcal{H}(\beta) \in \mathbb{R}^{p\times p}$ its Hessian matrix at $\beta \in \mathbb{R}^p$. The L_q norm of any vector v is defined as $\|v\|_q = (\sum_i |v_i|^q)^{1/q}$ and for any vector $\beta \in \mathbb{R}^p$ with group structures, the block norm of β for any $0 \leq q \leq \infty$ is denoted as $\|\beta\|_{2,q} = (\sum_{l=1}^g \|\beta_{(l)}\|_2^q)^{1/q}$. In particular, $\|\beta\|_{2,0} = \sum_{l=1}^g 1_{\beta_{(l)}\neq 0}$ indicates the number of non-zero groups, $\|\beta\|_{2,1} = \sum_{l=1}^g \|\beta_{(l)}\|_2$ represents the form of Group Lasso, $\|\beta\|_{2,2} = \|\beta\|_2$ denotes the L_2 norm, and $\|\beta\|_{2,\infty} = \max_l \|\beta_{(l)}\|_2$ means the largest L_2 norm of all groups. Moreover $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution.

2. Penalized Weighted Score Function Method

Recall that model (1), the loss function (i.e., the negative loglikelihood), is given by

$$\ell(\beta) = \frac{1}{n}\sum_{i=1}^n \Big\{\log\Big(1+\exp(x_i^T\beta)\Big) - (x_i^T\beta)y_i\Big\},$$

leading to the score function

$$\nabla\ell(\beta) = \frac{1}{n}\sum_{i=1}^n (G(x_i^T\beta) - y_i)x_i.$$

Note that the solution $\hat{\beta}^{GL}$ of model (2) satisfies KKT conditions defined as follows

$$\begin{cases} \frac{1}{n}\sum\limits_{i=1}^n (G(x_i^T\hat{\beta}^{GL}) - y_i)x_{i(l)} = -\lambda\omega_l\hat{\beta}^{GL}_{(l)}/\|\hat{\beta}^{GL}_{(l)}\|_2, & \text{if } \hat{\beta}^{GL}_{(l)} \neq 0, \\ |\frac{1}{n}\sum\limits_{i=1}^n (G(x_i^T\hat{\beta}^{GL}) - y_i)x_{i(l)}| \leq \lambda\omega_l, & \text{if } \hat{\beta}^{GL}_{(l)} = 0, \end{cases} \tag{3}$$

for all $l = 1, \cdots, g$. The left side of Equation (3) is the score function for logistic regression with a group structure, which shows that $\hat{\beta}^{GL}$ is actually a penalized score function estimator. To obtain a good estimator, we usually require that the inequality $\lambda\omega_l \geq c\|\nabla\ell(\beta^0)\|_{2,\infty}$ for all $l = 1, \cdots, g$ and some constant $c \geq 1$ holds with high probability (Meier et al. [13] and Kwemou [16]). However, the random part $G(x_i^T\beta^0) - y_i$ for $\nabla\ell(\beta^0)$, the score function valued at $\beta = \beta^0$, has variance $G(x_i^T\beta^0)(1 - G(x_i^T\beta^0))$, which is also the variance of the binary random variable $Y_i|X_i = x_i$. Obviously, binary noises are not homogeneous like the noise in linear regression models; a unique tuning parameter for all of the different coefficients is not a good choice.

We apply the idea from Group square-root Lasso to solve the above problem for choosing a tuning parameter, and develop our method as follows. Huang and Wang [27] formed a class of root-consistent estimating functions by a weighted score function for logistic regression

$$\nabla\ell_\psi(\beta) = \frac{1}{n}\sum_{i=1}^n \psi(x_i^T\beta)(G(x_i^T\beta) - y_i)x_i, \tag{4}$$

where $\psi(\cdot)$ is the weighted function of $x_i^T\beta$. This requires choosing a suitable weighed function to ensure that $\nabla\ell_\psi(\beta)$ is almost integrable for β. Then, replacing the score function in Equation (3) with the weighted score function, we develop a penalized weighted score function estimate $\hat{\beta}$, which is a solution of the following equation:

$$\begin{cases} \frac{1}{n}\sum\limits_{i=1}^n \psi(x_i^T\hat{\beta})(G(x_i^T\hat{\beta}) - y_i)x_{i(l)} = -\lambda\omega_l\hat{\beta}_{(l)}/\|\hat{\beta}_{(l)}\|_2, & \text{if } \hat{\beta}_{(l)} \neq 0, \\ |\frac{1}{n}\sum\limits_{i=1}^n \psi(x_i^T\hat{\beta})(G(x_i^T\hat{\beta}) - y_i)x_{i(l)}| \leq \lambda\omega_l, & \text{if } \hat{\beta}_{(l)} = 0. \end{cases} \tag{5}$$

Let $\ell_\psi(\beta)$ be the loss function corresponding to the weighted score function (4); the solution to Equation (5) is equivalent to solving the following optimization problem:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \ell_\psi(\beta) + \lambda \sum_{l=1}^{g} \omega_l \|\beta_{(l)}\|_2 \right\}. \tag{6}$$

Our method is motivated by Bunea et al.'s [26] minimization of the Group square-root Lasso for the linear model:

$$\hat{\beta}^{GSL} := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|\mathbb{Y} - \mathbb{X}\beta\|_2}{\sqrt{n}} + \frac{\lambda}{n} \sum_{l=1}^{g} \omega_l \|\beta_{(l)}\|_2 \right\},$$

where $\mathbb{Y} \in \mathbb{R}^{n \times 1}$ and $\mathbb{X} \in \mathbb{R}^{n \times p}$. When $\|\mathbb{Y} - \mathbb{X}\hat{\beta}^{GSL}\|_2$ is non-zero, the Group square-root Lasso estimator $\hat{\beta}^{GSL}$ satisfies the KKT condition

$$\begin{cases} \sqrt{n} \sum_{i=1}^{n} (\|\mathbb{Y} - \mathbb{X}\hat{\beta}^{GSL}\|_2)^{-1} (y_i - x_i^T \hat{\beta}^{GSL}) x_{i(l)} = \lambda \omega_l \hat{\beta}^{GSL}_{(l)} / \|\hat{\beta}^{GSL}_{(l)}\|_2, & \text{if } \hat{\beta}^{GSL}_{(l)} \neq 0, \\ |\sqrt{n} \sum_{i=1}^{n} (\|\mathbb{Y} - \mathbb{X}\hat{\beta}^{GSL}\|_2)^{-1} (y_i - x_i^T \hat{\beta}^{GSL}) x_{i(l)}| \leq \lambda \omega_l, & \text{if } \hat{\beta}^{GSL}_{(l)} = 0. \end{cases} \tag{7}$$

Compared with the KKT conditions for Group square-root Lasso and Group Lasso, the Group square-root Lasso adds the weighted function $(\sqrt{n}\|\mathbb{Y} - \mathbb{X}\hat{\beta}^{GSL}\|_2)^{-1}$ to estimate the homogeneous noise variance, which allows the tuning parameter λ to be independent of the homogeneous noise variance. Thus, the Group square-root Lasso is able to estimate for the grouped variables and influence the choice of the tuning parameter simultaneously.

A drawback of Group square-root Lasso is that it can only directly select the tuning parameter in linear regression models. However, in logistic regression models, there is no direct way to select the tuning parameter. The penalized weighted score function method uses this scheme. We will discuss this in more detail in the next section.

3. Statistical Properties

In this section, we will establish non-asymptotic oracle inequalities for the penalized weighted score function estimate and present a direct choice for tuning parameter.

Throughout this paper, we consider a fixed design setting (i.e., $x_1, \cdots, x_n$ are considered as deterministic), and we make the following assumptions:

(A1) There exists a positive constant $\mathcal{M} < \infty$ such that $\max_{1 \leq i \leq n} \max_{1 \leq l \leq g} \sqrt{\sum_{j \in G_l} x_{ij}^2} \leq \mathcal{M}$.

(A2) n, p satisfy that $n \leq p = o(e^{n^{1/3}})$, and $p/\epsilon > 2$ for $\epsilon \in (0, 1)$.

(A3) There exists $\mathcal{N}(\beta^0) > 0$ such that

$$\mathcal{N}^2(\beta^0) = \max_{1 \leq j \leq p} \left\{ \frac{1}{n} \sum_{1 \leq i \leq n} \psi^2(x_i^T \beta^0) G(x_i^T \beta^0)(1 - G(x_i^T \beta^0)) x_{ij}^2 \right\}.$$

(A4) Let $\ell_\psi(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}$ be a convex three-times differentiable function such that for all $u, v \in \mathbb{R}^P$, the function $g(t) = \ell_\psi(u + tv)$ satisfies $|g'''(t)| \leq \tau_0 \max_{1 \leq i \leq n} |x_i^T v| g''(t)$ for all $t \in \mathbb{R}$, where $\tau_0 > 0$ is a constant.

Assumption (A1) strictly controls the bounds of predictors, since the real data we collected were often bounded. Assumption (A2) controls the sparsity of the data and the lower bound on the probability that the non-asymptotic property holds. Assumption (A3) makes sure the variance of each component of $\nabla \ell_\psi(\beta^0)$ is bounded by choosing a suitable weighted function $\psi(\cdot)$. Assumption (A4) is similar to Proposition 1 proposed by Bach [28].

Under Assumption (A4), we can obtain lower and upper Taylor expansions of the loss function $\ell_\psi(\cdot)$, which can be used to derive non-asymptotic results.

