SELECTION OF DECISION BOUNDARIES FOR LOGISTIC REGRESSION

Matsui, Hidetoshi
Faculty of Mathematics, Kyushu University

https://doi.org/10.5109/1909526

出版情報：Bulletin of informatics and cybernetics. 47, pp.83-95, 2015-12. 統計科学研究会
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 47, 2015

SELECTION OF DECISION BOUNDARIES FOR LOGISTIC REGRESSION

By

Hidetoshi Matsui*

Abstract

We propose a method that selects decision boundaries for the logistic regression model by applying sparse regularization. We can investigate which decision boundaries are truly necessary for the multinomial logistic regression model by letting some of the coefficient parameters or the differences between them approach zero. The model is estimated by the maximum penalized likelihood method with a fused lasso-type penalty. We also introduce various model selection criteria for evaluating models estimated by the penalized likelihood method. Simulation studies are conducted in order to evaluate the effectiveness of the proposed method. Real data analysis provides new insights into how each of the predictors contributes to the classification.

Key Words and Phrases: Classification, Fused lasso, Logistic regression model, Model selection, Regularization

1. Introduction

The least absolute shrinkage and selection operator (lasso) was proposed by Tibshirani (1996), and it can simultaneously estimate models and select variables in linear regression models by imposing an L_1 penalty on the parameters. Because of the usefulness of the lasso, it has been widely applied in various fields of statistical science and machine learning (see, e.g., Bühlmann and van de Geer, 2011; Hastie et al., 2015), and variations or refinements of the lasso have been proposed (Frank and Friedman, 1993; Fan and Li, 2001; Zou and Hastie, 2005; Zhang, 2010).

Such sparse regularization methods may be used to select variables which affect classification procedures, such as support vector machines (Zhu et al., 2004; Wang and Shen, 2006) or Fisher's discriminant analysis (Witten and Tibshirani, 2011). The aim of this paper is to introduce multinomial logistic regression modeling into classification problems, via sparse regularization. The logistic regression model is a useful classification tool, since it provides posterior probabilities about the group to which the data belong (see, e.g., McCullagh and Nelder, 1989). It can be easily extended to classification problems of three or more groups by introducing dummy variables which follow a multinomial distribution. The use of sparse regularization for variable selection problems for the logistic regression model has been discussed in Krishnapuram et al. (2005), Park and Hastie (2007), Meier et al. (2008), and Friedman et al. (2010). However, if we impose an L_1 penalty directly on the multinomial logistic regression model, there may

* Faculty of Mathematics, Kyushu University, 744 Motooka, Nishi-Ku, Fukuoka 819-0395, Japan. tel +81–92–802–4444 hmatsui@math.kyushu-u.ac.jp

be difficulties in interpreting the results. Although we can obtain coefficients, some of which approach zero, these results are inadequate from the viewpoint of variable selection. In order to solve this problem, the R package `glmnet` uses a group lasso (Yuan and Lin, 2006) to combine multiple coefficients for each variable and then treat them as a grouped parameter. Furthermore, Matsui (2014) extended this idea to the logistic regression model for functional data.

We can also consider this problem as one of selecting the decision boundaries rather than variables, that is, we can investigate which decision boundaries are truly necessary for each variable, using the technique of sparse regularization. When we have L classes, the logistic regression model represents the relationship between linear predictors and the log odds of the two classes l $(< L)$ and L, by considering the class L as the reference class. Therefore, we can remove the decision boundaries between classes l and L by allowing the corresponding coefficients to shrink toward zero. However, a lasso-type penalty is not sufficient for selecting the boundaries between l and l' $(< L)$.

We propose a method for estimating and selecting decision boundaries simultaneously using an L_1-type penalty. We apply a fused lasso-type penalty (Tibshirani et al., 2005), which embeds first-order differences into the L_1 penalty so that neighboring coefficients are forced to have the same value. Tibshirani and Taylor (2011) extended the fused lasso such that it could include second- and higher-order differences. Related penalties that consider the correlation of predictors are discussed in Bondell and Reich (2008), Daye and Jeng (2009), Tutz and Ulbricht (2009), and Lin et al. (2013). Furthermore, She (2010) extended the fused lasso and then proposed a clustered lasso which penalizes the differences between all combinations of coefficients in a linear regression model. Jang et al. (2013) applied a penalty similar to the clustered lasso, for the clustering of highly correlated variables. We consider the application of the clustered lasso to the logistic regression model. Due to the effect of the proposed penalty, we can select multiple decision boundaries at one time. The model is estimated by the maximum penalized likelihood method with a fused lasso-type penalty. Since it is difficult to analytically derive an estimator of the model, we used the iterative procedure proposed by Ulbricht (2010). Furthermore, we need to select the regularization parameters to include in the penalized likelihood method. To evaluate the estimated model, we derived various model selection criteria based on information theory and the Bayesian approach (Konishi and Kitagawa, 2008). The proposed method was investigated through simulation studies and real data analysis, and we will show that the proposed method selects adequate decision boundaries.

