

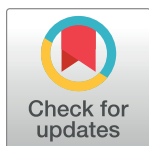


RESEARCH ARTICLE

Logistic regression with image covariates via the combination of L_1 and Sobolev regularizations

Baiguo An ^{*}, Beibei Zhang ^{*}

School of Statistics, Capital University of Economics and Business, Beijing, China

 These authors contributed equally to this work.^{*} anbg200@163.com

Abstract

The use of image covariates to build a classification model has lots of impact in various fields, such as computer science, medicine, and so on. The aim of this paper is to develop an estimation method for logistic regression model with image covariates. We propose a novel regularized estimation approach, where the regularization is a combination of L_1 regularization and Sobolev norm regularization. The L_1 penalty can perform variable selection, while the Sobolev norm penalty can capture the shape edges information of image data. We develop an efficient algorithm for the optimization problem. We also establish a nonasymptotic error bound on parameter estimation. Simulated studies and a real data application demonstrate that our proposed method performs very well.

OPEN ACCESS

Citation: An B, Zhang B (2020) Logistic regression with image covariates via the combination of L_1 and Sobolev regularizations. PLoS ONE 15(6): e0234975. <https://doi.org/10.1371/journal.pone.0234975>

Editor: Mihye Ahn, University of Nevada, Reno, UNITED STATES

Received: July 31, 2019

Accepted: June 6, 2020

Published: June 26, 2020

Copyright: © 2020 An, Zhang. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The real ZIP Code Dataset from the following website: https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/zipcode.html.

Funding: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

As one of the most important issues in machine learning field, classification plays a prominent role throughout various disciplines. Until now people have developed a large number of classification methods, such as KNN, Linear (Quadratic) discriminant analysis, logistic regression, naive bayes, decision tree, SVM, neural network, deep learning, and many others [1, 2].

Among all those methods logistic regression has a long history [3], and is one of the most popular approaches. Logistic regression model is a typical representative of generalized linear models and linear classification methods. Therefore, this article takes logistic regression as the research object. Traditionally, maximum likelihood method is usually used to obtain an estimator of the parameter in logistic regression model [4–6].

However, the big data era brings us massive complex data, one of whose most prominent characteristics is high dimensionality. The maximum likelihood estimation method in logistic regression model faces serious problems such as non-existence, non-uniqueness [7] in high dimensional settings. Regularization is a popular strategy to handle high dimensional problems [8]. Many regularized methods have been proposed over the past decades, including LASSO [9], the smoothly clipped absolute deviation (SCAD) penalty method [10], the mini-max concave penalty (MCP) method [11], and so on. For high dimensional logistic regression, [12] considered L_1 -regularization path algorithm. [13] proposed an interior-point method for

large-scale L_1 -regularized logistic regression. [14] proposed the group lasso for logistic regression. The $L_{1/2}$ regularized logistic regression is considered by [15] for gene selection in cancer classification.

Image data is a very popular form of data, and generated in many fields, such as computer science, medicine, and so on. In addition to high dimensionality, image data usually contains spatially smooth regions with relatively sharp edges, which leads to its own characteristics including local smoothness [16], jump discontinuity [17], and many others. Local smoothness leads to highly correlated features, which makes the image classification problem more challenging [18]. Jump discontinuity makes conventional smoothing techniques inefficient [17]. On the other hand, using these characteristics in modeling process is often helpful for model efficiency enhancement, and has received a lot of attention recently. For example, [19] introduced a locally adaptive smoothing method for image restoration. [16] proposed Propagation-Separation approach for local likelihood estimation, which can handle local smoothness of image data. [20] developed an adaptive regression model for the analysis of neuroimaging data, which is a generalization of the PS approach. [21] studied theoretical performance of nonlocal means for noise removal of image data. [17] considered a spatially varying coefficient model for neuroimaging data with jump discontinuities. [18] proposed a spatially weighted principal component analysis (SWPCA) for imaging classification. [22] developed a generalized scalar-on-image regression models via total variation regularization, which can keep the piecewise smooth nature of imaging data. [23] proposed an efficient nuclear norm penalized estimation method for matrix linear discriminant analysis.

In this paper, we consider a logistic regression model with image covariates, and develop a regularized estimation approach, which combines the L_1 regularization and the Sobolev norm regularization. The L_1 penalty performs variable selection and removes covariates unrelated to the response from models [9]. The Sobolev norm penalty keeps characteristics of image data (e.g. local smoothness) in model fitting. In fact, the Sobolev regularization is a popular technology in image data analysis, such as image denoising [24], edge detection of images [25], and many others. The proposed regularization method is different from the aforementioned regularized logistic regression models. It is also different from the elastic net method [26], which is a combination of Lasso and ridge regression. The elastic net encourages the grouping effect, where strongly correlated predictors tend to be in or out of the model. However, the elastic net can not exploit structure information of image covariates, and is not suitable for models with image covariates. There are differences between our proposed method and the fused lasso method [27]. In many real data analysis, such as gene expression data, covariates have a order. Adjacent covariates are often highly correlated and have similar effects on the response variable. The fused lasso tends to make adjacent covariates share common effect on the response. The proposed method can be treated as the extended version of the fused lasso from one dimension to multidimensions. Moreover, the fusion term here is Sobolev norm penalty. Furthermore, we develop a novel algorithm to solve the optimization problem. The theoretical property of our estimator is also studied, and a nonasymptotic estimate error bound is given. Numerical studies including simulations and a real data analysis are also considered to verify the performance of our method.

The rest of the article is organized as follows. Section 2 presents the methodology, including model setup, algorithm, and theoretical property. Section 3 is numerical studies, where simulated studies and a real data application are presented. Lastly, we make a short conclusion in Section 4. The proof details of theoretical studies are put in Appendix Section.

Methodology

Model setup

Suppose that we have observations (\mathbf{X}_i, Y_i) with $1 \leq i \leq n$, where $Y_i \in \{-1, +1\}$ is the class label, and $\mathbf{X}_i = (x_{jk}^{(i)} : j = 1, \dots, p; k = 1, \dots, q) \in \mathbb{R}^{p \times q}$ is the corresponding image covariate. We further assume that (\mathbf{X}_i, Y_i) with $1 \leq i \leq n$ are independent and identically distributed. In order to predict Y_i with \mathbf{X}_i , the following logistic regression model is assumed

$$\log \frac{P_i}{1 - P_i} = \langle \mathbf{X}_i, \mathbf{B} \rangle, \tag{1}$$

where $P_i = P(Y_i = +1 | \mathbf{X}_i)$, $\mathbf{B} = (b_{jk}) \in \mathbb{R}^{p \times q}$ is the corresponding coefficient image, and $\langle \mathbf{X}_i, \mathbf{B} \rangle = \sum_{j=1}^p \sum_{k=1}^q x_{jk}^{(i)} b_{jk}$ is the inner operator of two matrices. Let $\beta = \text{vec}(\mathbf{B}) = (\beta_1, \dots, \beta_{pq})^T$ and $X_i = \text{vec}(\mathbf{X}_i) = (x_{ij}, j = 1, \dots, pq)^T$ for $i = 1, \dots, n$, then $\langle \mathbf{X}_i, \mathbf{B} \rangle = \beta^T X_i$, and the model (1) is equivalent to $P_i = 1 / (1 + e^{-\beta^T X_i})$. The true value of β is denoted by $\beta^* = (\beta_1^*, \dots, \beta_{pq}^*)^T$.

Traditionally, maximum likelihood method is usually used to estimate coefficient image \mathbf{B} . The likelihood function is

$$L(\beta) = \prod_{i=1}^n P_i^{I(Y_i=+1)} (1 - P_i)^{I(Y_i=-1)} = \prod_{i=1}^n (1 + e^{-Y_i \beta^T X_i})^{-1},$$

and the corresponding log-likelihood function is $\ln(L(\beta)) = -\sum_{i=1}^n \log(1 + e^{-Y_i \beta^T X_i})$.