Moreover, the restricted eigenvalue condition plays a key role in deriving oracle inequalities. For the Group Lasso problem of high-dimensional linear regression models, the oracle property under the group restricted eigenvalue condition was discussed by Hu et al. [29] and extended to logistic regression models by Zhang et al. [18]. To establish the desired group restricted eigenvalue condition, we introduce the following group restricted set

$$\Theta_\alpha =: \Big\{\vartheta \in \mathbb{R}^p \ : \ \|W_{I^C}\vartheta_{(I^C)}\|_{2,1} \le \alpha\|W_I\vartheta_{(I)}\|_{2,1},\ \alpha > 0\Big\}, \tag{8}$$

which is a grouped version of the restricted set $\theta_\alpha =: \{\vartheta \in \mathbb{R}^p \ : \ \|\vartheta_{I^C}\|_1 \le \alpha\|\vartheta_I\|_1\}$ mentioned in Bickel et al. [30], where W_I is a diagonal matrix with the jth diagonal element ω_j if $j \in I$ and 0 otherwise. Based on the group restricted set (8), we propose the following group restricted eigenvalue condition:

(A5) For some integer s such that $1 < s < g$ and a positive number α, the following condition holds

$$\mu(s,\alpha) \triangleq \min_{\substack{I\subseteq\{1,\cdots,g\}\\ |I|\le s}} \min_{\substack{\delta\neq 0\\ \delta\in\Theta_\alpha}} \frac{(\delta^T\mathcal{H}_\psi(\beta^0)\delta)^{1/2}}{\|W_I\delta_{(I)}\|_{2,2}} > 0, \tag{9}$$

where $\mathcal{H}_\psi(\beta^0)$ is the Hessian matrix for $\ell_\psi(\beta^0)$. In contrast to the restricted eigenvalue condition mentioned in Bickel et al. [30] for linear regression models, the group restricted eigenvalue condition for logistic regression is converted from the L_2 norm to the block norm for the denominator part and from the Gram matrix to the Hessian matrix $\mathcal{H}_\psi(\beta^0)$ for the numerator part of (9).

Remark 1. *The Hessian matrix of $\ell_\psi(\beta)$ is given by*

$$\begin{aligned}\mathcal{H}_\psi(\beta) &= \frac{1}{n}\sum_{i=1}^n\left\{\nabla\psi(x_i^T\beta)\left[\frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} - y_i\right] + \psi(x_i^T\beta)\frac{\exp(x_i^T\beta)}{(1+\exp(x_i^T\beta))^2}\right\}x_ix_i^T\\ &= \frac{1}{n}\sum_{i=1}^n\Big\{\nabla\psi(x_i^T\beta)\Big[G(x_i^T\beta) - y_i\Big] + \psi(x_i^T\beta)G(x_i^T\beta)(1-G(x_i^T\beta))\Big\}x_ix_i^T.\end{aligned}$$

Bach [28] has already shown the Hessian matrix of $\ell(\beta)$ is positive definite on some restricted sets. If the chosen weighted function $\psi(x_i^T\beta)$ makes the loss function $\ell_\psi(\beta)$ satisfy the assumption (A3), $\mathcal{H}_\psi(\beta)$ is also positive definite on the group restricted set (8). Such weighted functions in fact exist and will be described later. In addition, the group restricted eigenvalue condition can effectively control the estimation error, enabling estimations with good statistical properties and reliable results.

Theorem 1. *Assume that (A1)–(A4) are satisfied. Let $\lambda < \frac{k(1-z)\mu(s,\alpha)}{4\tau_0\mathcal{M}s}, z \in (0,1)$ and $k < \min_{1\le l\le g}\omega_l$. Let λ be a tuning parameter chosen such that*

$$\lambda\omega_l = \frac{\mathcal{N}(\beta^0)}{z}\sqrt{\frac{|G_l|}{n}}\Phi^{-1}(1-\frac{\epsilon}{2p}). \tag{10}$$

Then, with probability of at least $1-\epsilon(1+o(1))$, we have the following:
1. A group restricted set $\hat\beta - \beta^0 \in \Theta_\alpha$ with $\alpha = \frac{1+z}{1-z}$.
2. Under the group restricted eigenvalue condition (A5), the block norm estimation errors are

$$\|\hat\beta - \beta^0\|_{2,1} \le \frac{2k\lambda s}{(\min\limits_{1\le l\le g}\omega_l - k)(1-z)\mu(s,\alpha)}, \tag{11}$$

$$\|\hat{\beta}-\beta^0\|_{2,q}^q \leq \left(\frac{2k\lambda s}{(\min\limits_{1\leq l\leq g}\omega_l - k)(1-z)\mu(s,\alpha)}\right)^q, \quad \text{for all } 1<q<2, \tag{12}$$

respectively, and the error of the loss function ℓ_ψ *is*

$$|\ell_\psi(\hat{\beta})-\ell_\psi(\beta^0)| \leq \frac{2\min\limits_{1\leq l\leq g}\omega_l\lambda^2 s}{(\min\limits_{1\leq l\leq g}\omega_l - k)(1-z)\mu(s,\alpha)}. \tag{13}$$

The non-asymptotic oracle inequalities for the true coefficient β^0 are provided in (11) and (12). Unfortunately, the parameter $\mathcal{N}(\beta^0)$ is influenced by the true coefficient β^0, so that the choice of λ also depends on β^0. Therefore, we will choose a suitable $\psi(x_i^T\beta^0)$ to solve this problem in the next theorem.

Theorem 2. *Choose the weight function in the following form*

$$\psi(x_i^T\beta^0) = \frac{1}{2}\left(\exp(\frac{x_i^T\beta^0}{2}) + \exp(-\frac{x_i^T\beta^0}{2})\right). \tag{14}$$

Under Assumptions (A2) and (A3), we choose the tuning parameter as

$$\lambda\omega_l = \frac{\sqrt{|G_l|\max\limits_{1\leq j\leq p}(\sum\limits_{i=1}^n x_{ij}^2)}}{2nz}\Phi^{-1}(1-\frac{\epsilon}{2p}). \tag{15}$$

Then, under the assumptions of Theorem 1 with the probability at least $1-\epsilon(1+o(1))$*, we have inequalities (11)–(13).*

In Theorem 2, Yin [11] presents a discussion about the order of $\Phi^{-1}(1-\frac{\epsilon}{2p})$ in (15), proving that $\Phi^{-1}(1-\frac{\epsilon}{2p}) \sim \mathcal{O}(\sqrt{\log(2p/\epsilon)})$. When $|G_l| = 1$ for $l = 1, 2, \cdots, g$, our estimate $\hat{\beta}$ is a Lasso estimate and its theoretical properties have been well studied by Yin [11].

Remark 2. *If* $\psi(x_i^T\beta^0)$ *is given as in Theorem 2, the loss function, weighted score function and the Hessian matrix, respectively, are given by*

$$\begin{cases} \ell_\psi(\beta^0) = \frac{1}{n}\sum\limits_{i=1}^n\left\{(1-y_i)\exp(\frac{x_i^T\beta^0}{2}) + y_i\exp(-\frac{x_i^T\beta^0}{2})\right\}, \\ \nabla\ell_\psi(\beta^0) = \frac{1}{2n}\sum\limits_{i=1}^n\left\{(1-y_i)\exp(\frac{x_i^T\beta^0}{2}) - y_i\exp(-\frac{x_i^T\beta^0}{2})\right\}x_i, \\ \mathcal{H}_\psi(\beta^0) = \frac{1}{4n}\sum\limits_{i=1}^n\left\{(1-y_i)\exp(\frac{x_i^T\beta^0}{2}) + y_i\exp(-\frac{x_i^T\beta^0}{2})\right\}x_ix_i^T. \end{cases}$$

Clearly, the Hessian matrix given as a weighting function in the form in Theorem 2 is positive definite.

4. Weighted Block Coordinate Descent Algorithm

We apply the techniques of the block coordinate descent algorithm to the penalized weighted score function. Choose the weighted function with the form of (14) and set $\beta = \hat{\beta} + \zeta$; then, a second-order Taylor expansion of the loss function $\ell_\psi(\beta)$ in Equation (6) gives

$$\mathcal{D}(\hat{\beta}+\zeta) = \left\{\left(\ell_\psi(\hat{\beta}) + \zeta^T\nabla\ell_\psi(\hat{\beta}) + \frac{1}{2}\zeta^T\mathcal{H}_\psi(\hat{\beta})\zeta\right) + \lambda\|W(\hat{\beta}+\zeta)\|_{2,1}\right\}, \tag{16}$$

Now, we consider minimization $\mathcal{D}(\hat{\beta}+\zeta)$ with respect to the lth group of penalized parameters. This means that

$$\nabla \ell_\psi(\hat{\beta})_{(l)} + \mathcal{H}_\psi(\hat{\beta})_{(l)}\zeta_{(l)} + \lambda\omega_l \frac{\hat{\beta}_{(l)}+\zeta_{(l)}}{\|\hat{\beta}_{(l)}+\zeta_{(l)}\|_2} = 0. \tag{17}$$

Inspired by Meier et al.'s [13] assumptions, we set the sub-matrix $\mathcal{H}_\psi(\hat{\beta})_{(l)}$ in the form of $\mathcal{H}_\psi(\hat{\beta})_{(l)} = h_\psi(\hat{\beta})_{(l)} I_{(l)}$, which means that $h_\psi(\hat{\beta})_{(l)} = -\max\{\mathrm{diag}(-\mathcal{H}_\psi(\hat{\beta})_{(l)}), r_0\}$, where r_0 is a lower bound to ensure convergence. Then, simplifying Equation (17) gives

$$\left(\frac{\lambda\omega_l}{\|\hat{\beta}_{(l)}+\zeta_{(l)}\|_2} + h_\psi(\hat{\beta})_{(l)}\right)(\hat{\beta}_{(l)}+\zeta_{(l)}) = h_\psi(\hat{\beta})_{(l)}\hat{\beta}_{(l)} - \nabla\ell_\psi(\hat{\beta})_{(l)}.$$