This paper is organized as follows. Section 2 introduces the multinomial logistic regression model for classifying data into three or more groups. In Section 3, we derive a fused lasso-type penalty for the logistic regression model and show the effect of this penalty. We also present an iterative procedure for estimating models, and we derive four model selection criteria for evaluating the estimated models. The results of simulation studies and two real data analyses are presented in Sections 4 and 5, respectively. Finally, concluding remarks are given in Section 6.

2. Multinomial logistic regression model

Suppose we have n observations $\{(\chi_i, g_i);\, i = 1, \ldots, n\}$, where $\chi_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of p predictors for the ith subject, and $g_i \in \{1, \ldots, L\}$ is the class label to which χ_i belongs. In the classification setting, we apply the Bayes rule which assigns

χ_i to class $g_i = l$ with the maximum posterior probability given by χ_i; this is denoted as $\Pr(g_i = l|\chi_i)$. Then, the logistic regression model is given by the log odds of the posterior probabilities:

$$\log \frac{\Pr(g_i = l|\chi_i)}{\Pr(g_i = L|\chi_i)} = \boldsymbol{x}_i^T \boldsymbol{\beta}_l, \quad l = 1, \ldots, L-1, \tag{1}$$

where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$ and $\boldsymbol{\beta}_l = (\beta_{l0}, \beta_{l1}, \ldots, \beta_{lp})^T$ are vectors of the predictors and coefficients, respectively. The coefficient β_{lj} relates to the decision boundary between classes l and L for the jth variable. Therefore, if $\beta_{lj} = 0$, then the jth variable has no effect on the classification between the classes l and L. Here, we consider the case in which the coefficient vector with $l = L$, denoted by $\boldsymbol{\beta}_L$, is equal to $\mathbf{0}$. The class L is then referred to as the reference class.

It follows from (1) that the posterior probability $\pi_l(\boldsymbol{x}_i; \boldsymbol{\beta}) = \Pr(g_i = l|\boldsymbol{x}_i)$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{L-1}^T)^T$ is given by

$$\begin{aligned} \pi_l(\boldsymbol{x}_i; \boldsymbol{\beta}) &= \frac{\exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_l\right)}{1 + \sum_{h=1}^{L-1} \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_h\right)} \quad (l = 1, \ldots, L-1), \\ \pi_L(\boldsymbol{x}_i; \boldsymbol{\beta}) &= \frac{1}{1 + \sum_{h=1}^{L-1} \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_h\right)}. \end{aligned} \tag{2}$$

We define the vectors of the binary responses which indicate the class labels as

$$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i(L-1)})^T = \begin{cases} (0, \ldots, 0, \overset{(l)}{1}, 0, \ldots, 0)^T & \text{if } g_i = l, \quad l = 1, \ldots, L-1, \\ (0, \ldots, 0)^T & \text{if } g_i = L. \end{cases} \tag{3}$$

Then, we can construct a joint probability function for the multinomial distribution:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\beta}) = \prod_{l=1}^{L-1} \pi_l(\boldsymbol{x}_i; \boldsymbol{\beta})^{y_{il}} \pi_L(\boldsymbol{x}_i; \boldsymbol{\beta})^{1 - \sum_{h=1}^{L-1} y_{ih}}.$$

3. Estimation and evaluation

In this section, we establish a strategy for constructing the logistic regression model. In the following subsections, we first introduce a method for estimating the model, and then we derive four model selection criteria for evaluating the estimated models.

3.1. Penalized likelihood method with a fused lasso-type penalty

We consider the estimation of a logistic regression model (1) with the maximum penalized likelihood method, which maximizes the penalized log-likelihood function given as

$$\ell_\lambda(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{\beta}) - nP_\lambda(\boldsymbol{\beta}),$$

where $P_\lambda(\cdot)$ is a penalty function. We apply a penalty function given as

$$P_\lambda(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} \sum_{l=1}^{L-1} |\beta_{lj}| + \lambda_2 \sum_{j=1}^{p} \sum_{1 \le l' < l < L} |\beta_{lj} - \beta_{l'j}|, \tag{4}$$

Table 1: Estimates of the coefficients β_{lj} for a toy example. (a) True, (b) estimates with $\lambda_1 = 1.0 \times 10^{-3}$ and $\lambda_2 = 0$ (lasso), (c) estimates with $\lambda_1 = 1.0 \times 10^{-3}$ and $\lambda_2 = 1.0 \times 10^{-3}$ (proposed).