Denote logistic loss function as $l(\beta) = \log(1 + e^{-Y_i \beta^T X_i})$, and the associated risk is denoted by $\mathbb{P}l(\beta) = El(\beta)$. We assume that $\beta^* = \arg \min_{\beta} \mathbb{P}l(\beta)$. The empirical risk is denoted by $\mathbb{P}_n l(\beta) = n^{-1} \sum_{i=1}^n \log(1 + e^{-Y_i \beta^T X_i})$. Hence, maximizing the likelihood function is equivalent to minimizing the empirical risk

$$\min_{\beta} \mathbb{P}_n l(\beta).$$

Many optimization methods, such as Newton-Raphson method [6], can be used to solve the above problem.

However, in the image covariate case, the corresponding coefficient image \mathbf{B} is usually assumed to be a piecewise smooth image with unknown edges. This assumption is widely used in the imaging literature, and is critical for addressing various scientific questions [22]. The maximum likelihood method does not take advantage of these characteristics. Moreover, image covariate is usually high dimensional, and not every element of \mathbf{X}_i is useful to predict Y_i . But the maximum likelihood method can not perform variable selection. Consequently, we propose a novel estimation method for \mathbf{B} in the next subsection, which can keep characteristics of image covariate such as local smoothing, and perform variable selection simultaneously.

Estimation

For the coefficient image \mathbf{B} , we define its discrete gradient $\nabla \mathbf{B} \in \mathbb{R}^{p \times q \times 2}$ as

$$(\nabla \mathbf{B})_{jk} = \begin{cases} (b_{j+1,k} - b_{j,k}, b_{j,k+1} - b_{j,k}), & j < p, k < q, \\ (0, b_{j,k+1} - b_{j,k}), & j = p, k < q, \\ (b_{j+1,k} - b_{j,k}, 0), & j < p, k = q, \\ (0, 0), & j = p, k = q. \end{cases}$$

$(\nabla \mathbf{B})_{jk} = (b_{j+1,k} - b_{j,k}, b_{j,k+1} - b_{j,k})$ is the discrete gradient at the position (j, k) . Furthermore, $b_{j+1,k} - b_{j,k}$ is the discrete gradient in the vertical direction, and $b_{j,k+1} - b_{j,k}$ indicates the discrete gradient in the horizontal direction. The Sobolev norm of \mathbf{B} is the L_2 norm of $\nabla \mathbf{B}$, which is written as

$$\|\mathbf{B}\|_{Sob} = \left(\sum_{j=1}^p \sum_{k=1}^q (\nabla \mathbf{B})_{jk}^2 \right)^{1/2}.$$

In fact, we can rewrite $\|\mathbf{B}\|_{Sob}^2$ as a quadratic form of β . Specifically, we define a matrix $D = (d_{ij}) \in \mathbb{R}^{(2pq-p-q) \times pq}$ with d_{ij} defined in the following formula (2)

$$d_{ij} = \begin{cases} -1, & i = (2p-1)(k-1) + s, j = p(k-1) + (s+1)/2, \\ & k = 1, \dots, (q-1), s = 1, 3, \dots, (2p-3); \\ 1, & i = (2p-1)(k-1) + s, j = p(k-1) + (s+1)/2 + 1, \\ & k = 1, \dots, (q-1), s = 1, 3, \dots, (2p-3); \\ -1, & i = (2p-1)(k-1) + s, j = p(k-1) + s/2, \\ & k = 1, \dots, (q-1), s = 2, 4, \dots, (2p-2); \\ 1, & i = (2p-1)(k-1) + s, j = p(k-1) + s/2 + p, \\ & k = 1, \dots, (q-1), s = 2, 4, \dots, (2p-2); \\ -1, & i = (2p-1)k, j = kp, k = 1, \dots, (q-1); \\ 1, & i = (2p-1)k, j = kp + p, k = 1, \dots, (q-1); \\ -1, & i > (2p-1)(q-1), j = i - (p-1)(q-1); \\ 1, & i > (2p-1)(q-1), j = i - (p-1)(q-1) + 1, \\ 0, & \text{else.} \end{cases} \tag{2}$$

Then one can easily verify that $\|\mathbf{B}\|_{Sob}^2 = \beta^T D^T D \beta$. We also present the matrix D with a graph in the case $p = q = 3$ for the purpose of understanding. Please see Fig (1).

We then consider the following optimization problem

$$\min_{\beta} Q(\beta), \tag{3}$$

where $Q(\beta) = \mathbb{P}_n l(\beta) + \lambda_1 \beta^T D^T D \beta + \lambda_2 \|\beta\|_1$, and $\|\cdot\|_1$ is the L_1 norm. The term $\lambda_1 \beta^T D^T D \beta$ shrinks adjacent elements of \mathbf{B} to be similar, hence it can capture the local smoothing of \mathbf{B} . The term $\lambda_2 \|\beta\|_1$ shrinks the elements of \mathbf{B} to 0, and performs variable selection. We next propose an algorithm to solve the optimization problem (3).

Algorithm

For the optimization problem (3), we define $K = D^T D = (k_{jl})$ and $H(\beta) = \mathbb{P}_n l(\beta) + \lambda_1 \beta^T K \beta$, then one can see that $Q(\beta) = H(\beta) + \lambda_2 \|\beta\|_1$. This indicates that the function $Q(\beta)$ is a convex function with the separable structure [28]. [29] shows that the coordinate descent algorithm can be guaranteed to converge to the global minimizer for any convex optimization function

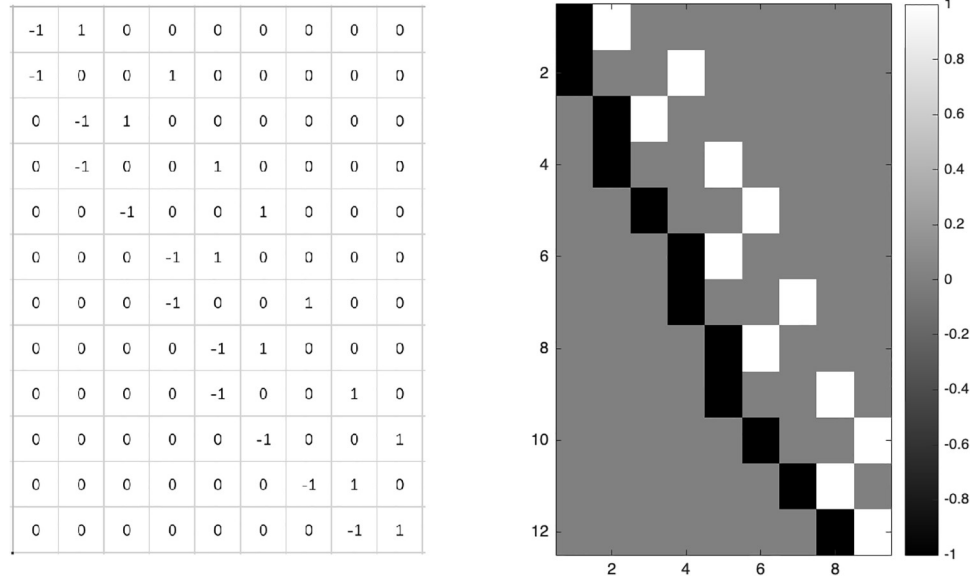


Fig 1. The matrix D in the case $p = q = 3$.

<https://doi.org/10.1371/journal.pone.0234975.g001>

with the separable structure. Hence we here propose a coordinate descent algorithm to obtain the solution of the optimization problem (3).