This leads to the following equivalence equation

$$\frac{\hat{\beta}_{(l)}+\zeta_{(l)}}{\|\hat{\beta}_{(l)}+\zeta_{(l)}\|_2} = \frac{h_\psi(\hat{\beta})_{(l)}\hat{\beta}_{(l)} - \nabla\ell_\psi(\hat{\beta})_{(l)}}{\|h_\psi(\hat{\beta})_{(l)}\hat{\beta}_{(l)} - \nabla\ell_\psi(\hat{\beta})_{(l)}\|_2}. \tag{18}$$

According to Equation (15) and Remark 2, it is obtained that:
If $\|h_\psi(\hat{\beta})_{(l)}\hat{\beta}_{(l)} - \nabla\ell_\psi(\hat{\beta})_{(l)}\|_2 \le \lambda\omega_l$, the value of ζ at the k-th iteration is given by

$$\zeta_{(l)}^{(k)} = -\hat{\beta}_{(l)}^{(k)},$$

otherwise

$$\zeta_{(l)}^{(k)} = -\frac{1}{h_\psi(\hat{\beta}^{(k)})_{(l)}}\left(\nabla\ell_\psi(\hat{\beta}^{(k)})_{(l)} + \lambda\omega_l \frac{h_\psi(\hat{\beta}^{(k)})_{(l)}\hat{\beta}_{(l)}^{(k)} - \nabla\ell_\psi(\hat{\beta}^{(k)})_{(l)}}{\|h_\psi(\hat{\beta}^{(k)})_{(l)}\hat{\beta}_{(l)}^{(k)} - \nabla\ell_\psi(\hat{\beta}^{(k)})_{(l)}\|_2}\right).$$

where $\lambda\omega_l = \sqrt{|G_l|\max_{1\le j\le p}(\sum_{i=1}^n x_{ij}^2)}\Phi^{-1}(1-\frac{\epsilon}{2p})/2nz$. If $\zeta_{(l)}^{(k)} \neq 0$, we use the Armijo rule of Tseng and Yun [31] to select the step factor $\sigma^{(k)}$ as follows:
Armijo rule

Choose $\sigma_0 > 0$ and let $\sigma^{(k)}$ be the largest value of $\{\sigma_0\theta^j\}_{j\ge 0}$ satisfying

$$\mathcal{D}(\hat{\beta}_{(l)}^{(k)} + \sigma^{(k)}\zeta_{(l)}^{(k)}) - \mathcal{D}(\hat{\beta}_{(l)}^{(k)}) \le \sigma^{(k)}\varrho\triangle_l^{(k)},$$

where $0<\theta<1$, $0<\varrho<1$, and

$$\triangle_l^{(k)} = -{\zeta_{(l)}^{(k)}}^T\nabla\ell_\psi(\hat{\beta}^{(k)})_{(l)} + \lambda\omega_l\|\hat{\beta}_{(l)}^{(k)}+\zeta_{(l)}^{(k)}\|_2 - \lambda\omega_l\|\hat{\beta}_{(l)}^{(k)}\|_2.$$

Finally, the update direction is calculated for the gradient of the parameters and the parameters are updated according to a certain step size

$$\hat{\beta}_{(l)}^{(k+1)} = \hat{\beta}_{(l)}^{(k)} + \sigma^{(k)}\zeta_{(l)}^{(k)}.$$

The weighted block coordinate gradient descent algorithm is given by Algorithm 1. An initial parameter setting of $\sigma_0 = 1, \theta = 0.5$ and $\varrho = 0.1$ was given by Tseng and Yun [31]. In the next simulations, we set the convergence criterion of step 3 in Algorithm 1 to be $\sigma^{(k)} \le 10^{-10}$. In general, selecting the tuning parameter λ using the cross-validation method is complicated. As we know from Algorithm 1, the algorithm eliminates the selection process for the tuning parameter $\lambda\omega_l$. Given an initial value $\hat{\beta}^{(0)}$, we can then iterate directly over $\hat{\beta}^{(0)}$ until it converges to the range which we expect.

Algorithm 1 Weighted block coordinate gradient descent algorithm

Step 1: Let $\hat{\beta}^{(0)} \in \mathbb{R}^p$ be an initial parameter vector
Step 2: For $l = 1, \cdots, g$
$\mathcal{H}_\psi(\hat{\beta}^{(k)})_{(l)} = h_\psi(\hat{\beta}^{(k)})_{(l)} I_{(l)}$,
$\zeta^{(k)} = \arg\min_{\zeta \in \mathbb{R}^p}\{\mathcal{D}(\hat{\beta}^{(k)} + \zeta)\}$,
if $\zeta^{(k)} = 0$
$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)}$,
else
Search $\sigma^{(k)}$ using Armijo rule,
$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \sigma^{(k)}\zeta^{(k)}$,
end
Step 3: Repeat step 2 until some convergence criterion is met

It is worth noting that we have given a direct choice (15) for λ under a specific weight function $\psi(x_i^T\beta^0)$ given by (14), so the weighted block coordinate gradient descent algorithm will be computationally faster than working iteratively on a fixed grid of tuning parameters λ (see Meier et al. [13]). If choosing other weight functions, the weighted block coordinate gradient descent algorithm can still be used to solve (6). However, then the tuning parameter λ depends on β^0 (unknown); some cross-validation can be used for choosing the parameter λ.

5. Simulations

In this section, we use simulated datasets to evaluate the performance of the penalized weighted score function estimator. Meier [13] describes the block coordinate gradient descent algorithm using the R package R 4.3.1 **grplasso** (https://cran.r-project.org/web/packages/grplasso/grplasso.pdf, accessed on 6 July 2023). While the **grplasso** algorithm offers 20 predefined values of the tuning parameter λ, it lacks an optimal design for λ. We improved **grplasso** by providing a scheme for directly selecting the tuning parameters, named **wgrplasso**, and we use it to describe the weighted block coordinate gradient descent algorithm. We compare the performance of the **wgrplasso** algorithm, the R package **grpreg** (https://cran.r-project.org/web/packages/grpreg/grpreg.pdf, accessed on 6 July 2023) developed by Breheny [20] and the R package **gglasso** (https://cran.r-project.org/web/packages/gglasso/gglasso.pdf, accessed on 6 July 2023) developed by Yang and Zou [32]. Three main aspects of model performance are considered: the correctness of variable selection, the accuracy of coefficient estimation and the running time of the algorithm. The evaluation indicators for the model include the following:

- *TP*: the number of predicted non-zero values in the non-zero coefficient set when determining the model.
- *TN*: the number of predicted zero values in the zero coefficient set when determining the model.
- *FP*: the number of predicted non-zero values in the zero coefficient set when determining the model.
- *FN*: the number of predicted zero values in the non-zero coefficient set when determining the model.
- *TPR*: the ratio of predicted non-zero values in the non-zero coefficient set when determining the model, which is calculated by the following formula:

$$TPR = \frac{TP}{TP + FN}.$$

- *Accur*: the ratio of accurate predictions when determining the model, which is calculated by the following formula:

$$Accur = \frac{TP + TN}{TP + TN + FP + FN}.$$

- *Time*: the running time of the algorithm.
- *BNE*: the block norm of the estimation error, which is calculated by the following formula:

$$BNE = \|\hat{\beta} - \beta\|_{2,1}.$$

The sample size was 200. We set values of $p = 300$, 600 and 900, and generated 500 random datasets to repeat the simulation. We set ϵ to 0.01 and 0.05 and uniformly specified the true non-zero coefficient parameters of the logistic regression models as

$$\beta = (1, \underbrace{\underbrace{1, \cdots, 1}_{3}, \cdots, \underbrace{1, \cdots, 1}_{3}}_{30}, \underbrace{0, \cdots, 0}_{p-30}).$$

For the log odd η setting, we considered the following four different models.

(a) In Model I, the observed data X are assumed to be sampled from a multivariate normal distribution and the log odd η is considered to be the linear case, where the data between groups are independent but the data within groups are correlated. We set the size of each group to 3 and assume that the data within the groups obey $X_i \sim N(0, \Sigma_{i,jk})$, where $\Sigma_i = 0.5^{|j-k|}$. Thus, the observed data can then be defined as $X \sim N(0, \Sigma)$, where $\Sigma = diag(\Sigma_1, \cdots, \Sigma_{\frac{p}{3}})$.

(b) In Model II, the observed data X are assumed to be the sum of two uniform distributions and the log odd η is considered to be the linear case. Assume that the p-dimensional vectors $Z_1, \cdots, Z_p$ and W are generated independently and through a uniform distribution of $[-1, 1]$. Thus, the observed data can be defined as $X_i = Z_i + W$.

The log odds η for Models I and II are then defined as follows

$$\eta = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p.$$

(c) In Model III, the observed data X are assumed to follow a standard multivariate normal distribution and the log ogg η is considered to be additive case. Assuming that X obeys the $\frac{p}{3}$-dimensional standard normal distribution, the observed data can therefore be defined as $X \sim N(0, I_{\frac{p}{3}})$.

(d) In Model IV, the observed data X are assumed to be the sum of two uniform distributions and the log odd η is considered to be the additive case. This means that the $\frac{p}{3}$-dimensional vectors $Z_1, \cdots, Z_{\frac{p}{3}}$ and W are assumed to be generated independently by a uniform distribution of $[-1, 1]$. Thus, the observed data can be defined as $X_i = Z_i + W$.