(a)

$j \backslash l$	1	2
1	1.000	5.000
2	2.000	4.000
3	3.000	3.000
4	0.000	2.000
5	0.000	0.000

(b)

$j \backslash l$	1	2
1	0.735	11.244
2	0.000	8.906
3	2.824	5.908
4	0.000	5.762
5	0.000	0.000

(c)

$j \backslash l$	1	2
1	0.719	6.839
2	0.001	4.805
3	2.953	2.953
4	0.000	2.987
5	0.000	0.000

where $\lambda_1 > 0$ and $\lambda_2 > 0$, respectively, are regularization parameters which adjust the degrees of the penalties for the first and the second terms of (4). The first term gives the effect of individually shrinking some of coefficients; the result is that there are no boundaries between classes l and L. On the other hand, the second term encourages the coefficients to have the same value for all combinations of l and l', that is, it penalizes so that $\beta_{lj} = \beta_{l'j}$ for each j. The result of this is that there is no boundary between the classes l and l', because the log odds of these classes are represented by

$$\log \frac{\pi_l(\boldsymbol{x}_i; \boldsymbol{\beta})}{\pi_{l'}(\boldsymbol{x}_i; \boldsymbol{\beta})} = \boldsymbol{x}_i^T (\boldsymbol{\beta}_l - \boldsymbol{\beta}_{l'}).$$

In order to see the effect of our method, we consider a toy example with simulated data, where $n = 50$, $p = 5$, and $L = 3$. The data were generated as follows. Predictors X_j $(j = 1, \ldots, p)$ were generated from $N_p(\mathbf{0}, \Sigma)$, where $\Sigma_{st} = 0.5^{|s-t|}$, and the true coefficients are given in Table 1 (a). The response Y was obtained by assigning (3) to the data whose class maximized (2). Coefficients β_{1j} and β_{2j} correspond to the boundaries between classes $l = 1$ and $l = 3$ and classes $l = 2$ and $l = 3$, respectively. Table 1 (b) shows the lasso estimates, and we can see that some coefficients are estimated to be exactly zero. This reveals that there is no boundary in the corresponding part, and, in particular, there is no boundary for $j = 5$. In other words, these are irrelevant for the classification. However, this result cannot lead to the exclusion of the boundary between the classes $l = 1$ and $l = 2$ for the third variable, even though it is truly irrelevant. On the other hand, the results of the proposed method, the underlined values in Table 1 (c), show that the coefficients for the third variable are estimated to be equal to each other, but they are not zero. This indicates that the classification between $l = 1$ and $l = 2$ is irrelevant for the third variable. Note that when $L = 3$, as in this example, the proposed penalty corresponds to the fused lasso penalty.

3.2. Local quadratic approximation algorithm

Since the penalized log-likelihood function involves the L_1 norm of the coefficients, it is difficult to derive estimates analytically, and therefore, iterative calculations are required. There are several algorithms for the fused lasso regularization (Tibshirani et al., 2005; Friedman et al., 2007; Höefling, 2010). More recently, Tibshirani and Taylor

(2011) proposed a unified algorithm for the generalized lasso, including the fused lasso. However, the proposed penalty (4) does not satisfy the "boundary lemma" conditions that would allow us to execute the algorithm of Tibshirani and Taylor (2011). Thus, we apply the local quadratic approximation (LQA) algorithm, which was originally used by Tibshirani (1996) and Fan and Li (2001) for L_1-type regularization problems. Ulbricht (2010) extended the LQA to a wider class of problems, including fused lasso-type regularization, and it is implemented for a generalized linear model in the R package `lqa`. In order to apply it to the estimation of a multinomial logistic regression model, we define a function

$$p_\lambda(|\boldsymbol{a}_h^T\boldsymbol{\beta}|) = \begin{cases} \lambda_1|\boldsymbol{a}_h^T\boldsymbol{\beta}|, & h = 1,\ldots,p(L-1), \\ \lambda_2|\boldsymbol{a}_h^T\boldsymbol{\beta}|, & h = p(L-1)+1,\ldots,H, \end{cases} \tag{5}$$

where $H = p(L-1) + p(L-1)(L-2)/2$, and the $\boldsymbol{a}_h$ are $p(L-1)$-dimensional vectors given in the following form:

$$\begin{pmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_H^T \end{pmatrix} = \begin{pmatrix} & & I_{p(L-1)} & & \\ -I_p & I_p & O & \cdots & O \\ -I_p & O & I_p & \cdots & O \\ \vdots & & \ddots & & \vdots \\ O & \cdots & -I_p & O & I_p \\ O & \cdots & O & -I_p & I_p \end{pmatrix}.$$

The upper and lower parts of (5) correspond to the first and second terms of (4), respectively. Thus, the penalty function can be expressed as

$$P_\lambda(\boldsymbol{\beta}) = \sum_{h=1}^{H} p_\lambda(|\boldsymbol{a}_h^T\boldsymbol{\beta}|).$$

Ulbricht (2010) showed that if the penalty function can be expressed in the above form, then we can use the LQA to estimate the parameters. When the initial value of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(0)}$, the $(k+1)$-th update is