For $j = 1, \dots, pq$, we successively minimize $Q(\beta)$ along β_j direction with other parameters fixed. Specifically, denote the current value of β as β^c , and $p_i^c = P(Y_i | X_i, \beta^c) = 1 / (1 + e^{-Y_i X_i^T \beta^c})$ for $i = 1, \dots, n$. For $j = 1, \dots, pq$, we use the second order Taylor expansion to approximate function $H(\beta_{-j}^c, \beta_j)$ with $\beta_{-j}^c = (\beta_1^c, \dots, \beta_{j-1}^c, \beta_{j+1}^c, \dots, \beta_{pq}^c)^T$ fixed. Specifically,

$$\frac{\partial H(\beta_{-j}^c, \beta_j)}{\partial \beta_j} \Big|_{\beta_j = \beta_j^c} = n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{l=1}^{pq} k_{jl} \beta_l^c,$$

$$\frac{\partial^2 H(\beta_{-j}^c, \beta_j)}{\partial \beta_j^2} \Big|_{\beta_j = \beta_j^c} = n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj}.$$

Hence,

$$\begin{aligned} H(\beta_{-j}^c, \beta_j) &\approx H(\beta_{-j}^c, \beta_j^c) + (n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{l=1}^{pq} k_{jl} \beta_l^c) (\beta_j - \beta_j^c) \\ &\quad + \frac{1}{2} (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj}) (\beta_j - \beta_j^c)^2. \end{aligned}$$

Moreover,

$$\begin{aligned} Q(\beta_{-j}^c, \beta_j) &\approx (n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{l=1}^{pq} k_{jl} \beta_l^c) (\beta_j - \beta_j^c) \\ &\quad + \frac{1}{2} (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj}) (\beta_j - \beta_j^c)^2 + \lambda_2 |\beta_j| + C, \end{aligned}$$

where C is a constant containing no information about β_j . Denote $(n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{i=1}^{pq} k_{ji} \beta_i^c)(\beta_j - \beta_j^c) + \frac{1}{2} (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})(\beta_j - \beta_j^c)^2 + \lambda_2 |\beta_j|$ by $\tilde{Q}(\beta_j)$. One can update β_j through minimizing $\tilde{Q}(\beta_j)$. Specifically, by

$$\begin{aligned} \frac{\partial \tilde{Q}(\beta_j)}{\partial \beta_j} &= (n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{i=1}^{pq} k_{ji} \beta_i^c) \\ &+ (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})(\beta_j - \beta_j^c) + \lambda_2 \frac{\partial |\beta_j|}{\partial \beta_j} = 0, \end{aligned}$$

where $\frac{\partial |\beta_j|}{\partial \beta_j}$ is the subderivative, that is $\frac{\partial |\beta_j|}{\partial \beta_j} = \text{sign}(\beta_j)$ if $\beta_j \neq 0$ and $\frac{\partial |\beta_j|}{\partial \beta_j} \in [-1, 1]$ otherwise, we have that

$$\begin{aligned} &\beta_j - \beta_j^c + (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} \\ &\cdot (n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{i=1}^{pq} k_{ji} \beta_i^c) = \lambda_2 (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} \frac{\partial |\beta_j|}{\partial \beta_j}. \end{aligned}$$

Consequently, one can update β_j as

$$\beta_j \leftarrow \text{sign}(\Delta_j^c) \left(|\Delta_j^c| - \lambda_2 (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} \right)_+,$$

where $\Delta_j^c = \beta_j^c - (n^{-1} \sum_{i=1}^n p_i^c (1 - p_i^c) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} (n^{-1} \sum_{i=1}^n (p_i^c - 1) Y_i x_{ij} + 2\lambda_1 \sum_{i=1}^{pq} k_{ji} \beta_i^c)$.

We summarize the algorithm as follows.

Coordinate

- Step 1. Initialization. Given initial value β .
- Step 2. For $t = 1, 2, \dots$, update β .
For $j = 1, \dots, p$
 - Compute p_i for $1 \leq i \leq n$;
 - Let $\Delta_j = \beta_j - (n^{-1} \sum_{i=1}^n p_i (1 - p_i) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} (n^{-1} \sum_{i=1}^n (p_i - 1) Y_i x_{ij} + 2\lambda_1 \sum_{i=1}^{pq} k_{ji} \beta_i)$;
 - Update $\beta_j \leftarrow \text{sign}(\Delta_j) \left(|\Delta_j| - \lambda_2 (n^{-1} \sum_{i=1}^n p_i (1 - p_i) x_{ij}^2 + 2\lambda_1 k_{jj})^{-1} \right)_+$;
- Step 3. Repeat Step 2 until convergence.

By the proposed algorithm, we can obtain the solution of (3), which is denoted by $\hat{\beta}$. As the estimator for β , the theoretical properties of $\hat{\beta}$ are studied in the next subsection.

Theoretical properties

In this subsection, we consider the properties of $\hat{\beta}$. A nonasymptotic error bound of $\hat{\beta}$ is given. We assume that the true value β^* is sparse. Let $I^* = \{1 \leq j \leq pq : \beta_j^* \neq 0\}$, and $k^* = \sum_{j=1}^{pq} I(\beta_j^* \neq 0)$ be the cardinality of I^* . For the purpose of theoretical studies, we make the following assumptions.

Assumption 1. Assume that there exists a constant L such that $|x_{ij}| \leq L$ for every $1 \leq i \leq n, 1 \leq j \leq pq$.

Assumption 2. Assume that there exists a constant C such that $\|\beta^*\|_1 \leq C$.

Assumption 3. For the matrix K , assume that there exists a constant C_0 such that $\lambda_{\max}(K) \leq C_0$, where $\lambda_{\max}(K)$ is the largest eigenvalue of K .

Assumption 4. Let $\Sigma = E(X_i X_i^T)$. Define the set $V_{\alpha, \epsilon} = \left\{ \beta \in \mathbb{R}^{pq} : \sum_{j \in I^*} |\beta_j| \leq \alpha \sum_{j \in I^*} |\beta_j| + \epsilon \right\}$ for some α, ϵ . Assume that there exists a constant $0 < b \leq 1$ such that for every $\beta \in V_{\alpha, \epsilon}$

$$P(\beta^T \Sigma \beta \geq b \sum_{j \in I^*} \beta_j^2 - \epsilon) = 1.$$

Assumption 1 makes a common bound L for all x_{ij} with $i = 1, \dots, n, j = 1, \dots, pq$. Assumption 2 gives a bound for $\|\beta^*\|_1$. Combining Assumptions 1 and 2, one can make sure that P_i with $1 \leq i \leq n$ are bounded away from zero and one. P_i equalling zero or one will cause the i -th subject to be either ignorable or dominant in the likelihood function, that is not expected to appear in statistical analysis. This case can be avoided by Assumptions 1 and 2. In Assumption 3, we assume that the largest eigenvalue of K is bounded. Assumption 4 is called Condition Stabil, which can be regarded as a stability requirement on the correlation structure [30]. Under these assumptions, we have the following theorem.

Theorem 1 Assume that Assumptions 1-3 are true and Assumption 4 holds for $\alpha = 5, \epsilon = \frac{\ln 2}{2^d} \times \frac{3}{\lambda_2}$ with $d = \max\{pq, n\}$, let $\lambda_1 = \lambda_2/(6CC_0)$, if

$$\lambda_2 \geq 3 \left(7L \sqrt{\frac{2 \ln(2d)}{n}} + \frac{L}{2d} + 2L \sqrt{\frac{-2 \ln \delta}{n}} \right), \tag{4}$$

then we have that

$$P \left(\|\hat{\beta} - \beta^*\|_1 \leq \frac{3k^* \lambda_2}{sb} + \left(1 + \frac{3s}{\lambda_2} \right) \epsilon \right) > 1 - \delta,$$

where $s = (1 + e^A)^{-4}$ with $A = 8CL$ is a constant.

The proof of Theorem 1 is put in the appendix section. The theorem shows that with a high probability, the L_1 norm of estimate error is bounded by $3k^* \lambda_2/(sb) + (1 + 3s/\lambda_2)\epsilon$. One can see that the term $(1 + 3s/\lambda_2)\epsilon = O(d/2^d)$, which can be negligible for large d . Hence, the term $3k^* \lambda_2/(sb)$ dominates the upper bound, which becomes larger when b becomes smaller. If further assume that $\ln(pq) = o(n)$, by the condition (4) one can see that λ_2 can tend to 0. Further $3k^* \lambda_2/(sb) \rightarrow 0$, that means the upper bound can tend to zero. Consequently, the consistency of $\hat{\beta}$ can be guaranteed.

