

Article

Group Logistic Regression Models with $l_{p,q}$ Regularization

Yanfang Zhang, Chuanhua Wei * and Xiaolin Liu

College of Science, Minzu University of China, 27 Zhongguancun South Street, Haidian District, Beijing 100081, China; zhangyanfang@muc.edu.cn (Y.Z.); 20301920@muc.edu.cn (X.L.)

* Correspondence: chwei@muc.edu.cn

Abstract: In this paper, we proposed a logistic regression model with $l_{p,q}$ regularization that could give a group sparse solution. The model could be applied to variable-selection problems with sparse group structures. In the context of big data, the solutions for practical problems are often group sparse, so it is necessary to study this kind of model. We defined the model from three perspectives: theoretical, algorithmic and numeric. From the theoretical perspective, by introducing the notion of the group restricted eigenvalue condition, we gave the oracle inequality, which was an important property for the variable-selection problems. The global recovery bound was also established for the logistic regression model with $l_{p,q}$ regularization. From the algorithmic perspective, we applied the well-known alternating direction method of multipliers (ADMM) algorithm to solve the model. The subproblems for the ADMM algorithm were solved effectively. From the numerical perspective, we performed experiments for simulated data and real data in the factor stock selection. We employed the ADMM algorithm that we presented in the paper to solve the model. The numerical results were also presented. We found that the model was effective in terms of variable selection and prediction.

Keywords: logistic regression; group sparse; oracle inequality; $l_{p,q}$ regularization; ADMM algorithm; multi-factors stock selection

MSC: 62J07; 62J12



Citation: Zhang, Y.; Wei, C.; Liu, X.

Group Logistic Regression Models with $l_{p,q}$ Regularization. *Mathematics* **2022**, *10*, 2227. <https://doi.org/10.3390/math10132227>

Academic Editors: Huiming Zhang and Ting Yan

Received: 24 April 2022

Accepted: 23 June 2022

Published: 25 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

For regression models, the categorical variables are important for applications and the explanatory variables are always thought of as grouped. Considering the interpretability and the accuracy of the models, the information regarding the group should be considered for the modeling, especially for the high-dimensional settings where sparsity and variable selection can play a very important role in estimation accuracy. Generally speaking, the regression model with the penalized regularizations gives a good result for variable-selection problems; we found lots of research in the literature for this kind of problem [1–12]. When the explanatory variables have a group structure, penalized regularization also plays an important role, such as with group least absolute shrinkage and selection operators (LASSOs) [13], the group smoothly clipped absolute deviation (SCAD) penalty [14] and the group minimax concave penalty (MCP) [15] models.

As we all know, the l_p ($0 < p < 1$) norm is seen as a good approximation of the l_0 norm, and it can recover a more sparse solution than the l_1 norm [16]. For the group sparsity of variables with a group structure, the $l_{p,q}$ norm plays an important role in the sparse aspect. The $l_{p,q}$ norm with a group structure is described as follows:

$$\|\beta\|_{p,q} := \left(\sum_{i=1}^r \|\beta_{G_i}\|_p^q \right)^{\frac{1}{q}}, \quad (1)$$

where $\beta := (\beta_{\mathcal{G}_1}^T, \dots, \beta_{\mathcal{G}_r}^T)^T$, and $\{\beta_{\mathcal{G}_i} \in R^{n_i} : i = 1, \dots, r\}$ is the grouping of the variable β . Here $\sum_{i=1}^r n_i = d + 1$, and \mathcal{G}_i denotes the index set corresponding to the i -th group. We denote \mathcal{S}, \mathcal{N} to be the index sets, and $\mathcal{G}_{\mathcal{S}}$ is to denote the index set $\{\mathcal{G}_i : i \in \mathcal{S}\}$.

For the linear regression problem with $l_{p,q}$ regularization, the oracle inequality and the global recovery bound were established in the paper [17]. The goodness of $l_{p,q}$ regularization was very obvious. For the logistic regression model, we employed penalized regularization in the variable-selection problems.

We assume that $\beta = (\beta_0, \beta_1, \dots, \beta_d)^T \in R^{d+1}$ are the coefficients of the explanatory variables. The matrix X denotes the explanatory variables, and it is given as follows:

$$X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ 1 & X_{21} & \cdots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix} \in R^{n \times (d+1)}.$$

X_i denotes the i -th row of the matrix X , $y = (y_1, \dots, y_n)^T$ are the categorical variables and $y_i \in \{0, 1\}$.

In this paper, we considered the logistic regression model with the $l_{p,q}$ norm, described as follows:

$$\min_{\beta} f(\beta) + \lambda \|\beta\|_{p,q}^q, \tag{2}$$

where $f(\beta) := -\frac{1}{n}y^T X\beta + \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(X_i\beta))$ is the loss function and $\|\cdot\|_{p,q}$ is defined by formulation (1). Moreover, we know $f(\beta) > 0$ from the properties of the logistic regression model, and $\|\beta\|_{p,q}^q = \sum_{i=1}^r \|\beta_{\mathcal{G}_i}\|_p^q$, $p \geq 1, 0 < q < 1$ and λ is the penalized parameter.

The group LASSO for the logistic regression model [18] is able to perform variable selection on groups of variables, and the model has the following form:

$$\min_{\beta} f(\beta) + \lambda \sum_{i=1}^r s(df_{\mathcal{G}_i}) \|\beta_{\mathcal{G}_i}\|_2, \tag{3}$$

where $\lambda \geq 0$ controls the amount of penalization, and $s(df_{\mathcal{G}_i})$ is used to rescale the penalty with respect to the dimensionality of the parameter vector $\beta_{\mathcal{G}_i}$.

Moreover, a quite general composite absolute penalty for the group sparsity problem is considered in [19], and this model includes the group LASSO as a special case. The group LASSO is an important extension of regularization, and it proposes an l_2 regularization for each group and, ultimately, gives the sparsity in a group manner. This property can be found in the numerical experiments.

We found that models (2) and (3) were logistic regression models with different penalized regularizations. They aimed to give a solution with group sparsity. The logistic regression model with $l_{p,q}$ regularization is different with the LASSO logistic regression, and it can give a more sparse solution within a group or between groups by adjusting the values of p and q . This could be found in the numerical experiments.

To illustrate the goodness of model (2), we also introduced the logistic regression problem with the elastic net penalty that could give a sparse solution of the problem, and it is described as follows:

$$\min_{\beta} f(\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2, \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the penalized parameters. Model (4) was not good at group sparsity, and we showed this in the numerical parts.