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} - \left\{ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\beta^{(k)}} - n\Omega(\boldsymbol{\beta}^{(k)}) \right\}^{-1} \left\{ \frac{\partial \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\beta^{(k)}} - n\Omega(\boldsymbol{\beta}^{(k)})\boldsymbol{\beta}^{(k)} \right\} \\ &= \left\{ \left(\tilde{X}^T W \tilde{X} + n\Omega(\hat{\boldsymbol{\beta}}^{(k)}) \right)^{-1} \tilde{X}^T \right\} W \left\{ \tilde{X}\boldsymbol{\beta}^{(k)} + W^{-1}\Delta \mathbf{1}_{n(L-1)} \right\}, \end{aligned}$$

where

$$\tilde{X} = (\mathbf{1}_{L-1}\mathbf{1}_{L-1}^T) \otimes X, \quad X = (\boldsymbol{x}_1^T,\ldots,\boldsymbol{x}_n^T),$$

$$W^{(k)} = \begin{pmatrix} W_{11}^{(k)} & \cdots & W_{1(L-1)}^{(k)} \\ \vdots & \ddots & \vdots \\ W_{(L-1)1}^{(k)} & \cdots & W_{(L-1)(L-1)}^{(k)} \end{pmatrix},$$

$$W_{hl}^{(k)} = \begin{cases} \operatorname{diag}\{\pi_l(\boldsymbol{x}_1;\boldsymbol{\beta})(1-\pi_l(\boldsymbol{x}_1;\boldsymbol{\beta})),\ldots,\pi_l(\boldsymbol{x}_n;\boldsymbol{\beta})(1-\pi_l(\boldsymbol{x}_n;\boldsymbol{\beta}))\} & (h=l), \\ \operatorname{diag}\{-\pi_h(\boldsymbol{x}_1;\boldsymbol{\beta})\pi_l(\boldsymbol{x}_1;\boldsymbol{\beta}),\ldots,-\pi_h(\boldsymbol{x}_n;\boldsymbol{\beta})\pi_l(\boldsymbol{x}_n;\boldsymbol{\beta})\} & (h\neq l), \end{cases}$$

$$\Omega(\boldsymbol{\beta}) = \sum_{h=1}^{H} \frac{p_\lambda(|\boldsymbol{a}_h^T\boldsymbol{\beta}|)}{\sqrt{(\boldsymbol{a}_h^T\boldsymbol{\beta})^2+c}}\boldsymbol{a}_h\boldsymbol{a}_h^T,$$
$$\Delta = \text{blockdiag}\left\{\Delta_1, \ldots, \Delta_{L-1}\right\}, \quad \Delta_l = \text{diag}\left\{y_{1l} - \pi_{1l}, \ldots, y_{nl} - \pi_{nl}\right\},$$

where $\otimes$ represents the Kronecker product, and c in $\Omega(\boldsymbol{\beta})$ is a positive constant that is sufficiently small (e.g., 10^{-6}) to prevent the elements from diverging.

3.3. Model selection criteria

Since the estimated model depends on the regularization parameters λ_1 and λ_2 in (4), it is important to determine these values objectively. We will consider selecting these values by applying various model selection criteria based on information theory and the Bayesian approach.

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for evaluating models estimated by the maximum penalized likelihood method with the proposed penalty are respectively given by

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2\widetilde{df},$$
$$\text{BIC} = -2\ell(\hat{\boldsymbol{\beta}}) + \widetilde{df}\log n,$$

where $\widetilde{df}$ is given by

$$\widetilde{df} = \text{tr}\left\{W\tilde{X}_{\mathcal{A}}(\tilde{X}_{\mathcal{A}}^T W\tilde{X}_{\mathcal{A}} + n\Omega(\hat{\boldsymbol{\beta}}_{\mathcal{A}}))^{-1}\tilde{X}_{\mathcal{A}}^T\right\}.$$

Here, $\tilde{X}_{\mathcal{A}}$ indicates the active set of $\tilde{X}$, and $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ indicates the nonzero elements of $\hat{\boldsymbol{\beta}}$. The definition of $\widetilde{df}$ is due to the effective degrees of freedom, and it was originally proposed by Hastie and Tibshirani (1990) and was used for sparse regularization in Tibshirani (1996) and Fan and Li (2001). Tibshirani et al. (2005) used the number of distinct nonzero estimates as the degrees of freedom, applying the result described in Zou et al. (2007) for fused lasso regularization. However, this result holds true for the linear model with Gaussian noise, not for the logistic regression model. On the other hand, Zhang et al. (2010) showed that the BIC-type criterion with the effective degrees of freedom consistently selects variables when smoothly clipped absolute deviation (SCAD) regularization (Fan and Li, 2001) is applied.