The selection of tuning parameters

The optimization function (3) contains two tuning parameters λ_1 and λ_2 , which should be determined by some criteria, such as BIC, cross validation method. In our simulation studies, we select the tuning parameters by a validation set. And in real data analysis, the cross validation method is used. Before applying these methods, one should firstly determine the value range of tuning parameters. Specifically, we here make a transformation of λ_1 and λ_2 . Let $\lambda = \lambda_1 + \lambda_2$, and $\alpha = \lambda_2/\lambda$. Then the penalty terms in (3) can be rewritten as $\lambda(\alpha \|\beta\|_1 + (1 - \alpha)\beta^T K\beta)$. Because $\alpha \in [0, 1]$, the alternative values of α are set as 0.02κ for $\kappa = 1, \dots, 50$. With a

given α , we denote λ_0 as the threshold value. Once $\lambda \geq \lambda_0$, the solution of (3) is exactly zero. By

$$\frac{\partial Q(\beta)}{\partial \beta} = \frac{1}{n} \sum_{i=1}^n \frac{e^{-Y_i \beta^T X_i}}{1 + e^{-Y_i \beta^T X_i}} (-Y_i X_i) + 2\lambda(1 - \alpha)K\beta + \lambda\alpha \frac{\partial \|\beta\|_1}{\beta} = 0,$$

one can see that once $\beta = 0$ is the solution, every element of $1/(2\lambda\alpha n) \sum_{i=1}^n (-Y_i X_i)$ belongs to $[-1, 1]$. This means that $\lambda_0 = 1/(2\alpha n) \|\sum_{i=1}^n (Y_i X_i)\|_\infty$. Following the idea of [14], the alternative values of λ are set as 0.001 and $0.96^v \lambda_0$ for $v = 0, 1, \dots, 160$. For the validation set method, the prediction error on the validation set of our approach with tuning parameters α, λ is denoted by $PE_{(\alpha, \lambda)}$. The final α, λ are selected as the minimizer of $PE_{(\alpha, \lambda)}$.

For the M -fold cross validation method, the data are randomly divided into M folds of approximately equal size. For $m = 1, \dots, M$, we treat the m fold as the validation set, and fit the model with tuning parameters α, λ on the remaining $M - 1$ folds. The corresponding prediction error on the validation set is denoted by $PE_{(\alpha, \lambda)}^{(m)}$ and the cross validation prediction error is defined as

$$PE_{(\alpha, \lambda)}^{(cv)} = \frac{1}{M} \sum_{m=1}^M PE_{(\alpha, \lambda)}^{(m)}.$$

The α, λ are selected as the minimizer of the cross validation prediction error [6].

Numerical studies

In this section, we evaluate the performance of our proposed method by two simulated examples and a real data analysis. For the purpose of comparison, we also consider the logistic regression model with L_1 penalty [12, 13], the logistic regression model with fused lasso penalty, and linear support vector machine, which are denoted by LG- L_1 , LG-fused, and Linear SVM respectively for convenience. Meanwhile, our proposed method is abbreviated as LG-sob.

Simulation studies

Example 1. We generate data from the following model

$$\log \frac{P(Y_i = +1 | \mathbf{X}_i)}{P(Y_i = -1 | \mathbf{X}_i)} = \langle \mathbf{X}_i, B_0 \rangle, \quad i = 1, \dots, n,$$

where \mathbf{X}_i and B_0 both belong to $\mathbb{R}^{32 \times 32}$. One result caused by image covariates is that the corresponding regression coefficient can be treated as a image too. Hence we here just treat B_0 as images, while \mathbf{X}_i is generated from a multivariate normal distribution. Specifically, we define the vectorization of \mathbf{X}_i as X_i , and X_i is generated from a multivariate normal distribution with mean 0 and covariance $\text{cov}(x_{j_1}, x_{j_2}) = 0.5^{|j_1 - j_2|}$ for any $1 \leq j_1, j_2 \leq 1024$. The parameter image B_0 is considered in two cases, which have been shown in Fig 2. The first case of B_0 denoted by B_{01} is a bird picture, in which the blue region takes value 0, and the yellow region takes value 1. The other case of B_0 denoted by B_{02} is a butterfly picture, which is more complicated and takes values in interval $[-0.0197, 0.0628]$. Given \mathbf{X}_i and B_0 , the response Y_i is generated from a two-point distribution $P(Y_i = +1 | \mathbf{X}_i) = 1 - P(Y_i = -1 | \mathbf{X}_i) = 1/(1 + e^{-\langle \mathbf{X}_i, B_0 \rangle})$.

Example 2. In this example, the mechanism of data generation is similar to that for Example 1, the only difference is that we generate \mathbf{X}_i in a more complex way. In particular, we follow the simulation scheme of [22] and generate \mathbf{X}_i from a 32×32 phantom map with 1024 pixels

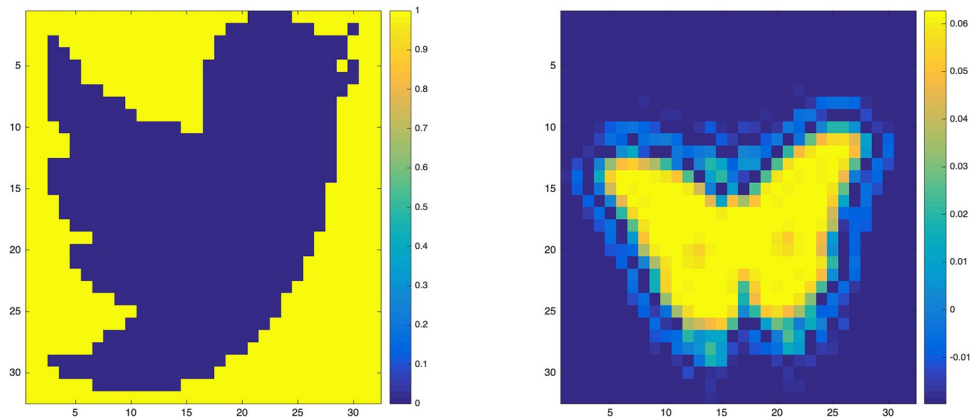


Fig 2. Simulated example. The true parameter images B_0 .

<https://doi.org/10.1371/journal.pone.0234975.g002>

according to a spatially correlated random process $X_i = \sum_l l^{-1} \eta_{il} \varphi_l$, among which the η_{il} are standard normal random variables and the φ_l are bivariate Haar wavelet basis functions.

For these two simulated examples, along with the training set with sample size n , we also generate a validation set and a test set with sample sizes both being 500. We train the model on the training set, select tuning parameters through the validation set, and calculate the classification accuracy on the test set to evaluate the performance of the model.

For every specification of the parameter B_0 and sample size n , we replicate the simulation 100 times for each example, and the average prediction errors are computed and summarized in Table 1 for Example 1, and Table 2 for Example 2 respectively. Besides the prediction errors, we also calculate the average estimation errors $\sum_{i=1}^{100} \|\hat{B}_i - B_0\|^2 / 100$ for LG-sob, LG- L_1 and LG-fused, where \hat{B}_i is the parameter image estimator in the i -th time. From the results, one can see that our proposed method performs better than the other three methods in all cases from the prediction perspective. As sample size n becomes larger, the prediction errors will become smaller, but the estimation errors do not decrease congruously. The reason may be that the tuning parameters are selected based on minimization of prediction error.

Table 1. Results of simulated example 1: Prediction error (PE) and estimation error (EE).

(n, B_0)	$(500, B_{01})$		$(1000, B_{01})$		$(500, B_{02})$		$(1000, B_{02})$	
	PE	EE	PE	EE	PE	EE	PE	EE
LG-sob	0.099	337.645	0.075	336.173	0.107	8.153	0.080	13.505
LG- L_1	0.272	404.589	0.199	375.926	0.272	17.354	0.204	23.143
LG-fused	0.248	423.242	0.190	406.866	0.248	7.016	0.190	9.400
Linear SVM	0.221	NA	0.174	NA	0.223	NA	0.172	NA

<https://doi.org/10.1371/journal.pone.0234975.t001>

Table 2. Results of simulated example 2: Prediction error (PE) and estimation error (EE).