This paper is organized as follows. In Section 2, we introduce the inequalities of $l_{p,q}$ regularization, the properties of the loss function for the logistic regression model, the (p, q) -group restricted eigenvalue condition relative to (S, N) ((p, q) -GREC(S, N)) and establish the oracle inequality and the global recovery bound for model (2). In Section 3, we apply the ADMM to solve model (2), and we found that the subproblems of the algorithm could be solved efficiently. In Section 4, we employ two numerical experiments that are always used for variable-selection problems to show the goodness of model (2) and the ADMM algorithm. The results for the LASSO logistic regression model (3) and the logistic regression model with the elastic net penalty (4) are compared with those of model (2), and we determine the advantages of model (2). We also give the results for model (2), which produced the real data in Section 5, and the results showed the effectiveness of the model and the algorithm that we gave in the paper. The last section draws conclusions and presents future work.

We introduce some notations that we will use in the following analysis. Let $\mathcal{S} := \{i \in \{1, \dots, r\} : \beta_{\mathcal{G}_i} \neq 0\}$ be the index set of nonzero groups of β , $\mathcal{S}^c := \{1, \dots, r\} \setminus \mathcal{S}$ be the complement of \mathcal{S} , $S := |\mathcal{S}|$ be the group sparsity of β . For a variable $\beta \in R^{d+1}$ and $S \subseteq \{1, \dots, d+1\}$, we employ β_S to denote the subvector of β corresponding to \mathcal{S} . For a group $\beta_{\mathcal{G}_i}$, we employ $\beta_{\mathcal{G}_i} = 0$ to describe a zero group, where $\beta_{\mathcal{G}_i} = 0$ means that $\beta_j = 0$ for all $j \in \mathcal{G}_i$. We give $S \leq N \ll r$, $\beta \in R^{d+1}$ and $\mathcal{J} \subseteq \{1, \dots, r\}$, and we use $\text{rank}_i(\beta)$ to denote the rank of $\|\beta_{\mathcal{G}_i}\|_p$ among $\{\|\beta_{\mathcal{G}_j}\|_p : j \in \mathcal{J}^c\}$ (in a decreasing order). We employ $\mathcal{J}(\beta; N)$ to denote the index set of the first N largest groups in the value of $\|\beta_{\mathcal{G}_i}\|_p$ among $\{\|\beta_{\mathcal{G}_j}\|_p : j \in \mathcal{J}^c\}$, which means

$$\mathcal{J}(\beta; N) := \{i \in \mathcal{J}^c : \text{rank}_i(\beta) \in \{1, \dots, N\}\}.$$

Moreover, we let $R := \lceil \frac{r-|\mathcal{J}|}{N} \rceil$, and we denote

$$\mathcal{J}_k(x; N) := \begin{cases} \{i \in \mathcal{J}^c : \text{rank}_i(\beta) \in \{kN + 1, \dots, (k+1)N\}\}, & k = 1, \dots, R-1, \\ \{i \in \mathcal{J}^c : \text{rank}_i(\beta) \in \{RN + 1, \dots, r - |\mathcal{J}|\}\}, & k = R. \end{cases} \tag{5}$$

2. Theoretical Analysis

In this section, we analyze the oracle property and the global recovery bound of the penalized regression model (2). Firstly, we introduce the following inequalities of the $l_{p,q}$ norm and the properties of the loss function $f(\beta)$.

Lemma 1 ([17], p. 8). *Let $0 < q \leq p \leq 2$, $\beta \in R^{d+1}$ and K be the smallest integer such that $2^{K-1}q \geq 1$. Then the following relation holds:*

$$\|\beta\|_{p,q}^q \leq r^{1-2^{-K}} \|\beta\|_{p,2}^q.$$

Lemma 2 ([17], p. 9). *Let $0 < q \leq 1 \leq p$ and $\beta_1, \beta_2 \in R^{d+1}$. Then we have*

$$\|\beta_1\|_{p,q}^q - \|\beta_2\|_{p,q}^q \leq \|\beta_1 - \beta_2\|_{p,q}^q.$$

Lemma 3 ([17], p. 13). *Let $0 < q \leq 1 \leq p$, $\tau \geq 1$ and $\beta \in R^{d+1}$, $\mathcal{N} := \mathcal{J}(\beta; N) \cup \mathcal{J}$ and $\mathcal{J}_k := \mathcal{J}_k(\beta; N)$ for $k = 1, \dots, R$. Then the following inequalities hold:*

$$\|\beta_{\mathcal{G}_{\mathcal{N}^c}}\|_{p,\tau} \leq \sum_{k=1}^R \|\beta_{\mathcal{G}_{\mathcal{J}_k}}\|_{p,\tau} \leq N^{\frac{1}{\tau}-\frac{1}{q}} \|\beta_{\mathcal{G}_{\mathcal{J}^c}}\|_{p,q}.$$

Moreover, the following properties are about the Lipschitz continuity and the convexity of the loss function $f(\beta)$.

Proposition 1. For $\beta_1, \beta_2 \in R^{d+1}$, we have

$$|f(\beta_1) - f(\beta_2)| \leq 2n\|X\|_2\|\beta_1 - \beta_2\|_2. \tag{6}$$

Proof. For $\beta_1, \beta_2, \tilde{\beta} \in R^{d+1}$, based on the differential mean value theorem and the properties of the norms we can get

$$\begin{aligned} |f(\beta_1) - f(\beta_2)| &= \left| -\frac{1}{n}y^T X(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^n (\ln(1 + \exp(X_i\beta_1)) - \ln(1 + \exp(X_i\beta_2))) \right| \\ &= \left| -\frac{1}{n}y^T X(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^n \frac{\exp(X_i\tilde{\beta})X_i^T}{1 + \exp(X_i\tilde{\beta})}(\beta_1 - \beta_2) \right| \\ &\leq \left| -\frac{1}{n}y^T X(\beta_1 - \beta_2) \right| + \left| \frac{1}{n} \sum_{i=1}^n X_i^T(\beta_1 - \beta_2) \right| \\ &\leq (n + \|y\|_2)\|X\|_2\|\beta_1 - \beta_2\|_2 \\ &\leq 2n\|X\|_2\|\beta_1 - \beta_2\|_2. \end{aligned} \tag{7}$$

Hence, we obtain our desirable result. \square

Proposition 2. For $\beta \in R^{d+1}$, the function $f(\beta)$ is convex.