While the AIC- and BIC-type criteria are widely used for evaluating various types of statistical models, they were initially derived for evaluating models estimated by the maximum likelihood method, not the maximum penalized likelihood method. On the other hand, Konishi and Kitagawa (1996) derived a generalized information criterion (GIC) for evaluating models estimated by the M-estimates, including the maximum penalized likelihood estimates. Using this result, the GIC for evaluating logistic regression models estimated by the regularization method with the penalty given in (4) is given by

$$\text{GIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2\text{tr}\left\{R^{-1}(\hat{\boldsymbol{\beta}})Q(\hat{\boldsymbol{\beta}})\right\},$$

where $R(\hat{\boldsymbol{\beta}})$ and $Q(\hat{\boldsymbol{\beta}})$ are respectively defined by

$$R(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\tilde{X}_{\mathcal{A}}^T W\tilde{X}_{\mathcal{A}} + \Omega(\hat{\boldsymbol{\beta}}_{\mathcal{A}}),$$
$$Q(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\tilde{X}_{\mathcal{A}}^T\tilde{\Delta}\tilde{X}_{\mathcal{A}} - \Omega(\hat{\boldsymbol{\beta}}_{\mathcal{A}})\hat{\boldsymbol{\beta}}_{\mathcal{A}}\mathbf{1}_{n(L-1)}^T\tilde{\Delta}\tilde{X}_{\mathcal{A}},$$

where $\tilde{\Delta} = (\Delta_1 \cdots \Delta_{L-1})^T(\Delta_1 \cdots \Delta_{L-1})$. Furthermore, Konishi et al. (2004) derived a generalized Bayesian information criterion (GBIC) for evaluating models estimated by the maximum penalized likelihood method. By applying the result of Konishi et al., the GBIC for evaluating models estimated by our method is given by

$$\text{GBIC} = -2\ell(\hat{\boldsymbol{\beta}}) + n\hat{\boldsymbol{\beta}}^T\Omega(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}} + q\log n + \log|R(\hat{\boldsymbol{\beta}})| - \log|\Omega(\hat{\boldsymbol{\beta}})|_+ - q\log(2\pi),$$

where $q = p(L-1) - \text{rank}\Omega(\hat{\boldsymbol{\beta}})$ and $|\Omega(\hat{\boldsymbol{\beta}})|_+$ denotes a product of the nonzero eigenvalues of $\Omega(\hat{\boldsymbol{\beta}})$. These criteria are repeatedly computed for various candidate values for λ_1 and λ_2, and then we determine which values of λ_1 and λ_2 minimize these criteria.

4. Simulation study

We performed Monte Carlo simulations to investigate the effectiveness of the proposed method. We simulated the data according to the logistic regression model (1). The predictors X_j were generated from $N_p(\mathbf{0}, \Sigma)$, where $\Sigma_{st} = 0.5^{|s-t|}$, and the true coefficients were given in the following two settings:

1. $p = 4$, $L = 4$, $\boldsymbol{\beta}_1 = (3, 1.5, 0, 0)^T$, $\boldsymbol{\beta}_2 = (1.5, 1.5, 0, 0)^T$, $\boldsymbol{\beta}_3 = (0, 1.5, 0, 0)^T$,

2. $p = 5$, $L = 4$, $\boldsymbol{\beta}_1 = (1, 2, 3, 0, 0)^T$, $\boldsymbol{\beta}_2 = (5, 4, 3, 0, 0)^T$, $\boldsymbol{\beta}_3 = (1, 2, 0, 0, 0)^T$.

After calculating the posterior probabilities, we obtained the true responses $\boldsymbol{y}_i^{(t)}$ as in (3), so that each subject was assigned to the class with the maximum posterior probability. We then used the posterior probability to generate random observations from the multinomial distribution for the responses $\boldsymbol{y}_i$. We estimated the logistic regression model by the proposed method, and then the tuning parameters were selected using the four model selection criteria described in Section 3.2. We repeated this strategy for 100 simulated data sets with $n = 50, 100, 200$ for each setting. After obtaining the estimated coefficients $\hat{\beta}_{lj}$ and responses $\hat{\boldsymbol{y}}_i$, we calculated the misclassification rate MCR $= \sum_{i=1}^n I(\boldsymbol{y}_i^{(t)} \neq \hat{\boldsymbol{y}}_i)/n$, where $I(\cdot)$ is an indicator function. Furthermore, we examined the selection accuracy by comparing the error rates between classes $l(< L)$ and L and between classes $l(< L)$ and $l'(< L)$, respectively defined by SER1 $= \sharp\{\hat{\beta}_{lj}|\ I(\hat{\beta}_{lj} = 0) \neq I(\beta_{lj} = 0)\}/\{p(L-1)\}$ and SER2 $= \sharp\{\hat{\beta}_{lj}|\ I(\hat{\beta}_{lj} = \hat{\beta}_{l'j}) \neq I(\beta_{lj} = \beta_{l'j}), l > l', \beta_{lj} \neq 0\}/\{p(L-1)(L-2)/2\}$.

Tables 2 and 3 give the results of the simulations using settings 1 and 2, respectively. These include the averaged values of 100 selected regularization parameters (λ_1 and λ_2 for penalty (4)), the misclassification rates MCR, and the selection error rates SER1 and SER2, for the proposed method and the lasso. Note that we omitted λ_2 for the lasso from the table, since the lasso corresponds to the case with $\lambda_2 = 0$. In most cases, the proposed method obtained MCRs that were smaller than those obtained by the lasso; this difference was largest when the sample size was large. On the other hand, there were smaller differences between the results of the four model selection criteria. The proposed method obtained SER1s and SER2s that were smaller than those obtained by the lasso. The model selection criteria BIC and GBIC tended to obtain smaller SER1s and SER2s than those obtained by the other criteria.