(n, B_0)	$(500, B_{01})$		$(1000, B_{01})$		$(500, B_{02})$		$(1000, B_{02})$	
	PE	EE	PE	EE	PE	EE	PE	EE
LG-sob	0.028	181.06	0.023	176.19	0.049	582.94	0.038	1096.1
LG- L_1	0.073	1712.6	0.058	1501.7	0.116	2190.9	0.090	3022.4
LG-fused	0.052	438.57	0.044	495.73	0.097	482.19	0.076	775.99
Linear SVM	0.050	NA	0.038	NA	0.084	NA	0.071	NA

<https://doi.org/10.1371/journal.pone.0234975.t002>

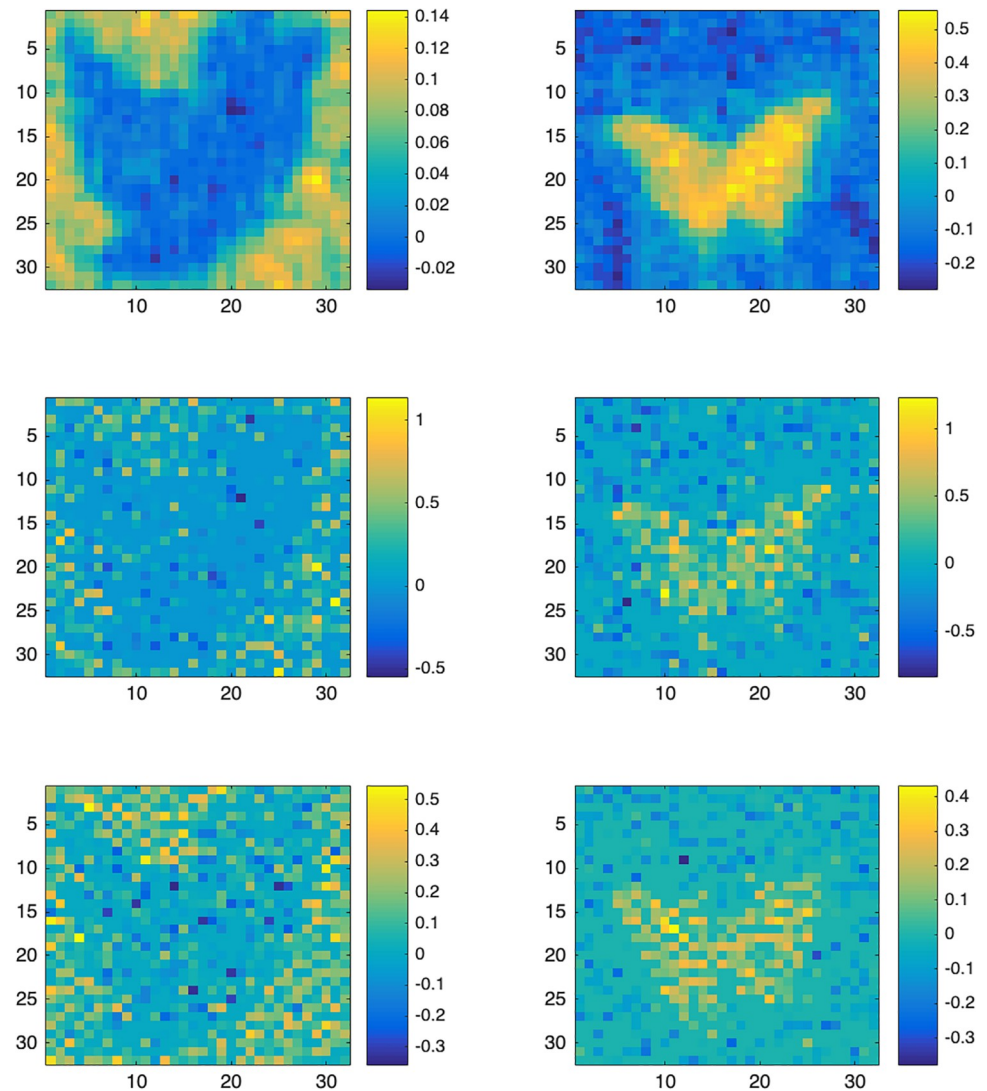


Fig 3. Simulated example. One of randomly selected parameter images estimations. The first row is the results of our proposed LG-sob, the second row is the results of LG- L_1 , the third row is the results of LG-fused.

<https://doi.org/10.1371/journal.pone.0234975.g003>

Moreover, we also randomly select one simulated result from the 100 replications of Example 1, and show the parameter image estimations in Fig 3. One can see that our proposed LG-sob method can capture the shapes of images, but LG- L_1 and LG-fused do not have this property.

A real data analysis

The classification of the ZIP Code Dataset is a benchmark problem in machine learning community [6]. One can obtain the ZIP Code Dataset from the following website https://web.stanford.edu/~hastie/StatLearnSparsity_files/DATA/zipcode.html [28]. The Dataset contains normalized handwritten digits, which are automatically scanned from envelopes by the U.S. Postal Service. Every observation is a handwritten digit, and is represented as a size normalized 16×16 grayscale image [31]. The purpose is to use the 256 pixel values to predict the corresponding digit. The Dataset contains a training set with 7291 observations and a test set with

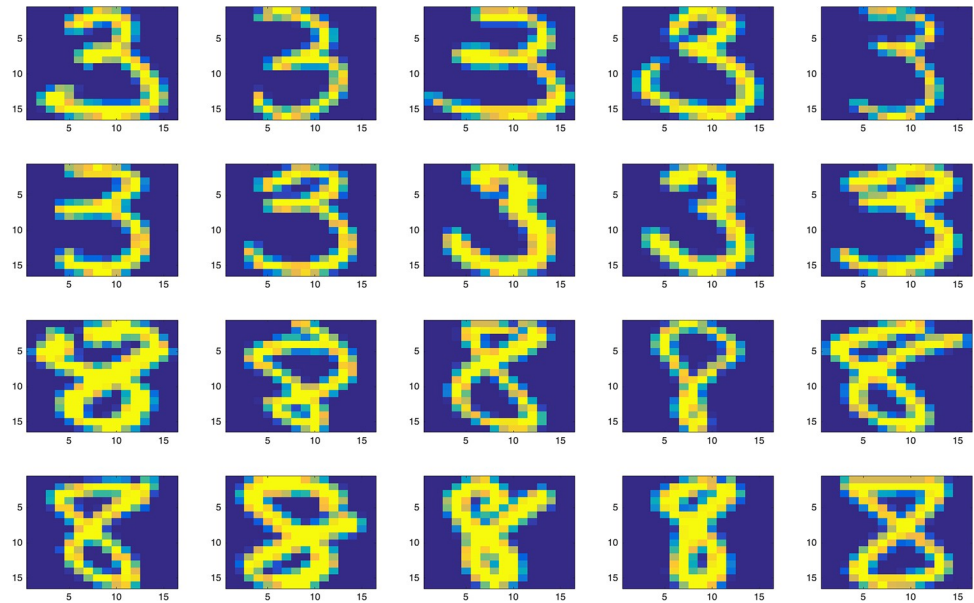


Fig 4. Real data analysis. Some examples of handwritten 3's and 8's.

<https://doi.org/10.1371/journal.pone.0234975.g004>

2007 observations. Because this article only considers the binary response prediction by logistic regression models, and it looks like that numbers 3 and 8 have more similar characteristics, hence we only consider handwritten 3's and 8's in this paper. The sizes of handwritten 3's and 8's are 658 and 542 respectively in the train set, while they are both 166 in the test set. Fig 4 shows some examples of handwritten 3's and 8's.

More specifically, we denote the i -th observation by $\mathbf{X}_i \in \mathbb{R}^{16 \times 16}$, and define the corresponding class label $Y_i = -1$ if \mathbf{X}_i represents handwritten 3 and $Y_i = +1$ if \mathbf{X}_i represents handwritten 8. Our proposed method is applied to construct the classifier for the prediction of Y_i (i.e. handwritten numeral) based on the grayscale image \mathbf{X}_i . We train the model on the training set, and evaluate the performance of the proposed method on the test set by classification accuracy. For the purpose of comparison, we also consider the logistic regression model with only L_1 penalty.

The tuning parameters are selected by 10-fold cross validation (CV) method. The CV prediction errors in various parameters setting are calculated and plotted in Fig 5. Finally, our proposed method selects the tuning parameters as $\alpha = 0.04$, $\lambda = 0.0118$, while the method with L_1 penalty selects the tuning parameter as $\lambda = 0.0014$. The parameter image estimations of the two methods are shown in Fig 6. One can see that our proposed method tends to make adjacent pixels have similar effects on the model. Meanwhile, LG- L_1 tries to obtain a more sparse parameter estimation, and LG-fused method tries to make pixels only adjacent in the vertical direction have similar effects. The top-left region of parameter image has positive effects on handwritten numeral 8, and the bottom-right region has positive effects on handwritten numeral 3. The classification accuracy on the test set of our proposed method is 96.99%, while the accuracies of LG- L_1 , LG-fused, and Linear SVM are 96.39%, 96.08% and 96.39%, respectively. The proposed method performs better.