Proof. From the definition of the function $f(\beta)$, we know the Hessian matrix of the function $f(\beta)$ is

$$H_f(\beta) = \frac{1}{n} \sum_{i=1}^n X_i^T X_i \frac{\exp(X_i\beta)}{(1 + \exp(X_i\beta))^2}.$$

Here, we find for $i = 1, \dots, n$ the matrix $X_i^T X_i \succeq 0$ and $\frac{\exp(X_i\beta)}{(1 + \exp(X_i\beta))^2} > 0$. Thus, the Hessian matrix $H_f(\beta)$ is a positive semi-definite matrix. Hence, the function $f(\beta)$ is convex. \square

The above lemmas and properties state the inequalities for $l_{p,q}$ regularization, and they will help the proof of the oracle inequality and the global recovery bound. The oracle inequalities for prediction error were discussed in [20,21], and they were derived without restricted eigenvalue conditions for LASSO-type estimators or sparsity. Moreover, for the group LASSO problems, the oracle inequalities were discussed in [22–24] under the restricted eigenvalue assumption. For linear regression with $l_{p,q}$ regularization, the oracle inequality was established in [17] with the help of the (p, q) -GREC(S,N).

Moreover, the (p, q) -GREC(S,N) was very important for the analysis of the oracle property and the global recovery bound of the $l_{p,q}$ norm. We define γ to be the smallest non-zero eigenvalue of the Hessian matrix of the function $f(\beta)$. We introduce it in the following definition.

Definition 1. Let $0 < q \leq p \leq 2$. The (p, q) -GREC(S,N) is said to be satisfied if

$$\phi_{p,q}(S, N) := \min \left\{ \frac{\gamma\|\beta\|_2}{\|\beta_{\mathcal{G}_N}\|_{p,2}} : |\mathcal{J}| \leq S, \|\beta_{\mathcal{G}_\mathcal{J}}\|_{p,q} \leq \|\beta_{\mathcal{G}_\mathcal{J}}\|_{p,q}, \mathcal{N} = \mathcal{J}(\beta, N) \cup \mathcal{J} \right\} > 0.$$

The oracle property is an important property for the variable selection, and it gives an upper bound on the square error of the logistic regression problem and the violation of the true nonzero groups for each point in the level set of the objective function of problem (2).

For $\tilde{\beta} \in R^{d+1}$, the level set is given as follows:

$$Lev_F(\tilde{\beta}) := \{\beta \in R^{d+1} : f(\beta) + \lambda\|\beta\|_{p,q}^q \leq \lambda\|\tilde{\beta}\|_{p,q}^q\}.$$

From the definition of the level set in ([25], p. 8), we know that many properties of the optimization problem (2) relate to the level set $Lev_F(\bar{\beta})$.

Theorem 1. Let $0 < q \leq 1 \leq p, S > 0$ and let the (p, q) -GREC(S, S) hold. Let $\bar{\beta}$ be the unique solution of $\min_{\beta} f(\beta)$ at a group sparsity level S , and \mathcal{S} be the index set of nonzero groups of $\bar{\beta}$. Let K be the smallest integer such that $2^{K-1}q \geq 1$. Then, for any $\beta^* \in Lev_F(\bar{\beta})$, which means that $f(\beta^*) + \lambda \|\beta^*\|_{p,q}^q \leq \lambda \|\bar{\beta}\|_{p,q}^q$, the following oracle inequality holds:

$$f(\beta^*) - f(\bar{\beta}) + \lambda \|\beta_{\mathcal{G}_{Sc}}^*\|_{p,q}^q \leq 2^{\frac{q}{2-q}} \lambda^{\frac{2}{2-q}} \gamma^{\frac{q}{2-q}} S^{\frac{2(1-2^{-K})}{2-q}} / \phi_{p,q}^{\frac{2q}{2-q}}(S, S). \tag{8}$$

Moreover, letting $\mathcal{N}_* := \mathcal{S} \cup \mathcal{S}(\beta^*; S)$, we have

$$\|\beta_{\mathcal{G}_{N_*}}^* - \bar{\beta}_{\mathcal{G}_{N_*}}\|_{p,2}^2 \leq 2^{\frac{2}{2-q}} \gamma^{\frac{2}{2-q}} \lambda^{\frac{2}{2-q}} S^{2(1-2^{-K})/(2-q)} / \phi_{p,q}^{\frac{4}{2-q}}(S, S).$$

Proof. Let $\beta^* \in Lev_F(\bar{\beta})$, and by the definition of the level set $Lev_F(\bar{\beta})$ we have

$$f(\beta^*) + \lambda \|\beta^*\|_{p,q}^q \leq \lambda \|\bar{\beta}\|_{p,q}^q.$$

Then by Lemmas 1 and 2 and the fact $f(\bar{\beta}) > 0$, one has the following formulation:

$$\begin{aligned} f(\beta^*) - f(\bar{\beta}) + \lambda \|\beta_{\mathcal{G}_{Sc}}^*\|_{p,q}^q &\leq f(\beta^*) + \lambda \|\beta_{\mathcal{G}_{Sc}}^*\|_{p,q}^q \\ &\leq \lambda (\|\bar{\beta}_{\mathcal{G}_S}\|_{p,q}^q - \|\beta_{\mathcal{G}_S}^*\|_{p,q}^q) \\ &\leq \lambda \|\bar{\beta}_{\mathcal{G}_S} - \beta_{\mathcal{G}_S}^*\|_{p,q}^q \\ &\leq \lambda S^{1-2^{-K}} \|\bar{\beta}_{\mathcal{G}_S} - \beta_{\mathcal{G}_S}^*\|_{p,2}^q. \end{aligned} \tag{9}$$

Moreover, we find

$$\|\beta_{\mathcal{G}_{Sc}}^* - \bar{\beta}_{\mathcal{G}_{Sc}}\|_{p,q}^q - \|\beta_{\mathcal{G}_S}^* - \bar{\beta}_{\mathcal{G}_S}\|_{p,q}^q \leq \|\beta_{\mathcal{G}_{Sc}}^*\|_{p,q}^q - (\|\bar{\beta}_{\mathcal{G}_S}\|_{p,q}^q - \|\beta_{\mathcal{G}_S}^*\|_{p,q}^q) = \|\beta^*\|_{p,q}^q - \|\bar{\beta}\|_{p,q}^q \leq 0.$$

Then, the (p, q) -GREC(S, S) implies the following:

$$\|\bar{\beta}_{\mathcal{G}_S} - \beta_{\mathcal{G}_S}^*\|_{p,2} \leq (\gamma \|\beta^* - \bar{\beta}\|_2) / \phi_{p,q}(S, S).$$

From the expansion of the Taylor formulation, we obtain the following relationship:

$$\begin{aligned} f(\beta^*) - f(\bar{\beta}) &= \nabla f(\bar{\beta})^T (\beta^* - \bar{\beta}) + \frac{1}{2} (\beta^* - \bar{\beta})^T \nabla^2 f(\bar{\beta}) (\beta^* - \bar{\beta}) \\ &\geq \frac{1}{2} (\beta^* - \bar{\beta})^T \nabla^2 f(\bar{\beta}) (\beta^* - \bar{\beta}) \\ &\geq \frac{1}{2} \gamma \|\beta^* - \bar{\beta}\|_2^2, \end{aligned} \tag{10}$$

where $\tilde{\beta} \in \{\bar{\beta} : \|\tilde{\beta} - \bar{\beta}\|_2 \leq \|\beta^* - \bar{\beta}\|_2\}$. The first inequality of formulation (10) is based on the fact that $\bar{\beta}$ is the unique optimal solution of $\min_{\beta} f(\beta)$ at group sparsity level S . Moreover, because of the uniqueness of $\bar{\beta}$, we obtain that the smallest eigenvalue of the Hessian matrix of the function $f(\beta)$ is positive. Hence, we obtain the second inequality.