The log odds η for Models III and IV are then defined as follows

$$\eta = \beta_0 + X_1\beta_1 + X_1^2\beta_2 + X_1^3\beta_3 + \cdots + X_{\frac{p}{3}}\beta_{p-2} + X_{\frac{p}{3}}^2\beta_{p-1} + X_{\frac{p}{3}}^3\beta_p.$$

Then, the dataset for the response variable Y was generated by the logistic regression models

$$\mathbb{P}(Y = 1|\eta) = \frac{1}{1 + \exp(\eta^{-1})}.$$

Table 1 shows the average simulation results of the three algorithms for the linear case, and Figure 1 shows the point–line plots of Model I and Model II for TPR, Accur, Time and MSE.

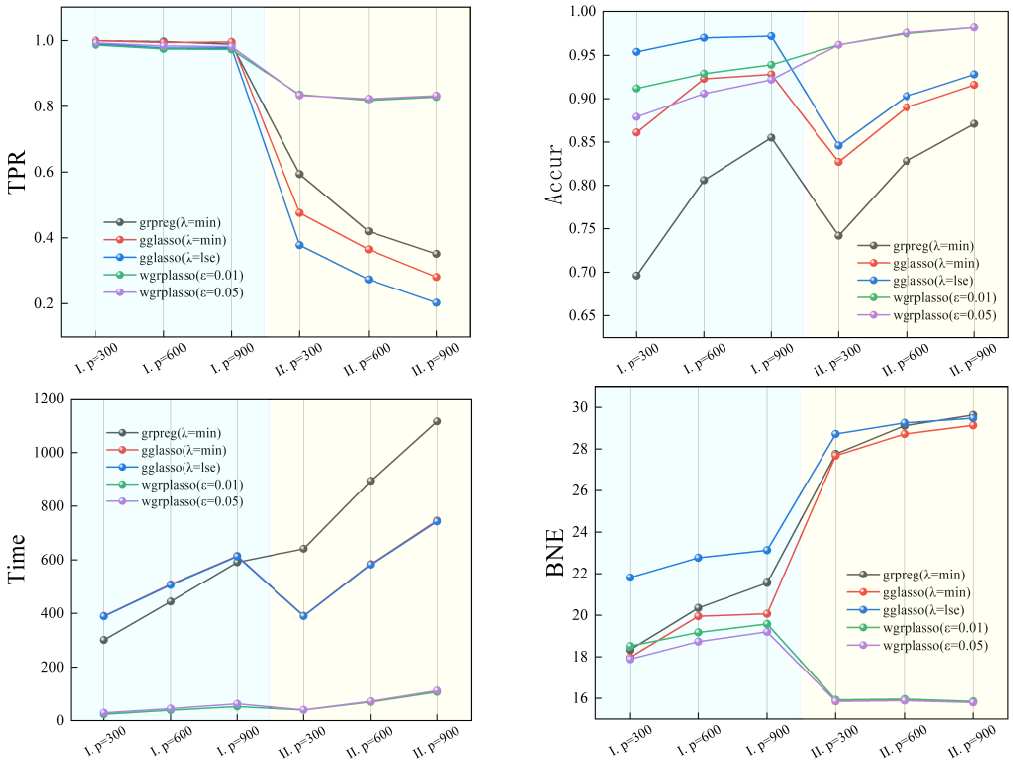

Figure 1. Average TPR, Accur, Time and BNE plots for 500 repetitions of the three algorithms in Model I and Model II.

First, from the TPR perspective, all three algorithms show excellent selection results when the normal distribution assumption is adopted. However, when the uniform distribution assumption is used, the **wgrplasso** algorithm shows higher correct selection in the nonzero set than the other algorithms, and the **wgrplasso** algorithm is also more stable in terms of variance.

Second, from the Accur perspective, compared to the **gepreg** algorithm, the **wgrplasso** and **gglasso** algorithms maintain a high selection effect under the assumption of a normal distribution. However, Accur is also affected by FP, and the **gepreg** algorithm and **gglasso** algorithm are not stable enough to control FP from the perspective of variance. In addition, under the assumption of a uniform distribution, both in terms of the effect of selection and the stability of variance, the **wgrplasso** algorithm has lower control over the FP aspect, which makes the **wgrplasso** algorithm perform better than the other algorithms in terms of Accur.

Third, from a Time perspective, using the **wgrplasso** algorithm saves a lot of time, both for the normal distribution assumption and the uniform distribution assumption.

Furthermore, lastly, from a BNE perspective, under the assumption of normal distribution, the BNE values obtained by the **wgrplasso** and **gglasso** algorithms are similar and smaller than that obtained by the **grpreg** algorithm. However, under the assumption of a uniform distribution, compared with the **gglasso** algorithm and the **grpreg** algorithm, the BNE obtained by the **wgrplasso** algorithm is smaller, which means that the **wgrplasso** algorithm performs better.

Table 2 presents the simulation results of the three algorithms for the additive case, and Figure 2 shows the point–line plots of Models III and IV for TPR, Accur, Time and BNE.

Table 1. Average results for 500 repetitions of the three algorithms in Models I and II.

		Model I					
		TP	**TPR**	**FP**	**Accur**	**Time**	**BNE**
$p = 300$	grpreg(λ = min)	30.00 (0.00)	1.000	91.28 (19.46)	0.696	300.63	18.32 (1.96)
	gglasso(λ = min)	30.00 (0.00)	1.000	41.64 (29.92)	0.861	390.56	17.96 (3.11)
	gglasso(λ = lse)	29.68 (1.10)	0.990	13.44 (14.73)	0.954	389.27	21.81 (2.29)
	wgrplasso(ϵ = 0.01)	29.61 (1.06)	0.987	26.15 (7.92)	0.912	23.53	18.51 (0.65)
	wgrplasso(ϵ = 0.05)	29.77 (0.85)	0.993	36.14 (9.80)	0.879	29.24	17.88 (0.70)
$p = 600$	grpreg(λ = min)	29.90 (0.55)	0.997	116.36 (26.51)	0.806	444.31	20.35 (1.73)
	gglasso(λ = min)	29.80 (0.91)	0.994	45.85 (34.78)	0.923	508.35	19.95 (2.41)
	gglasso(λ = lse)	29.32 (2.00)	0.978	17.37 (16.92)	0.970	506.27	22.77 (1.81)
	wgrplasso(ϵ = 0.01)	29.25 (1.40)	0.975	41.84 (11.33)	0.929	38.97	19.17 (0.71)
	wgrplasso(ϵ = 0.05)	29.50 (1.19)	0.984	55.81 (12.78)	0.906	45.16	18.73 (0.76)
$p = 900$	grpreg(λ = min)	29.66 (1.13)	0.989	130.12 (32.66)	0.855	590.55	21.56 (1.82)
	gglasso(λ = min)	29.88 (0.59)	0.996	64.84 (39.83)	0.928	614.64	20.07 (2.24)
	gglasso(λ = lse)	29.30 (1.53)	0.977	24.07 (21.79)	0.972	612.24	23.13 (1.80)
	wgrplasso(ϵ = 0.01)	29.19 (1.43)	0.973	54.10 (15.45)	0.939	52.63	19.58 (0.73)
	wgrplasso(ϵ = 0.05)	29.44 (1.21)	0.982	70.01 (15.98)	0.922	62.81	19.20 (0.78)
		Model II					
		TP	**TPR**	**FP**	**Accur**	**Time**	**BNE**
$p = 300$	grpreg(λ = min)	17.82 (4.36)	0.594	65.31 (10.55)	0.742	641.23	27.77 (1.32)
	gglasso(λ = min)	14.30 (4.92)	0.476	36.25 (10.33)	0.827	391.28	27.69 (1.43)
	gglasso(λ = lse)	11.36 (4.80)	0.378	27.70 (11.50)	0.846	389.83	28.73 (0.96)
	wgrplasso(ϵ = 0.01)	25.07 (2.67)	0.836	6.52 (4.83)	0.962	39.71	15.92 (1.09)
	wgrplasso(ϵ = 0.05)	25.02 (2.68)	0.834	6.28 (4.70)	0.962	40.24	15.85 (1.09)
$p = 600$	grpreg(λ = min)	12.61 (4.32)	0.420	85.84 (11.35)	0.828	894.47	29.13 (1.17)
	gglasso(λ = min)	10.95 (4.99)	0.365	47.08 (13.41)	0.890	584.32	28.73 (1.04)
	gglasso(λ = lse)	8.23 (4.76)	0.274	36.33 (13.85)	0.903	581.74	29.26 (0.72)
	wgrplasso(ϵ = 0.01)	24.57 (2.81)	0.819	9.43 (6.08)	0.975	69.48	15.96 (0.96)
	wgrplasso(ϵ = 0.05)	24.69 (2.80)	0.823	9.23 (6.26)	0.976	72.05	15.89 (0.99)
$p = 900$	grpreg(λ = min)	10.53 (4.60)	0.351	96.88 (12.79)	0.871	1115.73	29.64 (1.07)
	gglasso(λ = min)	8.43 (4.49)	0.281	53.67 (13.97)	0.916	746.62	29.14 (0.93)
	gglasso(λ = lse)	6.09 (4.20)	0.203	40.74 (15.09)	0.928	742.62	29.49 (0.58)
	wgrplasso(ϵ = 0.01)	24.86 (2.66)	0.829	10.80 (6.39)	0.982	106.940	15.85 (1.01)
	wgrplasso(ϵ = 0.05)	24.99 (2.71)	0.833	11.05 (6.23)	0.982	111.95	15.80 (1.00)

Reported numbers are the averages and standard errors (show in parentheses).

Table 2. Average results for 500 repetitions of the three algorithms in Models III and IV.