Conclusion

We have developed a novel estimation method for logistic regression with image covariates. This method can not only perform variable selection, but also capture the shape features of

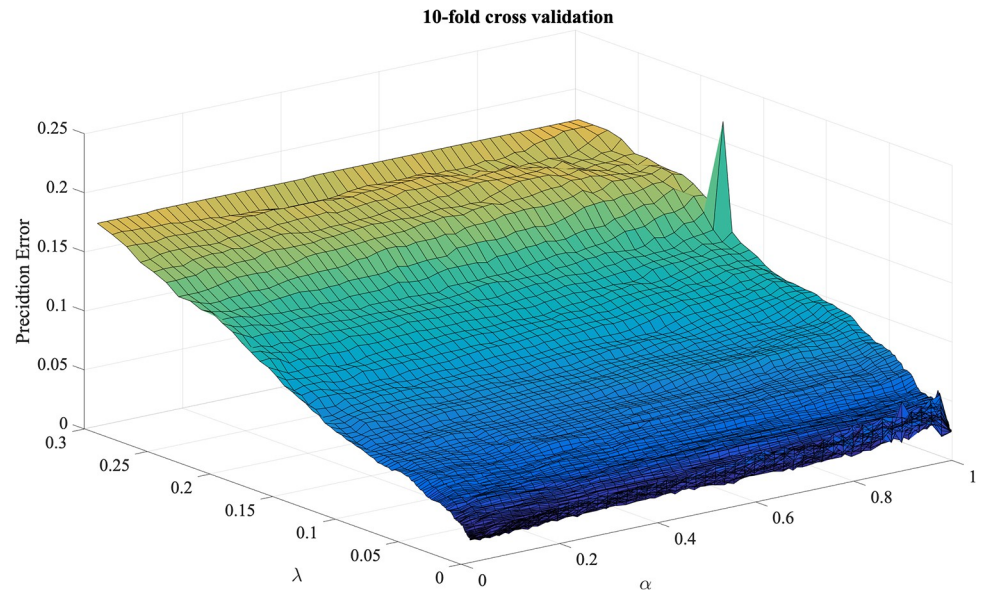


Fig 5. Real data analysis. The results of 10-fold CV: Prediction error in various parameters settings.

<https://doi.org/10.1371/journal.pone.0234975.g005>

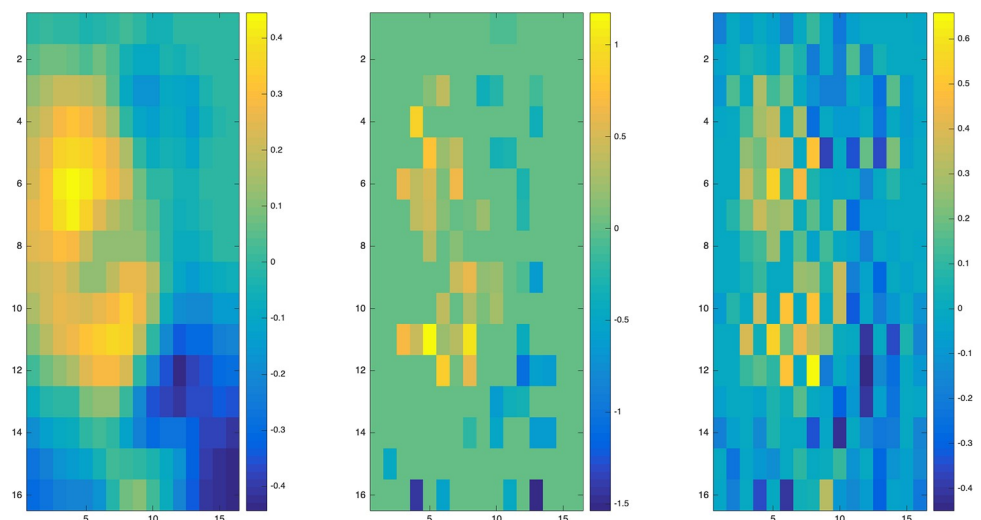


Fig 6. Real data analysis. The parameter image estimations. Left: the estimation of our proposed LG-sob; Middle: the estimation of LG- L_1 ; Right: the estimation of LG-fused.

<https://doi.org/10.1371/journal.pone.0234975.g006>

the parameter images. Both theoretical results and numerical studies show that our method performs well. We have proposed a coordinated descent algorithm to solve the optimization problem, and the global convergence of the algorithm is guaranteed. However, as pointed out by one referee, the coordinated descent algorithm is very time consuming, especially in the case of high dimensional image covariates. Many more efficient optimization approaches, such as Nesterov accelerated gradient methods [32], interior-point methods [13], may be more suitable. We will research this issue in future. Furthermore, our method is mainly based on Sobolev norm regularization, compared to which total variation

regularization is more sensitive to capture sharp edges and jumps of parameter images. However, the algorithm of total variation regularization based estimation method is more complex, which can be our future work to study.

Appendix: Proof of Theorem 1

Before giving the proof of Theorem 1, we first list the bounded differences inequality as the following lemma without proof.

Lemma 1 (the Bounded Differences Inequality) *Suppose that $X_1, \dots, X_n \in \mathcal{H}$ are independent, and the function $f : \mathcal{H}^n \rightarrow \mathbb{R}$ satisfies the bounded difference assumption*

$$\sup_{x_1, \dots, x_n, x'_i \in \mathcal{H}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i,$$

for $i = 1, \dots, n$. Then for all $t > 0$,

$$P(f - E(f) \geq t) \leq e^{-2t^2 / \sum_{i=1}^n c_i^2}.$$

For more details about Lemma 1 and its proof, one can refer to [33]. The following is the proof of Theorem 1.

Proof of Theorem 1. By the definitions of $\hat{\beta}$ and β^* , one can see that

$$\mathbb{P}I(\hat{\beta}) \geq \mathbb{P}I(\beta^*),$$

and

$$\mathbb{P}_n I(\hat{\beta}) + \lambda_1 \hat{\beta}^T K \hat{\beta} + \lambda_2 \|\hat{\beta}\|_1 \leq \mathbb{P}_n I(\beta^*) + \lambda_1 \beta^{*T} K \beta^* + \lambda_2 \|\beta^*\|_1.$$

Hence, we have that

$$\begin{aligned} 0 &\leq \mathbb{P}I(\hat{\beta}) - \mathbb{P}I(\beta^*) \\ &\leq (\mathbb{P}_n - \mathbb{P})(I(\beta^*) - I(\hat{\beta})) + \lambda_1(\beta^{*T} K \beta^* - \hat{\beta}^T K \hat{\beta}) + \lambda_2(\|\beta^*\|_1 - \|\hat{\beta}\|_1). \end{aligned} \tag{5}$$

Moreover,

$$\begin{aligned} &\lambda_2/3 \|\hat{\beta} - \beta^*\|_1 \\ &\leq \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 + \lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) \\ &\leq \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 + \lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) + \mathbb{P}I(\hat{\beta}) - \mathbb{P}I(\beta^*) \\ &\leq \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 + \lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) + \lambda_1(\beta^{*T} K \beta^* - \hat{\beta}^T K \hat{\beta}) \\ &\quad + (\mathbb{P}_n - \mathbb{P})(I(\beta^*) - I(\hat{\beta})) + \lambda_2(\|\beta^*\|_1 - \|\hat{\beta}\|_1). \end{aligned} \tag{6}$$

We first consider the term $(\mathbb{P}_n - \mathbb{P})(I(\beta^*) - I(\hat{\beta}))$. Specifically, define $L_n = \sup_{\beta} \frac{(\mathbb{P}_n - \mathbb{P})(I(\beta^*) - I(\beta))}{\|\beta - \beta^*\|_1 + \epsilon}$.