Then, we get

$$\gamma \|\beta^* - \bar{\beta}\|_2^2 \leq 2(f(\beta^*) - f(\bar{\beta})). \tag{11}$$

Combining this with formulation (9), we get

$$f(\beta^*) - f(\bar{\beta}) + \lambda \|\beta_{\mathcal{G}_{Sc}}^*\|_{p,q}^q \leq 2^{\frac{q}{2}} \lambda \gamma^{\frac{q}{2}} S^{1-2^{-K}} (f(\beta^*) - f(\bar{\beta}))^{\frac{q}{2}} / \phi_{p,q}^q(S, S). \tag{12}$$

Thus, we have

$$f(\beta^*) - f(\bar{\beta}) \leq 2^{2-q} \gamma^{2-q} \lambda^{2-q} S^{2(1-2^{-K})/(2-q)} / \phi_{p,q}^{2-q}(S, S). \tag{13}$$

Hence, by formulations (12) and (13), we obtain the oracle inequality (8). Moreover, from the definition of \mathcal{N}_* , the (p, q) -GREC(S, S) implies that

$$\|\beta_{\mathcal{G}_{\mathcal{N}_*}}^* - \bar{\beta}_{\mathcal{G}_{\mathcal{N}_*}}\|_{p,2}^2 \leq 2\gamma(f(\beta^*) - f(\bar{\beta})) / \phi_{p,q}^2(S, S) \leq 2^{2-q} \gamma^{2-q} \lambda^{2-q} S^{2(1-2^{-K})/(2-q)} / \phi_{p,q}^{2-q}(S, S).$$

Thus, the proof is complete. \square

In the following, we established the global recovery bound for the $l_{p,q}$ regularization problem (2). The global recovery bound shows that the sparse solution $\bar{\beta}$ could be recovered by any point β^* in the level set $lev_F(\bar{\beta})$. Here β^* is a global optimal solution of problem (2) when the penalized parameter λ is small enough. We show the global recovery bound for the $l_{p,q}$ regularization problem (2) in the next theorem.

Theorem 2. Let $0 < q \leq 1 \leq p \leq 2$, $S > 0$ and suppose the (p, q) -GREC(S, S) holds. Let $\bar{\beta}$ be the unique solution of $\min_{\beta} f(\beta)$ at group sparsity level S , and S be the index set of nonzero groups of $\bar{\beta}$. Let K be the smallest integer such that $2^{K-1}q \geq 1$. Then, for any $\beta^* \in lev_F(\bar{\beta})$, the following global recovery bound for problem (2) holds:

$$\|\beta^* - \bar{\beta}\|_2^2 \leq 2 * 2^{2-q} \lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)}} / \phi_{p,q}^{2-q}(S, S). \tag{14}$$

More precisely,

$$\|\beta^* - \bar{\beta}\|_2^2 \leq \begin{cases} \mathcal{O}(\lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S), & 2^{K-1}q = 1, \\ \mathcal{O}(\lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{4-q}{2-q}}), & 2^{K-1}q > 1. \end{cases} \tag{15}$$

Proof. We suppose $\mathcal{N}_* := S \cup S(\beta^*; S)$ as defined in Theorem 1. Since $p \leq 2$, from Lemma 3 and Theorem 1, we get

$$\|\beta_{\mathcal{N}_*^c}^*\|_2^2 \leq \|\beta_{\mathcal{G}_{\mathcal{N}_*^c}}^*\|_{p,2}^2 \leq S^{1-2/q} \|\beta_{\mathcal{G}_{S^c}}^*\|_{p,q}^2 \leq 2^{2-q} \lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)}} / \phi_{p,q}^{2-q}(S, S).$$

Furthermore, from Theorem 1 and the fact $2^{K-1}q \geq 1$, we get

$$\begin{aligned} \|\beta^* - \bar{\beta}\|_2^2 &= \|\beta_{\mathcal{G}_{\mathcal{N}_*}}^* - \bar{\beta}_{\mathcal{G}_{\mathcal{N}_*}}\|_2^2 + \|\beta_{\mathcal{G}_{\mathcal{N}_*^c}}^*\|_2^2 \\ &\leq 2^{2-q} \gamma^{2-q} \lambda^{2-q} S^{2(1-2^{-K})/(2-q)} / \phi_{p,q}^{2-q}(S, S) + 2^{2-q} \lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)}} / \phi_{p,q}^{2-q}(S, S) \\ &\leq 2 * 2^{2-q} \lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)}} / \phi_{p,q}^{2-q}(S, S). \end{aligned} \tag{16}$$

Hence, formulation (14) holds.

Moreover, if $2^{K-1}q = 1$, we have $\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)} = 1$ and we get

$$\|\beta^* - \bar{\beta}\|_2^2 \leq \mathcal{O}(\lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S).$$

If $2^{K-1}q > 1$, we know $q > 2^{1-K}$. Hence $\frac{q-2}{q} + \frac{4(1-2^{-K})}{q(2-q)} \leq \frac{4-q}{2-q}$, and we get

$$\|\beta^* - \bar{\beta}\|_2^2 \leq \mathcal{O}(\lambda^{\frac{4}{q(2-q)}} \gamma^{2-q} S^{\frac{4-q}{2-q}}).$$

Thus, formulation (15) holds. \square

Remark 1. From Proposition 2, we know that $f(\beta)$ is convex. By the convexity of the function $f(\beta)$, we know it is suitable for the assumption for the variable β . The conditions of the variable β help us to obtain desirable results for Theorems 1 and 2.