		Model III					
		TP	**TPR**	**FP**	**Accur**	**Time**	**BNE**
$p = 300$	grpreg(λ = min)	29.39 (1.79)	0.980	73.59 (21.16)	0.753	447.46	27.52 (1.96)
	gglasso(λ = min)	29.91 (0.59)	0.997	74.11 (25.60)	0.753	812.03	24.06 (2.05)
	gglasso(λ = lse)	29.57 (2.32)	0.986	40.58 (21.48)	0.863	807.65	25.27 (1.69)
	wgrplasso($\epsilon = 0.01$)	27.69 (2.51)	0.923	24.02 (7.69)	0.912	35.92	28.99 (1.27)
	wgrplasso($\epsilon = 0.05$)	28.55 (2.06)	0.952	32.00 (8.15)	0.888	39.13	28.84 (1.38)
$p = 600$	grpreg(λ = min)	28.05 (2.96)	0.935	86.76 (28.04)	0.852	598.05	28.65 (1.70)
	gglasso(λ = min)	29.40 (2.37)	0.980	97.53 (36.13)	0.836	974.70	25.44 (1.92)
	gglasso(λ = lse)	27.62 (5.90)	0.920	45.84 (27.29)	0.920	968.57	26.65 (1.87)
	wgrplasso($\epsilon = 0.01$)	27.15 (2.69)	0.905	40.41 (10.68)	0.928	56.35	29.40 (1.22)
	wgrplasso($\epsilon = 0.05$)	28.18 (2.21)	0.940	51.31 (11.66)	0.911	63.67	29.34 (1.33)
$p = 900$	grpreg(λ = min)	25.66 (5.66)	0.856	82.92 (36.76)	0.903	745.82	29.33 (1.51)
	gglasso(λ = min)	28.77 (3.79)	0.959	105.48 (45.77)	0.881	1121.19	26.32 (1.87)
	gglasso(λ = lse)	24.33 (9.47)	0.811	42.12 (35.83)	0.947	1113.45	27.76 (2.14)
	wgrplasso($\epsilon = 0.01$)	26.85 (2.87)	0.895	50.99 (10.80)	0.940	68.74	29.70 (1.18)
	wgrplasso($\epsilon = 0.05$)	27.80 (2.38)	0.926	63.14 (12.27)	0.927	81.32	29.67 (1.27)
		Model IV					
		TP	**TPR**	**FP**	**Accur**	**Time**	**BNE**
$p = 300$	grpreg(λ = min)	21.94 (4.03)	0.732	63.80 (9.64)	0.760	466.73	35.16 (1.78)
	gglasso(λ = min)	19.88 (4.43)	0.662	52.83 (11.36)	0.790	409.92	28.30 (1.13)
	gglasso(λ = lse)	17.30 (4.74)	0.577	47.80 (11.44)	0.798	408.22	28.93 (0.74)
	wgrplasso($\epsilon = 0.01$)	28.75 (1.65)	0.959	25.96 (8.12)	0.909	218.10	26.09 (2.55)
	wgrplasso($\epsilon = 0.05$)	28.78 (1.65)	0.960	26.32 (8.14)	0.908	221.08	26.13 (2.57)
$p = 600$	grpreg(λ = min)	18.32 (4.40)	0.611	83.08 (12.48)	0.842	689.27	35.02 (1.79)
	gglasso(λ = min)	16.48 (5.10)	0.549	70.00 (14.34)	0.861	571.90	29.08 (1.01)
	gglasso(λ = lse)	14.05 (5.17)	0.468	62.39 (14.65)	0.869	567.98	29.37 (0.62)
	wgrplasso($\epsilon = 0.01$)	28.58 (1.80)	0.953	34.33 (10.12)	0.940	384.79	26.59 (2.69)
	wgrplasso($\epsilon = 0.05$)	28.58 (1.83)	0.953	34.76 (10.11)	0.940	380.57	26.63 (2.70)
$p = 900$	grpreg(λ = min)	15.66 (4.25)	0.522	94.71 (12.41)	0.879	356.36	34.90 (1.50)
	gglasso(λ = min)	13.80 (4.61)	0.460	80.03 (13.92)	0.893	289.06	29.45 (0.92)
	gglasso(λ = lse)	11.61 (4.49)	0.387	70.52 (15.85)	0.901	287.73	29.64 (0.54)
	wgrplasso($\epsilon = 0.01$)	28.55 (1.83)	0.952	39.33 (12.57)	0.955	184.13	26.53 (2.34)
	wgrplasso($\epsilon = 0.05$)	28.56 (1.80)	0.952	39.24 (12.45)	0.955	186.89	26.56 (2.36)

Reported numbers are the averages and standard errors (show in parentheses).

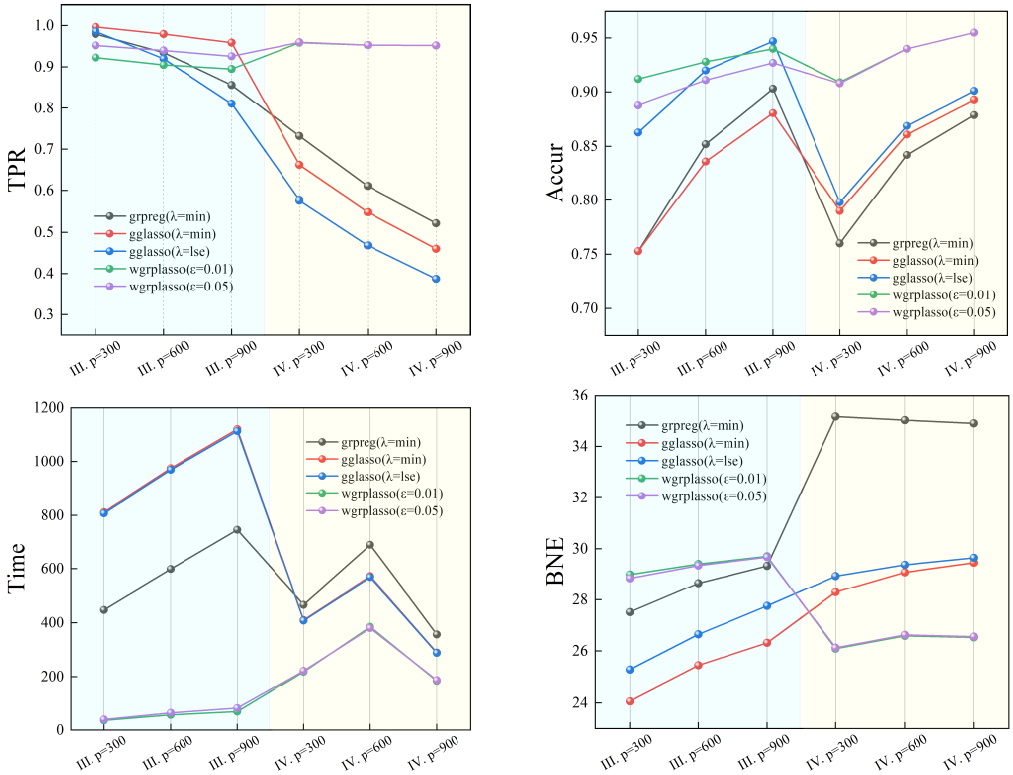

Figure 2. Average TPR, Accur, Time and BNE plots for 500 repetitions of the three algorithms in Model III and Model IV.

The simulation results show that the **grpreg** algorithm and the **gglasso** algorithm in the additive case are poorer both in terms of TPR and Accur, and also show through the variance that the **grpreg** algorithm and the **gglasso** algorithm also do not have a stable selection, as well as increasing computational time overheads and BNE values. However, **wgrplasso** obtains similar results in the additive case as in the linear case, and still maintains a better selection. Regardless of TPR, Accur and BNE, the **wgrplasso** algorithm performs better than the other algorithms, and the advantage in Time is even more obvious.

6. Real Data

In this section, we apply our proposed estimates to analyze two real data sets. The first data set comes from the molecular shape and conformation of musk. The second data set comes from histologically normal epithelial cells from breast cancer patients and cancer-free prophylactic mastectomy patients. As in the previous section, we set ϵ to 0.01 and 0.05, respectively. In Section 6.1, we compare the number of variables selected and the computation time of the three algorithms in the above simulation, and in Section 6.2, we compare the prediction accuracy and the computation time.

6.1. Studies on the Molecular Structure of Muscadine

The R package of **kernlab** (https://cran.r-project.org/web/packages/kernlab/kernlab.pdf, accessed on 12 July 2023) contains the molecular shape and conformation of musk in the native dataset musk. The data set contains a data frame of 476 observations for the following 167 variables. The first 162 of these variables are the distance characteristics of the rays, measured relative to the origin along which each ray was placed. Any experiment with the data should treat these features as being on any continuous scale. Variable 163 is the distance of the oxygen atom to a specified point in 3D space. Variable 164 is the x-displacement from the specified point. Variable 165 is the Y-displacement from the specified point. Variable 166 is the Z displacement from the specified point. Variable 167 has a value of 0 for no musk or 1 for musk.

We used 3/4 of the data for training and performed a third-order B-spline basis function expansion on the training data, and then we used the **wgrplasso**, **grpreg**, **gglasso**, and **glmnet** (https://cran.r-project.org/web/packages/glmnet/glmnet.pdf, accessed on 12 July 2023) algorithms for estimations using the expanded training data, respectively. The remaining 1/4 of the data were used as a test, and the estimated coefficients were used to predict the test data, comparing the prediction accuracy, model size and time for each of the four algorithms. Table 3 presents the experimental results of 100 repetitions.

Table 3. Average prediction accuracy, model size and time taken for 100 repetitions of the four algorithms in the musk dataset.

	wgrplasso (ϵ = **0.05**)	**grpreg** (λ = **min**)	**gglasso** (λ = **min**)	**glmnet** (λ = **min**)
Prediction accuracy	0.820	0.813	0.771	0.758
Model size	66.53	31.29	30.14	53.53
Time	0.69	3.04	2.70	2.12

The experimental results show that wgrplasso has the highest prediction accuracy among the four algorithms, indicating that the algorithm is able to identify the target class more accurately in the task of categorizing musk data, and wgrplasso also exhibits a shorter computation time without sacrificing accuracy. This makes the wgrplasso algorithm the preferred algorithm for dealing with the problem of categorizing musk datasets.