Table 2: Results for the simulation with setting 1. Values of λ_1 and λ_2 are multiplied by 10^3, and MCR, SER1, and SER2 are expressed as percentages.

		Proposed					LASSO			
		λ_1	λ_2	MCR	SER1	SER2	λ_1	MCR	SER1	SER2
$n = 50$	AIC	1.25	2.86	17.50	21.58	18.58	1.81	20.48	23.58	15.17
	BIC	2.02	4.29	17.38	19.67	15.25	3.47	17.82	27.00	17.25
	GIC	1.17	2.93	17.86	25.08	20.58	1.79	20.34	23.92	15.25
	GBIC	4.64	5.21	15.58	25.58	17.17	5.30	15.08	32.50	20.75
$n = 100$	AIC	0.85	2.20	11.86	8.16	6.94	1.04	18.33	10.65	7.83
	BIC	1.33	3.76	10.88	6.38	3.60	2.17	19.80	12.30	7.80
	GIC	0.93	2.61	11.83	8.67	7.38	1.05	18.33	10.60	7.77
	GBIC	2.61	4.09	13.49	7.68	5.31	4.32	17.56	16.81	10.49
$n = 200$	AIC	0.52	1.38	7.49	14.25	9.83	0.67	16.41	18.42	12.33
	BIC	0.71	2.94	6.44	7.50	4.50	1.08	17.12	18.67	11.17
	GIC	0.56	1.56	7.60	14.33	10.08	0.67	16.41	18.67	12.50
	GBIC	1.09	2.12	6.64	6.25	4.92	1.47	18.26	19.33	11.00

Table 3: Results for the simulation with setting 2. Values of λ_1 and λ_2 are multiplied by 10^3, and MCR, SER1, and SER2 are expressed as percentages.

		Proposed					LASSO			
		λ_1	λ_2	MCR	SER1	SER2	λ_1	MCR	SER1	SER2
$n = 50$	AIC	0.81	1.57	13.46	21.67	20.20	0.90	15.80	21.27	28.67
	BIC	1.37	2.88	13.24	19.53	19.00	2.16	16.58	26.33	24.80
	GIC	0.62	1.21	13.86	23.53	21.20	0.72	15.08	22.60	30.60
	GBIC	3.37	4.11	14.34	24.07	18.67	4.96	14.58	39.27	21.67
$n = 100$	AIC	0.56	1.23	8.82	15.80	12.33	0.70	15.13	18.00	26.27
	BIC	1.28	2.81	8.75	11.13	11.80	1.57	16.37	21.47	23.33
	GIC	0.64	1.36	8.83	15.47	12.67	0.74	14.99	18.60	27.47
	GBIC	1.87	2.25	9.85	12.60	12.73	2.85	17.35	27.80	21.80
$n = 200$	AIC	0.44	0.95	6.29	11.87	8.40	0.45	14.59	15.47	25.73
	BIC	0.71	2.15	6.03	6.13	5.80	0.81	15.31	16.73	23.60
	GIC	0.49	1.21	6.25	11.87	8.60	0.45	14.54	15.67	26.13
	GBIC	0.83	1.03	5.85	6.73	6.27	1.24	16.30	20.33	21.67

5. Real data analysis

We applied real data analysis to the proposed method, and then investigated the results. The data sets used here are vowel data and handwritten zip code data, both of which are available from the website[1] of Hastie et al. (2009). In these analyses, the tuning parameters included in the penalty were selected by the BIC, since it performed well in the numerical experiments.

[1] http://statweb.stanford.edu/~tibs/ElemStatLearn/

Table 4: Coefficient estimates $\hat{\beta}_{lj}$ for the vowel data set. Underlined values have the same estimated value for each variable j.

$j\backslash l$	1	2	3	4	5	6	7	8	9	10
1	−3.84	−0.21	−1.86	−1.86	6.05	1.24	4.75	4.44	1.24	1.24
2	−15.29	−2.94	−5.12	−2.66	5.95	1.70	5.31	4.97	2.25	1.70
3	1.39	−0.24	−5.34	−7.45	−0.89	0.08	1.71	−0.24	1.71	2.92
4	0.00	2.36	0.00	−7.26	0.00	−1.65	2.36	−1.28	0.00	2.36
5	5.62	5.62	0.27	−6.25	−5.91	−3.21	−3.55	−3.21	0.80	0.00
6	−2.67	0.80	−3.50	−10.66	2.64	0.80	2.64	0.80	1.74	1.74
7	−1.75	−0.36	−3.08	−5.60	−5.60	−2.44	−0.36	−0.36	−0.61	−0.61
8	1.32	1.32	−1.90	−7.62	−1.90	−1.90	0.00	1.32	−0.10	−1.90
9	0.00	2.07	0.00	0.00	2.07	0.00	0.00	2.07	2.07	1.48
10	0.00	1.53	0.00	−1.68	0.00	0.00	0.00	1.53	0.00	−2.44