Let $l(\beta; Y_i, X_i) = \log(1 + e^{-Y_i \beta^T X_i})$, and

$$\mathbb{P}_n I(\beta) = \frac{1}{n} \left(\sum_{i=1, i \neq l}^n l(\beta; Y_i, X_i) + l(\beta; Y'_l, X'_l) \right),$$

which is the empirical measure corresponding to replacing (Y_i, X_i) by (Y'_i, X'_i) . Then

$$\begin{aligned} & \frac{(\mathbb{P}_n - \mathbb{P})(l(\beta^*) - l(\beta))}{\|\beta - \beta^*\|_1 + \epsilon} - \frac{(\mathbb{P}_n - \mathbb{P})(l(\beta^*) - l(\beta))}{\|\beta - \beta^*\|_1 + \epsilon} \\ &= \frac{1}{n} \frac{l(\beta^*; Y_i, X_i) - l(\beta; Y_i, X_i) - l(\beta^*; Y'_i, X'_i) + l(\beta; Y'_i, X'_i)}{\|\beta - \beta^*\|_1 + \epsilon} \\ &\leq \frac{4L}{n} \frac{\|\beta - \beta^*\|_1}{\|\beta - \beta^*\|_1 + \epsilon} \leq \frac{4L}{n}, \end{aligned}$$

among which the inequality is obtained by a first order Taylor expansion and the assumption 1. Then by Lemma 1, we can obtain that

$$P(L_n - E(L_n) \geq u) \leq \exp\left\{-\frac{nu^2}{8L^2}\right\}.$$

Let $\delta = \exp\left\{-\frac{nu^2}{8L^2}\right\}$, then we have that $u = 2L\sqrt{\frac{-2\ln \delta}{n}}$, and $P(L_n - E(L_n) \geq u) \leq \delta$.

Let $d = \max\{pq, n\}$. Taking $\epsilon = \frac{\ln 2}{2^d} \times \frac{3}{\lambda_2}$, by the lemma 3 of [34] with $C_\varphi = 1, C_F = L$, we have

$$E(L_n) \leq 7L\sqrt{\frac{2 \ln(2d)}{n}} + \frac{L}{2d}.$$

Consequently, we have that

$$P(L_n \leq 7L\sqrt{\frac{2 \ln(2d)}{n}} + \frac{L}{2d} + 2L\sqrt{\frac{-2 \ln \delta}{n}}) \geq 1 - \delta.$$

By the condition of Theorem 1, we know $\lambda_2 \geq 3\left(7L\sqrt{\frac{2 \ln(2d)}{n}} + \frac{L}{2d} + 2L\sqrt{\frac{-2 \ln \delta}{n}}\right)$, hence $P(L_n \leq \lambda_2/3) \geq 1 - \delta$.

On the event $\{L_n \leq \lambda_2/3\}$, we have that

$$(\mathbb{P}_n - \mathbb{P})(l(\beta^*) - l(\hat{\beta})) \leq \frac{\lambda_2}{3}(\|\beta - \beta^*\|_1 + \epsilon). \tag{7}$$

Secondly, we consider the term $\lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) + \lambda_1(\beta^{*T} K \beta^* - \hat{\beta}^T K \hat{\beta})$. Based on Assumptions 2 and 3, one can see that

$$\begin{aligned} & \lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) + \lambda_1(\beta^{*T} K \beta^* - \hat{\beta}^T K \hat{\beta}) \\ &= 2\lambda_1 \beta^{*T} K(\beta^* - \hat{\beta}) \\ &\leq 2\lambda_1 \lambda_{\max}(K) \|\beta^*\|_2 \|\hat{\beta} - \beta^*\|_2 \\ &\leq 2\lambda_1 \lambda_{\max}(K) \|\beta^*\|_1 \|\hat{\beta} - \beta^*\|_1 \\ &\leq 2\lambda_1 C C_0 \|\hat{\beta} - \beta^*\|_1. \end{aligned}$$

One can see that if $\lambda_1 = \lambda_2/(6CC_0)$, we have

$$\lambda_1(\hat{\beta} - \beta^*)^T K(\hat{\beta} - \beta^*) + \lambda_1(\beta^{*T} K \beta^* - \hat{\beta}^T K \hat{\beta}) \leq \lambda_2/3 \|\hat{\beta} - \beta^*\|_1. \tag{8}$$

Consequently, on the event $\{L_n \leq \lambda_2/3\}$ we combine (6), (7), (8), and obtain

$$\begin{aligned} \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 &\leq \lambda_2 \|\hat{\beta} - \beta^*\|_1 + \lambda_2 (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \lambda_2/3\epsilon \\ &\leq \lambda_2 (\|\hat{\beta}\|_1 + \|\beta^*\|_1) + \lambda_2 (\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \lambda_2/3\epsilon \\ &\leq 2\lambda_2 \|\beta^*\|_1 + \lambda_2/3\epsilon. \end{aligned} \tag{9}$$

Hence we have that

$$\|\hat{\beta} - \beta^*\|_1 \leq 6\|\beta^*\|_1 + \epsilon \leq 7C. \tag{10}$$

By (9) one can also obtain that

$$\begin{aligned} \|\hat{\beta} - \beta^*\|_1 &\leq 3\|\hat{\beta} - \beta^*\|_1 + 3(\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \epsilon \\ &= 3\left(\sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \sum_{j \notin I^*} |\hat{\beta}_j| + \sum_{j \in I^*} |\beta_j^*| - \sum_{j=1}^{pq} |\hat{\beta}_j|\right) + \epsilon \\ &= 3\left(\sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \sum_{j \in I^*} |\beta_j^*| - \sum_{j \in I^*} |\hat{\beta}_j|\right) + \epsilon \\ &\leq 6\sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \epsilon. \end{aligned}$$

Consequently, we have $\sum_{j \notin I^*} |\hat{\beta}_j - \beta_j^*| \leq 5\sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \epsilon$. This means that $\hat{\beta} - \beta^* \in V_{5,\epsilon}$.

By the example 4.5 in [35], we have that $\mathbb{P}l(\hat{\beta}) - \mathbb{P}l(\beta^*) \geq E_X(P(\hat{\beta}) - P(\beta^*))^2$, where $P(\beta) = 1/(1 + e^{-X^T \beta})$ and $E_X(\cdot)$ is the expectation with respect to the distribution of X . Using Taylor expansion, one can obtain that

$$P(\hat{\beta}) - P(\beta^*) = \frac{e^{\tilde{\beta}^T X}}{(1 + e^{\tilde{\beta}^T X})^2} X^T (\hat{\beta} - \beta^*),$$

where $\tilde{\beta} = \tau \hat{\beta} + (1 - \tau)\beta^*$ for some $\tau \in (0, 1)$. Moreover, by (10) and Assumptions 1-2, we have $\tilde{\beta}^T X \leq \|\tilde{\beta}\|_1 L \leq (\tau \|\hat{\beta} - \beta^*\|_1 + \|\beta^*\|_1)L \leq 8CL$. This means that

$$(P(\hat{\beta}) - P(\beta^*))^2 \geq s(\hat{\beta} - \beta^*)^T X X^T (\hat{\beta} - \beta^*),$$

where $s = (1 + e^A)^{-4}$ and $A = 8CL$, then by Assumption 4 we have that

$$\mathbb{P}l(\hat{\beta}) - \mathbb{P}l(\beta^*) \geq E_X(P(\hat{\beta}) - P(\beta^*))^2 \geq s(\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*) \geq sb \sum_{j \in I^*} (\hat{\beta}_j - \beta_j^*)^2 - s\epsilon. \tag{11}$$