6.2. Gene Expression Studies in Epithelial Cells of Breast Cancer Patients

We obtained microarray data from the NCBI Gene Expression Omnibus for patient histological epithelial cells (https://www.ncbi.nlm.nih.gov/geo/, accessed on 31 August 2023) under accession GSE20437. The dataset consists of 42 samples with 22,283 variables. It consists of microarray gene expression data collected from the histologically normal epithelium (NlEpi) from 18 breast cancer patients (HN), 18 patients undergoing breast reduction (RM) and 6 cancer-free prophylactic mastectomy (PM) patients in high-risk women. Graham et al. [33] have shown that genes are differentially expressed between HM and RM samples. This is more fully discussed in Yang and Zou [32]. Here, we consider the effect of genes on HM and RM. Similar to Yang and Zou's [32] approach to the data, we fit the sparse additive logistic regression model using the Group Lasso penalty while selecting the significant additive components.

As with the setup in Section 6.1, we continue to train with 3/4 of the data and expand the training data using a third-order B-spline basis function and treated them as a group to reflect the role in the additive models, leading to a grouped regression problem with $n = 36$ and $p = 66849$. All data were then standardized so that the mean of each original variable was zero and the sample variance was in units. This experiment was repeated 100 times to obtain the prediction error. We built a complete observational model for one of experiments, and report the selected genes in **wgrplasso**, **grpreg** and **gglasso** algorithms. These results are listed in Table 4. We observe that the **wgrplasso** and **gglasso** algorithms select more variables than the **grpreg** algorithm, and **wgrplasso** has lower prediction errors. Summarizing the above results, our proposed penalized weighted score function method can pick much more meaningful variables for explanation and prediction.

Table 4. Average prediction error and model size for selected genes for 100 repetitions of three algorithms in microarray gene expression data from histological epithelial cells.

	wgrplasso (ϵ= 0.05)	**grpreg (λ = min)**	**gglasso (λ = min)**
Prediction Accuracy	0.73	0.63	0.71
Model Size	14	9	14
Selected genes	117_at 1255_g_at 200000_s_at 200002_at 200030_s_at 200040_at 200041_s_at 200655_s_at 200661_at 200729_s_at 201040_at 201465_s_at 202707_at 211997_x_at	201464_x_at 201465_s_at 201778_s_at 202707_at 204620_s_at 205544_s_at 211997_x_at 213280_at 217921_at	200047_s_at 200729_s_at 200801_x_at 201465_s_at 202046_s_at 202707_at 205544_s_at 208443_x_at 211374_x_at 211997_x_at 212234_at 213280_at 217921_at 220811_at

7. Conclusions

In our work, we propose the penalized weighted score function method for Group Lasso for logistic regression models. We determine an upper bound of the error of parameter estimation with a high probability and the direct choice of the tuning parameter under a specific weighted function. Under the direct choice of the tuning parameter, we improve the block coordinate descent algorithm to reduce the computational time and complexity. Simulation results show that our method not only exhibits better statistical accuracy, but also calculates faster than competing methods. Experimental results with real data also show that our method is effective in other fields such as biology and chemistry. Indeed, our approach can be extended to other generalized linear models with a sparse group structure, which will be future research.

Author Contributions: Conceptualization, Z.Y.; Methodology, M.Z., Z.Y. and Z.W.; Software, M.Z.; Data curation, Z.W.; Writing—original draft, M.Z. and Z.Y.; Writing—review & editing, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The authors' work was supported by the Educational Commission of Jiangxi Province of China (No.GJJ160927) and the National Natural Science Foundation of China (No.62266002).

Data Availability Statement: All data available in the paper with its related references.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Lemma A1 (Bach [28])**.** *Consider a three-times differentiable convex function $g : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$, $|g'''(t)| \leq Sg''(t)$, for some $S \geq 0$. Then, for all $t \geq 0$:*

$$\frac{g''(0)}{S^2}(\exp(-St) + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g''(0)}{S^2}(\exp(St) - St - 1).$$

Lemma A2 (Hu et al. [29])**.** *If the inequality $\sum_{i=1}^{n} a_i \leq b_0$ holds for all $a_i > 0$, we have $\sum_{i=1}^{n} a_i^q \leq b_0^q$ for $1 < q < 2$.*

Proof of Lemma A2. We first introduce the Holder inequality:

Set $m, n > 1$ and $\frac{1}{m} + \frac{1}{n} = 1$. Let a_i and b_i be non-negative real numbers, then

$$\sum_{i=1}^{n} a_i b_i \leq \left(\sum_{i=1}^{n} a_i^m \right)^{\frac{1}{m}} \left(\sum_{i=1}^{n} b_i^n \right)^{\frac{1}{n}}.$$

According to the Holder inequality and setting $m = \frac{1}{2-q}$ and $n = \frac{1}{q-1}$, we have

$$\begin{aligned} \sum_{i=1}^{n} a_i^q &= \sum_{i=1}^{n} \left(a_i^{2-q} a_i^{2q-2} \right) \\ &\leq \left(\sum_{i=1}^{n} a_i \right)^{2-q} \left(\sum_{i=1}^{n} a_i^2 \right)^{q-1}, \end{aligned}$$

because $\sum_{i=1}^{n} a_i^2 \leq (\sum_{i=1}^{n} a_i)^2 \leq b_0^2$. Then,

$$\sum_{i=1}^{n} a_i^q \leq b_0^{2-q} \left(b_0^2 \right)^{q-1} = b_0^q,$$

where $m, n > 1$, which means $q \in (1, 2)$. □

Lemma A3 (Sakhanenko [34]). *Let $\mathcal{F}_1, \cdots, \mathcal{F}_n$ be independent random variables with $\mathbb{E}(\mathcal{F}_i) = 0$ and $|\mathcal{F}_i| < 1$ for all $1 \leq i \leq n$. Denote $B_n^2 = \sum_{i=1}^{n} \mathbb{E}(\mathcal{F}_i^2)$ and $L_n = \sum_{i=1}^{n} \mathbb{E}(|\mathcal{F}_i|^3)/B_n^3$. Then, there exists a positive constant R such that for all $x \in [1, \frac{1}{R} \min\{B_n, L_n^{-1/3}\}]$*

$$\mathbb{P}\left(\sum_{i=1}^{n} \mathcal{F}_i > B_n x \right) = (1 + O(1)x^3 L_n)(1 - \Phi(x)).$$

Proof of Theorem 1. Define the event

$$A = \left\{ \max_{1 \leq l \leq g} \sqrt{\sum_{j \in G_l} \nabla \ell_\psi^2(\beta_j^0)} \leq z\lambda\omega_l \right\}.$$

We state the theorem result on the event A and find an lower bound of $\mathbb{P}(A)$.

Define $I = \left\{ k : \ \|\beta_{(k)}^0\|_2 \neq 0 \right\}$, and since $\hat{\beta}$ is the minimizer of $\ell_\psi(\beta) + \lambda \|W\beta\|_{2,1}$, we get

$$\ell_\psi(\hat{\beta}) + \lambda \|W\hat{\beta}\|_{2,1} \leq \ell_\psi(\beta^0) + \lambda \|W\beta^0\|_{2,1}. \tag{A1}$$

Adding $\lambda \|W(\hat{\beta} - \beta^0)\|_{2,1}$ to both sides of (A1) and rearranging the inequality, we obtain

$$\begin{aligned} \ell_\psi(\hat{\beta}) - \ell_\psi(\beta^0) + \lambda \|W(\hat{\beta} - \beta^0)\|_{2,1} &\leq \lambda \|W\beta^0\|_{2,1} - \lambda \|W\hat{\beta}\|_{2,1} + \lambda \|W(\hat{\beta} - \beta^0)\|_{2,1} \\ &\leq 2\lambda \|W_I(\hat{\beta} - \beta^0)_{(I)}\|_{2,1}. \end{aligned} \tag{A2}$$

According to the fact that $\ell_\psi(\beta^0)$ is a convex function, by applying the Cauchy–Schwarz inequality, its Taylor expansion is as follows

$$\begin{aligned}\ell_\psi(\hat{\beta}) - \ell_\psi(\beta^0) &\geq (\hat{\beta} - \beta^0)^T \nabla \ell_\psi(\beta^0) \\ &\geq -\sum_{l=1}^{g} \sqrt{\sum_{j \in G_l} \nabla \ell_\psi^2(\beta_j^0)} / \omega_l \cdot \omega_l \| (\hat{\beta} - \beta^0)_{(l)} \|_2 \\ &\geq -\max_{1 \leq l \leq g} \sqrt{\sum_{j \in G_l} \nabla \ell_\psi^2(\beta_j^0)} / \omega_l \cdot \sum_{l=1}^{g} \omega_l \| (\hat{\beta} - \beta^0)_{(l)} \|_2 \\ &\geq -z\lambda \| W(\hat{\beta} - \beta^0) \|_{2,1}. \end{aligned} \tag{A3}$$

Combining (A2) and (A3) and defining $\delta_{(l)} = \hat{\beta}_{(l)} - \beta^0_{(l)}$, we obtain the weighted restricted group

$$\| W_{I^C} \delta_{(I^c)} \|_{2,1} \leq \alpha \| W_I \delta_{(I)} \|_{2,1}.$$

Therefore, in the event A, we have $\mu(s, \alpha) > 0$ for $\alpha = \frac{1+z}{1-z}$.