3. ADMM Algorithm

In this section, we give an algorithm based on the ADMM algorithm [26,27] for solving the logistic regression model with $l_{p,q}$ regularization (2). The ADMM algorithm performs very well for problems where the variables can be separated. Model (2) can be equivalently described as follows:

$$\begin{aligned} \min_{\beta} \quad & f(\beta) + \lambda \|r\|_{p,q}^q, \\ \text{s.t.} \quad & \beta = r. \end{aligned} \tag{17}$$

The augmented Lagrange function of the above model is as follows:

$$L_{\rho}(\beta, r, u) = f(\beta) + \lambda \|r\|_{p,q}^q + u^T(\beta - r) + \frac{\rho}{2} \|\beta - r\|_2^2, \tag{18}$$

where $u \in R^{d+1}$ is the dual variable, and $\rho > 0$ is the augmented Lagrange multiplier.

Generally speaking, the structure of the ADMM algorithm is given as follows:

$$\beta^{k+1} = \operatorname{argmin}_{\beta} f(\beta) + (\beta - r^k)^T u^k + \frac{\rho}{2} \|\beta - r^k\|_2^2, \tag{19}$$

$$r^{k+1} = \operatorname{argmin}_r \lambda \|r\|_{p,q}^q + (\beta^{k+1} - r)^T u^k + \frac{\rho}{2} \|\beta^{k+1} - r\|_2^2, \tag{20}$$

$$u^{k+1} = u^k + \rho(\beta^{k+1} - r^{k+1}). \tag{21}$$

Based on the above structure and Propositions 1 and 2, the subproblem (19) is an unconstrained convex optimization problem and the objective function is Lipschitz continuous. Based on these good properties of the subproblem (19), we found that it could be effectively solved by many optimal algorithms, such as the trust region algorithm, the sequential quadratic programming algorithm, the algorithm based on the gradient and so on. Moreover, the first order optimal conditions of problem (19) are given by the following formulation:

$$\frac{\partial \left(f(\beta) + u^{kT}(\beta - r^k) + \frac{\rho}{2} \|\beta - r^k\|_2^2 \right)}{\partial \beta} = 0_{d+1}.$$

Thus, we obtain the optimal solution for subproblem (19) by solving the following nonlinear equations:

$$n\rho\beta - \sum_{i=1}^n \frac{X_i^T}{1 + \exp(X_i\beta)} = n\rho r^k - nu^k + X^T(y - e_n), \tag{22}$$

where $e_n = \underbrace{(1, 1, \dots, 1)}_n^T \in R^n$.

From this fact we know that the variable β is group sparse. Hence, we can divide the variables β^{k+1} and u^k by the group structure. $\beta_{G_i}^{k+1}$ and $u_{G_i}^{k+1}$ denote the G_i th group variables of β^{k+1} and u^k , respectively. Subproblem (20) is solved by employing a group structure. Then, for $i = 1, \dots, r$, $r_{G_i}^{k+1}$ can be given by solving the following optimization problem:

$$r_{G_i}^{k+1} = \operatorname{argmin}_{r_m} \lambda \|r_{G_i}\|_{p,q}^q + (\beta_{G_i}^{k+1} - r_{G_i})^T u_{G_i}^k + \frac{\rho}{2} \|\beta_{G_i}^{k+1} - r_{G_i}\|_2^2. \tag{23}$$

We can then obtain the solution for subproblem (20) by the following:

$$r^{k+1} = (r_{\mathcal{G}_1}^{k+1T}, \dots, r_{\mathcal{G}_r}^{k+1T})^T. \tag{24}$$

Moreover, we find that the problem (23) can be equivalently solved by the following one:

$$(r_{\mathcal{G}_i})^{k+1} = \operatorname{argmin}_{r_{\mathcal{G}_i}} \|r_{\mathcal{G}_i} - (\beta_{\mathcal{G}_i}^{k+1} + \frac{u_{\mathcal{G}_i}^k}{\rho})\|_2^2 + \frac{2\lambda}{\rho} \|r_{\mathcal{G}_i}\|_{p,q}. \tag{25}$$

The proximal gradient method given in [17] has proven very useful for solving (25). Based on the above analysis, we found that the ADMM algorithm was effective for solving the logistic regression problem with $l_{p,q}$ regularization, and we describe the structure of Algorithm 1 in the following.

Algorithm 1 ADMM algorithm for solving (2).

- Step 1: Initialization: give $\beta^0, u^0, r^0, \rho > 0, \lambda > 0$, and set $k = 0$;
- Step 2: for $k = 0, 1, \dots$, if the stop criteria is satisfied, the algorithm is stopped; otherwise go to Step 3;
- Step 3: Update β^{k+1} : β^{k+1} is given by solving nonlinear Equation (22);
- Step 4: Update r^{k+1} : for $i = 1, \dots, r$, we employ the proximal gradient method to solve the optimization problem (25) and give $r_{\mathcal{G}_i}^{k+1}$. Then, we have

$$r^{k+1} = (r_{\mathcal{G}_1}^{k+1T}, \dots, r_{\mathcal{G}_r}^{k+1T})^T.$$

- Step 5: Update u^{k+1} : $u^{k+1} = u^k + \rho(\beta^{k+1} - r^{k+1})$.
-

4. Simulation Examples

In this section, we employ the simulation data to illustrate the efficiency of the logistic regression model with $l_{p,q}$ regularization and the ADMM algorithm. $l_{1/2}$ regularization is shown to give a more sparse optimal solution than the l_1 norm [28]. Hence, we employed $q = 1/2$ to do our numerical experiments, and we used the ADMM algorithm that we designed in Section 3 to solve model (2). The environment for the simulations was Python 3.7.

In order to verify the effect of the prediction and the classification for penalized logistic regression model (2), we designed two simulation experiments with different data structures. At the same time, we employed the LASSO logistic regression model and the logistic regression model with the elastic net penalty to solve the numerical problems. The numerical results illustrated the advantages of the logistic regression model with $l_{p,q}$ regularization.

We mainly considered two aspects of the effects of the models: the ability of the model to select variables; the ability of the model to test the effects of the classifications and predictions for the penalized logistic regression models. The evaluation indexes for the models in this section mainly included the following:

- P : the number of non-zero coefficients in the variables that the model gives.
- TP : the number of coefficients predicted to be non-zero which are actually non-zero.
- TN : the number of coefficients predicted to be zero which are actually zero.
- FP : the number of coefficients predicted to be non-zero but which are actually zero.
- FN : the number of coefficients predicted to be zero but which are actually non-zero.
- PSR : the ratio of the number of non-zero coefficients in the variables for the predicted case to that of the true case, which is calculated by the following:

$$PSR = \frac{TP}{p}.$$

- *Accuracy*: the accuracy of the prediction for the test data, which is calculated by the following formulation:

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

- *AUC*: area under the curve.