Furthermore, we have

$$\begin{aligned}
 & \frac{\lambda_2}{3} \|\hat{\beta} - \beta^*\|_1 + sb \sum_{j \in I^*} (\hat{\beta}_j - \beta_j^*)^2 - s\epsilon \\
 \leq & \frac{\lambda_2}{3} \|\hat{\beta} - \beta^*\|_1 + \mathbb{P}l(\hat{\beta}) - \mathbb{P}l(\beta^*) \\
 \leq & \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 + \lambda_1(\beta^{*T}K\beta^* - \hat{\beta}^TK\hat{\beta}) + (\mathbb{P}_n - \mathbb{P})(l(\beta^*) - l(\hat{\beta})) \\
 & + \lambda_2(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \\
 \leq & \lambda_2/3 \|\hat{\beta} - \beta^*\|_1 + \lambda_1(\hat{\beta} - \beta^*)^TK(\hat{\beta} - \beta^*) + \lambda_1(\beta^{*T}K\beta^* - \hat{\beta}^TK\hat{\beta}) \\
 & + (\mathbb{P}_n - \mathbb{P})(l(\beta^*) - l(\hat{\beta})) + \lambda_2(\|\beta^*\|_1 - \|\hat{\beta}\|_1) \\
 \leq & \lambda_2 \|\hat{\beta} - \beta^*\|_1 + \lambda_2(\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \lambda_2/3\epsilon \\
 = & \lambda_2(\sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \sum_{j \notin I^*} |\hat{\beta}_j|) + \lambda_2(\sum_{j \in I^*} |\beta_j^*| - \sum_{j \in I^*} |\hat{\beta}_j| - \sum_{j \notin I^*} |\hat{\beta}_j|) \\
 & + \lambda_2/3\epsilon \\
 = & \lambda_2 \sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \lambda_2(\sum_{j \in I^*} |\beta_j^*| - \sum_{j \in I^*} |\hat{\beta}_j|) + \lambda_2/3\epsilon \\
 \leq & 2\lambda_2 \sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + \lambda_2/3\epsilon,
 \end{aligned}$$

where the first inequality follows by (11), the second inequality follows by (5), and the fourth inequality is obtained by combining the results of (7) and (8). Consequently, we have

$$\begin{aligned}
 \|\hat{\beta} - \beta^*\|_1 + \frac{3}{\lambda_2} sb \sum_{j \in I^*} (\hat{\beta}_j - \beta_j^*)^2 & \leq 6 \sum_{j \in I^*} |\hat{\beta}_j - \beta_j^*| + (1 + \frac{3s}{\lambda_2})\epsilon \\
 & \leq 9ak^* + \frac{1}{a} \sum_{j \in I^*} (\hat{\beta}_j - \beta_j^*)^2 + (1 + \frac{3s}{\lambda_2})\epsilon,
 \end{aligned}$$

where a is a positive constant and the second inequality follows by

$$6|\hat{\beta}_j - \beta_j^*| = 2 \cdot 3a^{1/2} \cdot a^{-1/2}|\hat{\beta}_j - \beta_j^*| \leq 6a + (\hat{\beta}_j - \beta_j^*)^2/a.$$

Let $a = \lambda_2/(3sb)$, then

$$\|\hat{\beta} - \beta^*\|_1 \leq \frac{3k^*\lambda_2}{sb} + (1 + \frac{3s}{\lambda_2})\epsilon.$$

This completes the proof of the Theorem.

Acknowledgments

We thank the Editor, the AE and four referees for their helpful comments and valuable suggestions, which make the article have a greatly improvement.

Author Contributions

Formal analysis: Baiguo An.

Funding acquisition: Baiguo An.

Investigation: Baiguo An.

Methodology: Baiguo An, Beibei Zhang.

Software: Baiguo An, Beibei Zhang.

Validation: Baiguo An.

Writing – original draft: Baiguo An, Beibei Zhang.

Writing – review & editing: Baiguo An, Beibei Zhang.

References

1. Bishop CM. Pattern Recognition and Machine Learning. Springer; 2007.
2. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press; 2016.
3. Cox DR. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society Series B (Methodological)*. 1958; 20(2):215–242. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>
4. Albert A, Anderson JA. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*. 1984; 71(1):1–10. <https://doi.org/10.1093/biomet/71.1.1>
5. Santner TJ, Duffy DE. A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*. 1986; 73(3):755–758. <https://doi.org/10.1093/biomet/73.3.755>
6. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd ed. Springer; 2009. Available from: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.
7. Silvapulle MJ. On the Existence of Maximum Likelihood Estimators for the Binomial Response Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1981; 43(3):310–313.
8. Sun Y. Regularization in High-dimensional Statistics. PhD dissertation stanford university. 2015.
9. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1996; 58(1):267–288.
10. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*. 2001; 96(456):1348–1360. <https://doi.org/10.1198/016214501753382273>
11. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. 2010; 38(2):894–942. <https://doi.org/10.1214/09-AOS729>
12. Park MY, Hastie T. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69(4):659–677. <https://doi.org/10.1111/j.1467-9868.2007.00607.x>
13. Koh K, Kim S, Boyd SP. An Interior-Point Method for Large-Scale l_1 -Regularized Logistic Regression. *Journal of Machine Learning Research*. 2007; 8:1519–1555.
14. Meier L, De Geer SV, Bühlmann P. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2008; 70(1):53–71. <https://doi.org/10.1111/j.1467-9868.2007.00627.x>
15. Liang Y, Liu C, Luan X, Leung K, Chan T, Xu Z, et al. Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. *BMC Bioinformatics*. 2013; 14:198. <https://doi.org/10.1186/1471-2105-14-198> PMID: 23777239
16. Polzehl J, Spokoiny V. Propagation-Separation Approach for Local Likelihood Estimation. *Probability Theory and Related Fields*. 2006; 135(3):335–362. <https://doi.org/10.1007/s00440-005-0464-1>
17. Zhu H, Fan J, Kong L. Spatially Varying Coefficient Model for Neuroimaging Data With Jump Discontinuities. *Journal of the American Statistical Association*. 2014; 109(507):1084–1098. <https://doi.org/10.1080/01621459.2014.881742> PMID: 25435598
18. Guo R, Ahn M, Zhu H. Spatially Weighted Principal Component Analysis for Imaging Classification. *Journal of Computational and Graphical Statistics*. 2015; 24(1):274–296. <https://doi.org/10.1080/10618600.2014.912135> PMID: 26089629
19. Polzehl J, Spokoiny V. Adaptive Weights Smoothing with Applications to Image Restoration. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2000; 62(2):335–354. <https://doi.org/10.1111/1467-9868.00235>
20. Li Y, Zhu H, Shen D, Lin W, Gilmore JH, Ibrahim JG. Multiscale adaptive regression models for neuroimaging data. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2011; 73(4):559–578. <https://doi.org/10.1111/j.1467-9868.2010.00767.x>

21. Ariascastro E, Salmon J, Willett R. Oracle inequalities and minimax rates for non-local means and related adaptive kernel-based methods. *SIAM Journal on Imaging Sciences*. 2012; 5(3):944–992. <https://doi.org/10.1137/110859403>
22. Wang X, Zhu H, for the Alzheimer's Disease Neuroimaging Initiative. Generalized Scalar-on-Image Regression Models via Total Variation. *Journal of the American Statistical Association*. 2017; 112(519):1156–1168. <https://doi.org/10.1080/01621459.2016.1194846> PMID: 29151658
23. Hu W, Shen W, Zhou H, Kong D. Matrix Linear Discriminant Analysis. *Technometrics*. 2019; 0(0):1–10.
24. Peyré G. Denoising by Sobolev and Total Variation Regularization. http://www.numerical-tours.com/matlab/denoisingsimp_4_denoiseregul/; 2019.
25. Qiu P. *Image Processing and Jump Regression Analysis*. Wiley-Interscience; 2005.
26. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2005; 67(2):301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
27. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*. 2005; 67(1):91–108. <https://doi.org/10.1111/j.1467-9868.2005.00490.x>
28. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC; 2015.
29. Tseng P. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications*. 2001; 109(3):475–494. <https://doi.org/10.1023/A:1017501703105>
30. Bunea F. Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics*. 2008; 2:1153–1194. <https://doi.org/10.1214/08-EJS287>
31. Lecun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*. 1989; p. 396–404.
32. Yu N. Gradient methods for minimizing composite functions. *Mathematical Programming*. 2013; 140(1):125–161. <https://doi.org/10.1007/s10107-012-0629-5>
33. Devroye L, Lugosi G. *Combinatorial Methods in Density Estimation*. Springer; 2001.
34. Wegkamp M. Lasso type classifiers with a reject option. *Electronic Journal of Statistics*. 2007; 1(3):155–168. <https://doi.org/10.1214/07-EJS058>
35. Steinwart I. How to Compare Different Loss Functions and Their Risks. *Constructive Approximation*. 2007; 26(2):225–287. <https://doi.org/10.1007/s00365-006-0662-3>