Then, due to $\ell_\psi(\beta^0)$ satisfying the condition of being three-times differentiable, define the function $g(t) = \ell_\psi(\beta^0 + t\delta)$. By applying the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |g'''(t)| &\leq \tau_0 \max_{1 \leq i \leq n} |x_i^T \delta| g''(t) \\ &\leq \tau_0 \max_{1 \leq i \leq n} \sum_{l=1}^{g} \left(\sqrt{\sum_{j \in G_l} x_{ij}^2} / \omega_l \right) \omega_l \| \delta_{(l)} \|_2 g''(t) \\ &\leq \tau_0 \max_{1 \leq i \leq n} \max_{1 \leq l \leq g} \left(\sqrt{\sum_{j \in G_l} x_{ij}^2} / \omega_l \right) \| W\delta \|_{2,1} g''(t) \\ &\leq \tau_0 \left(\mathcal{M} / \min_{1 \leq l \leq g} \omega_l \right) (\alpha + 1) \sqrt{s} \| W_I \delta_{(I)} \|_{2,2} g''(t). \end{aligned}$$

Make $\overline{\mathcal{M}} = \tau_0(\alpha + 1)\sqrt{s}\mathcal{M} / \min_{1 \leq l \leq g} \omega_l$, where ω_l is a real-valued constant; thus, $\overline{\mathcal{M}}$ is bounded, and this means that $|g'''(t)| \leq \overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2} g''(t)$. By Lemma A1, we have

$$\ell_\psi(\hat{\beta}) - \ell_\psi(\beta^0) \geq \delta^T \nabla \ell_\psi(\beta^0) + \frac{\delta^T \mathcal{H}_\psi(\beta^0) \delta}{\overline{\mathcal{M}}^2 \| W_I \delta_{(I)} \|_{2,2}^2} \left(e^{-\overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2}} + \overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2} - 1 \right). \tag{A4}$$

Combining (A3) and (A4), we have the following result

$$\begin{aligned} -z\lambda \| W\delta \|_{2,1} + \frac{\delta^T \mathcal{H}_\psi(\beta^0) \delta}{\overline{\mathcal{M}}^2 \| W_I \delta_{(I)} \|_{2,2}^2} \left(e^{-\overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2}} + \overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2} - 1 \right) \\ \leq \lambda \| W_I \delta_{(I)} \|_{2,1} - \lambda \| W_{I^C} \delta_{(I^c)} \|_{2,1}. \end{aligned}$$

Furthermore, using the group restricted eigenvalue condition, we obtain

$$\frac{\mu(s, \alpha)}{\overline{\mathcal{M}}^2} \left(e^{-\overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2}} + \overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2} - 1 \right) + (1 - z)\lambda \| W\delta \|_{2,1} \leq 2\lambda \sqrt{s} \| W_I \delta_{(I)} \|_{2,2}. \tag{A5}$$

This implies that

$$e^{-\overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2}} + \overline{\mathcal{M}} \| W_I \delta_{(I)} \|_{2,2} - 1 \leq \frac{2\lambda \sqrt{s}}{\mu(s, \alpha)} \overline{\mathcal{M}}^2 \| W_I \delta_{(I)} \|_{2,2}. \tag{A6}$$

In fact, we can reach the conclusion as follows under all $t \in [0, 1)$

$$\exp(\frac{-2t}{1-t}) + 2t - 1 \geq 0.$$

Therefore, we adopt $t = \overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2}/(2 + \overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2})$, which meets the above conditions, and then we obtain

$$e^{-\overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2}} + \overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2} - 1 \geq \frac{\overline{\mathcal{M}}^2\|W_I\delta_{(I)}\|_{2,2}^2}{2 + \overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2}}. \tag{A7}$$

Combining (A6) and (A7), we have

$$\frac{\|W_I\delta_{(I)}\|_{2,2}}{2 + \overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2}} \leq \frac{2\lambda\sqrt{s}}{\mu(s,\alpha)}.$$

Based on the group restricted eigenvalue condition, choose $\lambda \leq \frac{k(1-z)\mu(s,\alpha)}{8\tau_0 s\overline{\mathcal{M}}}$, for a positive constant $k < \min\limits_{1\leq l\leq g} \omega_l$ and substitute it into the above equation

$$\overline{\mathcal{M}}\|W_I\delta_{(I)}\|_{2,2} \leq \frac{2k}{\min\limits_{1\leq l\leq g} \omega_l - k}.$$

Then, substituting this equation into (A7), we have

$$e^{-\overline{\mathcal{M}}\|W\delta\|_{2,2}} + \overline{\mathcal{M}}\|W\delta\|_{2,2} - 1 \geq \frac{\min\limits_{1\leq l\leq g} \omega_l - k}{2\min\limits_{1\leq l\leq g} \omega_l}\overline{\mathcal{M}}^2\|W_I\delta_{(I)}\|_{2.2}^2. \tag{A8}$$

Combining (A5) and (A8) and because of the Cauchy–Schwarz inequality, we have that

$$\begin{aligned}\frac{\min\limits_{1\leq l\leq g} \omega_l - k}{2\min\limits_{1\leq l\leq g} \omega_l}\mu(s,\alpha)\|W_I\delta_{(I)}\|_{2,2}^2 + (1-z)\lambda\|W\delta\|_{2,1} &\leq 2\lambda\|W_I\delta_{(I)}\|_{2,1} \\ &\leq 2\lambda\sqrt{s}\|W_I\delta_{(I)}\|_{2,2} \\ &\leq a\lambda^2 s + \frac{1}{a}\|W_I\delta_{(I)}\|_{2,2}^2.\end{aligned}$$

Let $a = \frac{2\min\limits_{1\leq l\leq g} \omega_l}{(\min\limits_{1\leq l\leq g} \omega_l - k)\mu(s,\alpha)}$; then, we have the following conclusion under the event A

$$\|W\delta\|_{2,1} \leq \frac{2\min\limits_{1\leq l\leq g} \omega_l\lambda s}{(\min\limits_{1\leq l\leq g} \omega_l - k)(1-z)\mu(s,\alpha)},$$

which means that

$$\|\delta\|_{2,1} \leq \frac{2\lambda s}{(\min\limits_{1\leq l\leq g} \omega_l - k)(1-z)\mu(s,\alpha)}.$$

Furthermore, Equation (12) follows from (11) by applying Lemma A2.
Furthermore, by (A2) and (A3), we obtain

$$|\ell_\psi(\hat{\beta}) - \ell_\psi(\beta^0)| \leq \lambda\|W\delta\|_{2,1} \leq \frac{2\min\limits_{1\leq l\leq g} \omega_l\lambda^2 s}{(\min\limits_{1\leq l\leq g} \omega_l - k)(1-z)\mu(s,\alpha)}$$

Now, we prove the probability of event A

$$\begin{aligned}\mathbb{P}(A^c) &= \mathbb{P}\left\{\max_{1\le l\le g}\sqrt{\sum_{j\in G_l}\nabla\ell_\psi^2(\beta_j^0)}/\omega_l > z\lambda\right\}\\ &\le \mathbb{P}\left\{\max_{1\le l\le g}\max_{j\in G_l}|G_l|\frac{\nabla\ell_\psi^2(\beta_j^0)}{\omega_l^2} > (z\lambda)^2\right\}\\ &\le \mathbb{P}\left\{\max_{1\le j\le p}|\nabla\ell_\psi(\beta_j^0)| > \frac{z\lambda\omega_l}{\sqrt{|G_l|}}\right\},\end{aligned}$$

Take $\eta = \Phi^{-1}(1-\frac{\epsilon}{2p})$ and $\lambda\omega_l = \frac{\mathcal{N}(\beta^0)}{z}\sqrt{\frac{G_l}{n}}\eta$, then it follows that

$$\begin{aligned}\mathbb{P}(A^c) &\le p\max_{1\le j\le p}\mathbb{P}\left\{|\nabla\ell_\psi(\beta_j^0)| > \frac{z\lambda\omega_l}{\sqrt{|G_l|}}\right\}\\ &\le p\max_{1\le j\le p}\mathbb{P}\left\{|\frac{1}{n}\sum_{i=1}^n\left\{\psi(x_i^T\beta^0)[G(x_i^T\beta^0)-Y_i]x_{ij}\right\}| > \frac{z\lambda\omega_l}{\sqrt{|G_l|}}\right\}\\ &= p\max_{1\le j\le p}\mathbb{P}\left\{|\sum_{i=1}^n\kappa_{ij}| > \sqrt{n}\mathcal{N}(\beta^0)\eta\right\},\end{aligned}$$

where $\kappa_{ij} = \psi(x_i^T\beta^0)[G(x_i^T\beta^0)-Y_i]x_{ij}$. Furthermore, with assumptions, we obtain that

$$\begin{aligned}&\mathbb{E}(\kappa_{ij}) = \psi(x_i^T\beta^0)[G(x_i^T\beta^0)-\mathbb{E}(Y_i)]x_{ij} = 0,\\ &\mathbb{E}(\kappa_{ij}^2) = \mathrm{Var}(\kappa_{ij}) = \psi^2(x_i^T\beta^0)G(x_i^T\beta^0)(1-G(x_i^T\beta^0))x_{ij}^2 = \mathcal{N}^2(\beta^0),\end{aligned}$$

because of

$$|\kappa_{ij}| \le \psi(x_i^T\beta^0)[G(x_i^T\beta^0)-Y_i](\max_{i,j}|x_{ij}|) \le \mathcal{MR},$$

with a positive constant $\mathcal{R} = \max_{1\le i\le n}\psi(x_i^T\beta^0)$, $0 \le G(x_i^T\beta^0) \le 1$. $\mathcal{F}_{ij} = \kappa_{ij}/(\mathcal{MR})$, where $|\mathcal{F}_{ij}| \le 1$, $\mathbb{E}(\mathcal{F}_{ij}) = 0$.