A *P* value close to *TP* shows that the model is good. The greater the *Accuracy* and *AUC*, the better the model. *PSR* being close to 1 shows the goodness of the model.

Moreover, for the penalized parameter λ , we used the test set verification method to choose it. Firstly, we selected a value of λ that made all coefficients equal to 0, and we set it as λ_{max} . Secondly, we chose a number that was very close to 0, such as 0.0001, and set it as λ_{min} . We chose $\lambda \in [\lambda_{min}, \lambda_{max}]$ to do the numerical experiments. Finally, we gave $\lambda \in [\lambda_{min}, \lambda_{max}]$, and it produced a maximum value for the *AUC*. The augmented Lagrange multiplier ρ did not influence the convergence of the algorithm, but the proper one would give a fast convergence rate. When we performed the numerical experiments, we just chose a better one, which meant it made the algorithm converge quickly.

4.1. Simulation Experiment with Non-Sparse Variables in the Group

Firstly, we constructed the group structure features with similar features in the group and different features between different groups, and we obtained the data as it is given in [1]. The data was generated according to the following model:

$$y_i = \frac{1}{1 + \exp(-X_i\beta + \varepsilon)}.$$

Here, the explanatory variables for the groups followed the multivariate normal distributions, which meant $X_i \sim N(0, \Sigma_{p_s})$, and the error followed the standard normal distribution, which meant $\varepsilon \sim N(0, 1)$. The correlation coefficient of the variables $(X_{G_i})_i$ and $(X_{G_i})_j$ was $\rho_s^{|i-j|}$. Generally speaking, we employed $\rho_s = 0.2$ or $\rho_s = 0.7$ to denote the weak correlation or strong correlation for the variables in the group.

For this simulation experiment, we generated data using 10 groups independently, and each group contained five variables. Hence, the total number of variables was 50. There were three groups that were significant, and the other seven groups were not significant. The correlation coefficients within the groups were 0.2 and 0.7, respectively. The sample size was 500. We selected 80% of the data for the training set and the others were the test set. The experimental simulation was repeated 30 times. For this example, the penalized parameter was $\lambda = 0.001$ for the LASSO logistic regression model. For model (2) the penalized parameter was $\lambda = 0.02$ and the augmented Lagrange multiplier was chosen as $\rho = 0.1$. For the logistic regression model with the elastic net penalty, the parameters were set as $\lambda_1 = 0.0009$ and $\lambda_2 = 0.0001$.

The following table gives the numerical results for this example with different models and different correlation coefficients.

According to Table 1, the logistic regression model with the elastic net penalty, the LASSO logistic regression model and the logistic regression model with $l_{p,q}$ regularization could perform variable selection. When the correlation of variables was different, the criteria of the logistic regression model with $l_{p,q}$ regularization were better than the other two models. According to the indexes *P* and *TP*, it could be seen that all variables with non-zero coefficients in the logistic regression model with $l_{2,1/2}$ regularization were screened out when the correlation was different. The value of *P* was closer to that of *TP*, and all the selected variables were variables with non-zero coefficients. When choosing different parameters, the logistic regression model with $l_{p,q}$ regularization selected more variables with true non-zero coefficients than that of the other two models, and the logistic regression model with $l_{p,q}$ regularization gave a solution that was closer to the number of variables with true non-zero coefficients. The *PSR* was closer to 1 and could select significant

variables. From the prediction effect of the model, the AUC and accuracy of the logistic regression model with $l_{p,q}$ regularization showed better prediction effects with different correlation coefficients.

Table 1. The results for the simulations in experiment one.

The Correlation Coefficient	Models	P	TP	PSR	Accuracy	AUC
$\rho_s = 0.2$	Elastic Net	33.07	15.00	1.0000	0.9520	0.9503
	LASSO	32.17	15.00	1.0000	0.9503	0.9497
	$p = 2, q = \frac{1}{2}$	15.00	15.00	1.0000	0.9657	0.9658
	$p = 1, q = \frac{1}{2}$	15.20	14.93	0.9956	0.9580	0.9582
$\rho_s = 0.7$	Elastic Net	31.33	15.00	1.0000	0.9577	0.9571
	LASSO	29.47	14.83	0.9889	0.9577	0.9571
	$p = 2, q = \frac{1}{2}$	15.00	15.00	1.0000	0.9710	0.9716
	$p = 1, q = \frac{1}{2}$	16.03	14.90	0.9933	0.9637	0.9634

According to the variable-selection effect and the prediction effect of the models, the logistic regression model with $L_{2,1/2}$ regularization was better than those of the logistic regression model with the elastic net penalty and the LASSO logistic regression model. Since the data for this simulation experiment was designed to be zero or non-zero in one group, the logistic regression model with $L_{2,1/2}$ regularization performed better than the logistic regression model with $L_{1,1/2}$ regularization. The logistic regression model with $L_{2,1/2}$ regularization could select a set of variables or not, and it had an ideal group variable-selection effect. The logistic regression model with the elastic net penalty and the LASSO logistic regression model compressed the variables to achieve a sparse effect, but because they did not make full use of the group structure information of the variables, they selected too many variables with zero coefficients during variable selection, so the performance of the two models was not good.

4.2. Simulation Experiment with the Sparse Variables in the Group

The data give were similar to those in the above experiment. The difference between these two simulation experiments are given as follows. A total of six groups of variables with intragroup correlation were simulated. Each group contained 10 variables. A total of two groups of 12 variables were significant. One group of variables was completely significant, and the other group contained two significant variables. The sample size was 500. The ratio of the training set to the test set was 8 : 2. The correlation coefficients within the group were 0.2 and 0.7, respectively. For this example, the penalized parameter was $\lambda = 0.001$ for the LASSO logistic regression model, and for model (2) the penalized parameter was $\lambda = 0.032$ and the augmented Lagrange multiplier was 0.1. When $\rho_s = 0.2$, the parameters for the logistic regression model with the elastic net penalty were $\lambda_1 = 0.0018$ and $\lambda_2 = 0.0002$. Moreover, when $\rho_s = 0.7$, the parameters for the logistic regression model with the elastic net penalty were $\lambda_1 = 0.0021$ and $\lambda_2 = 0.0009$.

The following table gives us the numerical results for this example with different models and different correlation coefficients.