Table 5: Numbers of decision boundaries for each variable.

j	1	2	3	4	5	6	7	8	9	10
#	7	9	8	4	7	6	6	5	2	3

5.1. Vowel data

This data set consists of 462 observations, including 11 kinds of steady-state vowels of British English, and each of them has 10 voice features. The number of variables is $p = 10$, and the number of classes is $L = 11$. We applied the proposed modeling strategy to an analysis of the data, and we obtained estimates of the coefficients. In this analysis, we let the reference class $l = L$ be the last class of the data, that is, the logistic regression model consists of log odds ratios for classes $l(< 11)$ and $L\ (= 11)$. The tuning parameters included in the model were selected by the BIC.

Table 4 shows the coefficients estimated for the logistic regression model for the vowel data. Some of coefficients were estimated to be zero; this indicates that these variables have no effect on the classification between $l(< L)$ and L. Furthermore, some coefficients have the same value for each variable. This means that there are no decision boundaries between $l(< L)$ and $l'(< L)$. For example, for variable $j = 4$, there are no boundaries between classes $l = 1, 3, 5, 9$, and 11. Moreover, there also are no boundaries between classes $l = 2, 7$, and 10, since the corresponding coefficients of the model are estimated to be the same. As a result, only four decision boundaries are estimated for variable $j = 4$. The decision boundaries for all variables are summarized in Table 5. It can be seen that almost half of the decision boundaries have disappeared for most variables.

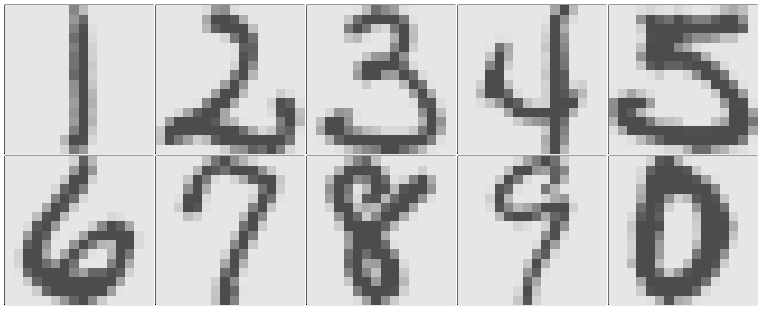

Figure 1: Example of handwritten zip code data.

0	0	0	0	1	0	0	2	1	2	0	1	1	1	0	1
0	0	0	1	1	2	1	3	2	2	1	1	0	0	0	0
0	0	1	1	0	0	0	1	1	0	1	1	0	0	0	1
0	0	0	1	1	1	1	3	0	1	1	1	0	0	0	0
1	0	1	0	2	0	0	1	1	1	1	1	0	0	0	0
1	0	0	1	0	3	0	1	2	0	1	3	1	2	2	0
0	0	1	1	2	0	0	1	1	0	0	0	0	0	0	0
0	1	1	1	0	1	1	2	2	1	1	1	0	0	0	0
0	0	1	0	2	1	0	1	0	1	0	2	0	0	0	0
0	0	0	1	1	1	0	1	1	0	1	0	0	0	0	0
1	1	1	0	2	1	1	0	0	0	0	2	0	2	1	0
1	0	0	1	0	2	1	1	1	0	0	1	0	0	1	0
0	1	0	0	1	1	1	0	0	0	3	1	2	0	1	0
0	1	1	1	0	2	0	2	2	0	0	1	0	1	1	0
0	2	0	1	1	0	2	0	4	1	0	1	1	1	0	0
1	0	0	0	2	0	3	1	2	1	0	0	0	0	0	0

Figure 2: Number of decision boundaries for each pixel for the zip code data.

5.2. Handwritten zip code data

We applied the proposed method to the analysis of data of handwritten digits. Each handwritten digit was represented as 16×16 pixels (examples are given in Figure 1), and the 256 luminance values were treated as an individual datum. This dataset has a sample size of $n = 7291$ with $p = 256$ dimensions and $L = 10$ classes. We applied our method to the analysis of this data set, and then used the BIC to select the regularization parameters.

Figure 2 shows numbers of selected decision boundaries for each pixel. The position of the lattice corresponds to that of the pixel of each digit. It can be seen in this figure that most pixels on the left and right sides have zero or one decision boundary. In other words, these pixels contain little information that is useful for classification. This result

makes sense, because most digits do not have a similar amount of structure on both sides. On the other hand, some pixels near the horizontal center have more decision boundaries; this is especially so at the center top and center bottom. This reveals that the pixels in those locations contain more information about the classification of digits.