$$\begin{aligned}B_{nj}^2 &= \sum_{j=1}^n\mathbb{E}(\mathcal{F}_{ij}^2) = \sum_{j=1}^n\mathbb{E}(\kappa_{ij}^2)/(\mathcal{MR})^2 \le n\mathcal{N}^2(\beta^0)/(\mathcal{MR})^2,\\ L_{nj} &= \sum_{j=1}^n\mathbb{E}(|\mathcal{F}_{ij}|^3)/B_{nj}^3 \le \sum_{j=1}^n\mathbb{E}(|\mathcal{F}_{ij}|^2)/B_{nj}^3 = \frac{1}{B_{nj}}.\end{aligned}$$

Then, $B_{nj} = O(\sqrt{n})$ and $L_{nj} = O(1/\sqrt{n})$. By Lemma A3, we have

$$\begin{aligned}\mathbb{P}\left\{|\sum_{i=1}^n\kappa_{ij}| > \sqrt{n}\mathcal{N}(\beta^0)\eta\right\} &= \mathbb{P}\left\{|\sum_{i=1}^n\mathcal{F}_{ij}| > \frac{\sqrt{n}\mathcal{N}(\beta^0)}{\mathcal{MR}}\eta\right\}\\ &\le \mathbb{P}\left\{|\sum_{i=1}^n\mathcal{F}_{ij}| > B_{nj}\eta\right\}\\ &= 2(1+O(1)\eta^3L_{nj})(1-\Phi(\eta))\\ &= \frac{\epsilon}{p}\Big(1+O(\eta^3/\sqrt{n})\Big).\end{aligned}$$

Note that for any $\eta > 0$, we have $1 - \Phi(\eta) \le \Phi(\eta)/\eta$; then,

$$\frac{\epsilon}{2p} = 1 - \Phi(\eta) \le \frac{\Phi(\eta)}{\eta} = \frac{\exp(-\eta^2/2)}{\sqrt{2\pi}\eta}.$$

Our default $p > 2$ has $p/\epsilon > 2$, which means that $\eta > \Phi^{-1}(3/4) > 1/\sqrt{2\pi}$, and so

$$\frac{\epsilon}{2p} \le \frac{\exp(-\eta^2/2)}{\sqrt{2\pi}\eta} < \exp(-\frac{\eta^2}{2}).$$

Here, we get

$$\eta < \sqrt{2\log\frac{2p}{\epsilon}}.$$

As $n, p \to \infty$ with $n \le p = o(e^{n^{1/3}})$, we have

$$\mathbb{P}(A^c) \le \epsilon(1 + o(1)).$$

which completes the proof of Theorem 1. □

Proof of Theorem 2. We only need to show that the action of the weight function in the form of (15) under logistic loss satisfies the Assumption (A3).

Denote $g(t) = \ell_\psi(u + tv; X, Y)$ for $u, v \in \mathbb{R}^p$, and then we have

$$\begin{aligned}
g'(t) &= \frac{1}{2n}\sum_{i=1}^{n}\left\{(1-Y_i)\exp(\frac{x_i^Tu + x_i^Ttv}{2}) - Y_i\exp(-\frac{x_i^Tu + x_i^Ttv}{2})\right\}v^Tx_i,\\
g''(t) &= \frac{1}{4n}\sum_{i=1}^{n}\left\{(1-Y_i)\exp(\frac{x_i^Tu + x_i^Ttv}{2}) + Y_i\exp(-\frac{x_i^Tu + x_i^Ttv}{2})\right\}(v^Tx_i)^2,\\
g'''(t) &= \frac{1}{8n}\sum_{i=1}^{n}\left\{(1-Y_i)\exp(\frac{x_i^Tu + x_i^Ttv}{2}) - Y_i\exp(-\frac{x_i^Tu + x_i^Ttv}{2})\right\}(v^Tx_i)^3.
\end{aligned}$$

It is not difficult to find that $|g''(t)| = g''(t)$, and then

$$\begin{aligned}
|g'''(t)| &= \frac{1}{8n}\left|\sum_{i=1}^{n}\left\{(1-Y_i)\exp(\frac{x_i^Tu + x_i^Ttv}{2}) - Y_i\exp(-\frac{x_i^Tu + x_i^Ttv}{2})\right\}(v^Tx_i)^3\right|\\
&\le \frac{1}{2}\max_{1\le i\le n}|x_i^Tv|\frac{1}{4n}\left\{\sum_{i=1}^{n}\left|(1-Y_i)\exp(\frac{x_i^Tu + x_i^Ttv}{2})\right| + \left|Y_i\exp(-\frac{x_i^Tu + x_i^Ttv}{2})\right|\right\}(v^Tx_i)^2\\
&= \frac{1}{2}(\max_{1\le i\le n}|x_i^Tv|)|g''(t)|.
\end{aligned}$$

which completes the proof of Theorem 2. □

References

1. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
2. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
3. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
4. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351.
5. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef] [PubMed]
6. Sur, P.; Chen, Y.; Candès, E.J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Relat. Fields* **2019**, *175*, 487–558. [CrossRef]

7. Ma, R.; Tony Cai, T.; Li, H. Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Am. Stat. Assoc.* **2021**, *116*, 984–998. [CrossRef]
8. Bianco, A.M.; Boente, G.; Chebi, G. Penalized robust estimators in sparse logistic regression. *Test* **2022**, *31*, 563–594. [CrossRef]
9. Abramovich, F.; Grinshtein, V. High-dimensional classification by sparse logistic regression. *IEEE Trans. Inf. Theory* **2018**, *65*, 3068–3079. [CrossRef]
10. Huang, H.; Gao, Y.; Zhang, H.; Li, B. Weighted Lasso estimates for sparse logistic regression: Non-asymptotic properties with measurement errors. *Acta Math. Sci.* **2021**, *41*, 207–230. [CrossRef]
11. Yin, Z. Variable selection for sparse logistic regression. *Metrika* **2020**, *83*, 821–836. [CrossRef]
12. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
13. Meier, L.; Van De Geer, S.; Bühlmann, P. The group lasso for logistic regression. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2008**, *70*, 53–71. [CrossRef]
14. Wang, L.; You, Y.; Lian, H. Convergence and sparsity of Lasso and group Lasso in high-dimensional generalized linear models. *Stat. Pap.* **2015**, *56*, 819–828. [CrossRef]
15. Blazere, M.; Loubes, J.M.; Gamboa, F. Oracle Inequalities for a Group Lasso Procedure Applied to Generalized Linear Models in High Dimension. *IEEE Trans. Inf. Theory* **2014**, *60*, 2303–2318. [CrossRef]
16. Kwemou, M. Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model. *ESAIM Probab. Stat.* **2016**, *20*, 309–331. [CrossRef]
17. Nowakowski, S.; Pokarowski, P.; Rejchel, W.; Sołtys, A. Improving group Lasso for high-dimensional categorical data. In *Proceedings of the International Conference on Computational Science*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 455–470.
18. Zhang, Y.; Wei, C.; Liu, X. Group Logistic Regression Models with Lp, q Regularization. *Mathematics* **2022**, *10*, 2227. [CrossRef]
19. Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **2001**, *109*, 475–494. [CrossRef]
20. Breheny, P.; Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **2015**, *25*, 173–187. [CrossRef]
21. Abramovich, F.; Grinshtein, V.; Levy, T. Multiclass classification by sparse multinomial logistic regression. *IEEE Trans. Inf. Theory* **2021**, *67*, 4637–4646. [CrossRef]
22. Chen, S.; Wang, P. Gene selection from biological data via group LASSO for logistic regression model: Effects of different clustering algorithms. In Proceedings of the 2021 40th Chinese Control Conference (CCC), Shanghai, China, 26–28 July 2021; pp. 6374–6379.
23. Ryan Kilcullen, J.; Castonguay, L.G.; Janis, R.A.; Hallquist, M.N.; Hayes, J.A.; Locke, B.D. Predicting future courses of psychotherapy within a grouped LASSO framework. *Psychother. Res.* **2021**, *31*, 63–77. [CrossRef] [PubMed]
24. Yang, Y.; Hu, X.; Jiang, H. Group penalized logistic regressions predict up and down trends for stock prices. *N. Am. J. Econ. Financ.* **2022**, *59*, 101564. [CrossRef]
25. Belloni, A.; Chernozhukov, V.; Wang, L. Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **2011**, *98*, 791–806. [CrossRef]
26. Bunea, F.; Lederer, J.; She, Y. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory* **2013**, *60*, 1313–1325. [CrossRef]
27. Huang, Y.; Wang, C. Consistent functional methods for logistic regression with errors in covariates. *J. Am. Stat. Assoc.* **2001**, *96*, 1469–1482. [CrossRef]
28. Bach, F. Self-concordant analysis for logistic regression. *Electron. J. Stat.* **2010**, *4*, 384–414. [CrossRef]
29. Hu, Y.; Li, C.; Meng, K.; Qin, J.; Yang, X. Group sparse optimization via lp, q regularization. *J. Mach. Learn. Res.* **2017**, *18*, 960–1011.
30. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* **2009**, *37*, 1705–1732. [CrossRef]
31. Tseng, P.; Yun, S. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.* **2009**, *117*, 387–423. [CrossRef]
32. Yang, Y.; Zou, H. A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.* **2015**, *25*, 1129–1141. [CrossRef]
33. Graham, K.; de Las Morenas, A.; Tripathi, A.; King, C.; Kavanah, M.; Mendez, J.; Stone, M.; Slama, J.; Miller, M.; Antoine, G.; et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br. J. Cancer* **2010**, *102*, 1284–1293. [CrossRef] [PubMed]
34. Sakhanenko, A. Berry-Esseen type estimates for large deviation probabilities. *Sib. Math. J.* **1991**, *32*, 647–656. [CrossRef]