According to Table 2, the logistic regression model with $L_{1,1/2}$ regularization performed very well. From the perspective of variable selection, the logistic regression model with the elastic net penalty, the LASSO logistic regression model and the logistic regression model with $l_{p,q}$ regularization could screen out all the variables with non-zero coefficients. The logistic regression model with the elastic net penalty and the LASSO logistic regression model screened out too many variables with non-zero coefficients. However, the logistic regression model with $l_{p,q}$ regularization not only gave all the variables with non-zero coefficients, but also the values of P and TP were very close, which meant the variables given

by the logistic regression model with $l_{p,q}$ regularization were close to the predicted ones. For this kind of data, we found that the logistic regression model with $l_{2,1/2}$ regularization tended to compress a group of variables to be zero or non-zero at the same time. Therefore, the logistic regression model with $l_{2,1/2}$ regularization selected a significant difference, but it did not filter out the important variables in the group. In terms of the prediction effect of the model, the logistic regression model with $l_{1,1/2}$ regularization performed well for different correlation coefficients, and its prediction ability also improved compared with the univariate selection model.

Table 2. The results for the simulations in experiment two.

The Correlation Coefficient	Models	P	TP	PSR	$Accuracy$	AUC
$\rho_s = 0.2$	Elastic Net	29.83	12.00	1.0000	0.9467	0.9467
	LASSO	36.00	12.00	1.0000	0.9480	0.9478
	$p = 2, q = \frac{1}{2}$	14.67	12.00	1.0000	0.9580	0.9650
	$p = 1, q = \frac{1}{2}$	12.37	12.00	1.0000	0.9651	0.9673
$\rho_s = 0.7$	Elastic Net	24.36	12.00	1.0000	0.9570	0.9561
	LASSO	29.40	12.00	1.0000	0.9590	0.9589
	$p = 2, q = \frac{1}{2}$	17.33	12.00	1.0000	0.9640	0.9635
	$p = 1, q = \frac{1}{2}$	12.90	12.00	1.0000	0.9700	0.9704

Combining the effects of variable selection and model prediction, the logistic regression model with $l_{p,q}$ regularization performed well when the variables were sparse in the group. Due to the “all in all out” mechanism, the logistic regression model with $l_{2,1/2}$ regularization could only screen out important variable groups. The important variables in the group could not be screened out, and the variable-selection ability was not good. But the logistic regression model with $l_{1,1/2}$ regularization could overcome this disadvantage.

From the above two simulation experiments, we found the goodness of the logistic regression model with $l_{p,q}$ regularization for variable selection and prediction for the data with a group structure. Moreover, for different data, we adjusted the value of p to adapt to the problems.

5. Real-Data Experiment

Data description and preprocessing are described as follows. In this section, the real data are considered. The data came from the excellent mining quantitative platform, and the website is <https://uqer.datayes.com/> (accessed on 31 October 2021). The data were the factor data and the yield data of constituent stocks for the Shanghai and Shenzhen 300 index in China’s stock market from 1 January 2010 to 31 December 2020. The advantages for using these data were good performance, large scale, high liquidity and active trading in the market. In order to ensure the accuracy and rationality of the analysis results, we needed to select and correct the range of samples. According to the development experience of China’s stock market and previous research experience, the empirical part located the sample starting point in 2010. Secondly, because the capital of companies in the financial industry has the characteristics of high leverage and high debt, the construction of some financial indicators is quite different from the other listed companies. Therefore, we excluded some financial companies based on our experience. In addition, ST and PT stocks in the market have abnormal financial conditions and performance losses. We also excluded such kinds of stocks with weak comparability. Among the 243 stock factors visible on the excellent mining quantitative platform, 34 factors that belong to nine groups were selected to evaluate model (2). The data were daily data, but in practice, if daily transaction data were used for investment, the frequent transactions would lead to a significant increase in transaction costs. The rise of transaction costs would affect the annualized rate of return.

In order to reduce the impact of transaction costs, we used monthly transaction data for modeling. Hence, it was necessary to conduct a monthly average processing for each factor’s data, and we used monthly data for stock selection. In the division of the data set, the data from 1 January 2010 to 30 April 2018 were selected as the training set, and the data from 1 May 2018 to 31 December 2020 were used for back testing.

In order to ensure the quality of subsequent factor screening and the effect of stock rise and fall prediction, we needed to preprocess the data before the analysis. The methods for data preprocessing included noise cleaning, missing value processing, data standardization, lag processing and so on.

Using the above factors and data processing methods to generate the stock factor matrix, we employed the logistic regression model with $l_{p,q}$ regularization and the ADMM algorithm proposed in Section 3 to estimate the coefficients of the factors. We also calculated the posterior probability of the stock, sorted the probability from large to small, and bought the top ten stocks with equal weight.

In the following, we introduce the historical back test evaluation index. When we performed the back test analysis for the solutions, in order to perform the objective reflective and comprehensive evaluation of the solutions, it was necessary to give the evaluation indicators. We selected nine indicators to do this, and they were the return rate of the year, the return rate of the benchmark year, the sharp ratio, the volatility, the return unrelated to the market fluctuations (α), the sensitivity to market changes (β), the information ratio, the maximum pullback, the turnover rate of the year, etc.

Based on the above evaluation criteria, we used the selected data to back test and verify the effectiveness of model (2), and we employed the above model to predict the stock trend. Moreover, we sorted the predicted yield data, and selected the top 10 stocks with the highest rise in probability as the stock portfolio for the month, according to the equal weight reconstruction portfolio. We held it until the end of the month to calculate the month’s income. The initial capital was set at 10 million yuan, the tax for buying was 0.003, and the tax for selling was 0.0013. The sliding point was 0. Moreover, the parameters for model (2) were given as $\lambda = 0.004$. $\lambda = 0.003$ was chosen for the LASSO logistic regression model. We adopted $\lambda_1 = \lambda_2 = 0.02$ for the logistic regression model with the elastic net penalty. We employed the ADMM algorithm that we gave in Section 3 to solve this example, and the Lagrange multipliers for the ADMM algorithm to solve these models were chosen as $\rho = 0.1$. After the calculation, we listed the following transaction back test results and the cumulative yield figure, which are listed in Table 3 and Figure 1, respectively.

Table 3. The back test results.

The Logistic Regression with Different Penalties	The $l_{1,1/2}$ Norm	LASSO	Elastic Net
The return rate of the year	22.7%	20.5%	18.8%
The return rate of the benchmark year	13.3%	13.3%	13.3%
α	10.5%	6.2%	8.0%
β	0.89	0.92	0.64
The sharp ratio	0.84	0.72	0.41
The volatility	22.8%	23.5%	23.7%
The information ratio	0.68	0.54	0.41
The maximum pullback	18.8%	27.7%	25.1%
The turnover rate of the year	9.14	7.67	9.33



Figure 1. The cumulative yield figure.

From Table 3, we found that the return rates for the year for these three models were all higher than the return rate of the benchmark year, 13.3%. For the return rate of the benchmark year, the logistic regression model with $l_{1,1/2}$ regularization was the highest. Model (2) gave a good strategy from the perspective of excess return α and the sharp ratio. Under the same risk coefficient, the investment strategy based on the logistic regression model with $l_{1,1/2}$ regularization could help investors make effective investment decisions and obtain higher yields. Moreover, the model could give a strategy with an acceptable range in the maximum pullback, and the strategy could also more effectively prevent the pullback risk.

From the graph of the cumulative rate of return, the line given by the logistic regression model with $l_{1,1/2}$ regularization was basically always above the benchmark annualized rate of the return curve, which indicated that the return of the portfolio constructed with the group information was always stable. The portfolio constructed based on the logistic regression model with $l_{1,1/2}$ regularization could not only screen out the important factor types affecting stock returns, but could also screen out the important factor indicators in the group, so as to more accurately predict the probability of stock returns rising. When constructing the portfolio based on this model, it had certain advantages over the other two regression models.

6. Conclusions

Combined with the data requirements, this paper proposed a logistic regression model with $l_{p,q}$ regularization. We showed the properties of the $l_{p,q}$ norm and the loss function of the logistic regression problem. Moreover, the oracle inequality for the $l_{p,q}$ norm and the global recovery bound for the penalized regression model were established with the help of the (p,q) -group restricted eigenvalue condition. These properties were important for variable selection. In Section 3, we showed the framework for the ADMM algorithm for solving the penalized logistic regression model. For the algorithm, we gave a method for solving the subproblems, so as to reduce the difficulty and complexity of solving the model (2).

By the numerical simulation results, since the logistic regression with $l_{p,q}$ regularization comprehensively considered the group structure information of variables, compared with the univariate selection method considering only a single variable, it could eliminate more redundant variables, so as to screen out the more important characteristics of dependent variables. It also had higher accuracy when we performed the test on the test set. Moreover,

for the real-data experiment, it was found that the logistic regression model with $l_{p,q}$ regularization could show a more stable effect in variable selection and prediction.

In future work, we could extend the group logistic regression model with $L_{p,q}$ regularization. In the theoretical part, we could do more analysis, such as local recovery bounds and design a more appropriate algorithm, which could give a convergence solution. Moreover, in the numerical parts, we could choose more p and q to illustrate the goodness of the model (2).

Author Contributions: Conceptualization and methodology, Y.Z. and C.W.; investigation and data curation, X.L.; writing—original draft preparation, writing—review and editing, Y.Z. and C.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors' work was supported by the National Natural Science Foundation of China (No. 12171027), National Statistical Science Research Project (2019LZ40), State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences, and the Youth Foundation of Minzu University of China.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [[CrossRef](#)]
2. Huang, J.; Ma, S.; Xie, H.; Zhang, C. A group bridge approach for variable selection. *Biometrika* **2009**, *96*, 339–355. [[CrossRef](#)]
3. Kim, S.; Koh, K.; Lustig, M.; Boyd, S.; Gorinevsky, D. An interior-point method for large-scale regularized least squares. *IEEE J. Sel. Top. Signal Process.* **2007**, *1*, 606–617. [[CrossRef](#)]
4. Meinshausen, N. Relaxed lasso. *Comput. Stat. Data Anal.* **2006**, *52*, 374–393. [[CrossRef](#)]
5. Nikolova, M.; Ng, M.K.; Zhang, S.; Ching, W. Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. *SIAM J. Imaging Sci.* **2008**, *1*, 2–25. [[CrossRef](#)]
6. Ong, C.S.; An, L.T.H. Leaning sparse classifiers with difference of convex functions algorithms. *Optim. Method Softw.* **2013**, *28*, 830–854. [[CrossRef](#)]
7. Soubies, E.; Blance-Fraud, L.; Aubert, G. A continuous exact penalty (cel0) for least squares regularized problem. *SIAM J. Imaging Sci.* **2015**, *8*, 1607–1639. [[CrossRef](#)]
8. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. B* **2011**, *73*, 273–282. [[CrossRef](#)]
9. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [[CrossRef](#)]
10. Zhang, T. Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **2010**, *11*, 1081–1107.
11. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **2005**, *67*, 301–320. [[CrossRef](#)]
12. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [[CrossRef](#)]
13. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* **2006**, *68*, 49–67. [[CrossRef](#)]
14. Wang, L.; Li, H.; Huang, J.Z. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Am. Stat. Assoc.* **2008**, *103*, 1556–1569. [[CrossRef](#)] [[PubMed](#)]
15. Huang, J.; Breheny, P.; Ma, S. A selective review of group selection in high-dimensional models. *Stat. Sci.* **2012**, *27*, 481–499. [[CrossRef](#)] [[PubMed](#)]
16. Chartrand, R.; Staneva, V. Restricted isometry properties and nonconvex compressive sensing. *Inverse Probl.* **2008**, *24*, 035020. [[CrossRef](#)]
17. Hu, Y.; Li, C.; Meng, K.; Qin, J.; Yang, X. Group sparse optimization via $l_{p,q}$ regularization. *J. Mach. Learn. Res.* **2017**, *18*, 1–52.
18. Meier, L.; Sara, G.; Bühlmann, P. The group lasso for logistic regression. *J. R. Statist. Soc. B* **2008**, *70*, 53–71. [[CrossRef](#)]
19. Zhao, P.; Rocha, G.; Yu, B. The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **2009**, *37*, 3468–3497. [[CrossRef](#)]
20. Bartlett, P.L.; Mendelson, S.; Neeman, J. L_1 regularized linear regression: Persistence and oracle inequalities. *Probab. Theory Relat. Fields* **2012**, *154*, 193–224. [[CrossRef](#)]
21. Greenshtein, E.; Ritov, Y.A. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **2012**, *10*, 971–988. [[CrossRef](#)]
22. Blazère, M.; Loubes, J.M.; Gamboa, F. Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Trans. Inform. Theory* **2014**, *60*, 2303–2318.

23. 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](#)]
24. Xiao, Y.; Yan, T.; Zhang, H.; Zhang, Y. Oracle inequalities for weighted group lasso in high-dimensional misspecified Cox models. *J. Inequal. Appl.* **2020**, *1*, 1–33. [[CrossRef](#)]
25. Rockafellar, R.T.; Wets, R.J.-B. *Variational Analysis*, 3rd ed.; Springer: New York, NY, USA, 2009.
26. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **2011**, *3*, 1–122. [[CrossRef](#)]
27. Han, D. A survey on some recent developments of alternating direction method of multipliers. *J. Oper. Res. Soc. China* **2022**, *10*, 1–52. [[CrossRef](#)]
28. Xu, Z.; Chang, X.; Xu, F.; Zhang, H. $l_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 1013–1027.