6. Concluding remarks and discussion

We have proposed a method for selecting decision boundaries rather than variables in a classification problem. We applied the logistic regression model, and then the model was estimated by the maximum penalized likelihood method with the fused lasso-type penalty; this was done in order to shrink some of the coefficients toward zero and to shrink the differences between some pairs of coefficients toward zero. The model was estimated by the maximum penalized likelihood method using the local quadratic approximation algorithm. Since it is crucial to select regularization parameters that are included in the penalty, we introduced four model selection criteria. Simulation results showed that the proposed method performed fairly well compared to the existing method. Furthermore, we applied the proposed method to the analysis of some real data sets, and then selected decision boundaries for each variable.

In this work, we constructed the logistic regression model by considering the reference class, but we did not care how it was selected. Hastie et al. (2015) described that the model provides different results when the reference class is selected in different ways. Furthermore, Kim et al. (2006) pointed out that when the sparse regularization method is applied to this model, too many variables may be dropped. Hastie et al. (2015) suggested using a logistic regression model without using the reference class. We will consider using this model in future investigations.

When we treat β_{lj}, $j = 1, \ldots, p$ as grouped parameters, as in the group lasso of Yuan and Lin (2006), we may be able to perform clustering, since each coefficient vector $\boldsymbol{\beta}_l$ coincides with the coefficient vector $\boldsymbol{\beta}_{l'}$. Bondell and Reich (2008) approached the clustering problem in the linear model by constructing an octagonal penalty. Our intended future work will include the construction of a clustering method that uses the logistic regression model and sparse regularization.

Acknowledgements

The author is grateful to the referee for valuable comments and suggestions for the refinement of this paper. This work was supported by Grant-in-Aid for Young Scientists (B) No. 25730017 of JSPS.

References

Bondell, H. D. and Reich, B. J. (2008), Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR, *Biometrics*, 64, 115–123.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Heidelberg: Springer.

Daye, Z. J. and Jeng, X. J. (2009), Shrinkage and model selection with correlated variables via weighted fusion, *Comput. Statist. Data Anal.*, 53, 1284–1298.

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, 96, 1348–1360.

Frank, I. and Friedman, J. (1993), A statistical view of some chemometrics regression tools, *Technometrics*, 35, 109–135.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), Pathwise coordinate optimization, *Ann. Appl. Statist.*, 1, 302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), Regularization paths for generalized linear models via coordinate descent, *J. Stat. Softw.*, 33, 1–22.

Hastie, T. and Tibshirani, R. (1990), *Generalized additive models*, London: Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The elements of statistical learning 2nd ed.*, New York: Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalization*, Boca Raton: Chapman & Hall/CRC.

Höefling, H. (2010), A path algorithm for the fused lasso signal approximator, *J. Comput. Graph. Statist.*, 19, 984–1006.

Jang, W., Lim, J., Lazar, N. A., Loh, J. M., and Yu, D. (2013), Regression shrinkage and grouping of highly correlated predictors with HORSES, *arXiv preprint*.

Kim, Y., Kwon, S., and Heun Song, S. (2006), Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data, *Comput. Statist. Data Anal.*, 51, 1643–1655.

Konishi, S., Ando, T., and Imoto, S. (2004), Bayesian information criteria and smoothing parameter selection in radial basis function networks, *Biometrika*, 91, 27–43.

Konishi, S. and Kitagawa, G. (1996), Generalised information criteria in model selection, *Biometrika*, 83, 875–890.

— (2008), *Information criteria and statistical modeling*, New York: Springer.

Krishnapuram, B., Carin, L., Figueiredo, M., and Hartemink, A. (2005), Sparse multinomial logistic regression: Fast algorithms and generalization bounds, *IEEE Trans. Pattern Anal. Mach. Intell.*, 27, 957–968.

Lin, Y., Wang, S., and Chappell, R. (2013), Lasso tree for cancer staging with survival data, *Biostatistics*, 14, 327–339.

Matsui, H. (2014), Variable and boundary selection for functional data via multiclass logistic regression modeling, *Comput. Statist. Data Anal.*, 78, 176–185.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear model*, London: Chapman & Hall/CRC.

Meier, L., van de Geer, S., and Bühlmann, P. (2008), The group lasso for logistic regression, *J. Roy. Statist. Soc. Ser. B*, 70, 53–71.

Park, M. and Hastie, T. (2007), L1-regularization path algorithm for generalized linear models, *J. Roy. Statist. Soc. Ser. B*, 69, 659–677.

She, Y. (2010), Sparse regression with exact clustering, *Electron. J. Stat.*, 4, 1055–1096.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), Sparsity and smoothness via the fused lasso, *J. Roy. Statist. Soc. Ser. B*, 67, 91–108.

Tibshirani, R. and Taylor, J. (2011), The solution path of the generalized lasso, *Ann. Statist.*, 39, 1335–1371.

Tutz, G. and Ulbricht, J. (2009), Penalized regression with correlation-based penalty, *Statist. Comput.*, 19, 239–253.

Ulbricht, J. (2010), *Variable selection in generalized linear models. Dissertation*, Ludwig-Maximilians-Universität, München.

Wang, L. and Shen, X. (2006), Multi-category support vector machines, feature selection and solution path, *Statist. Sinica*, 16, 617–633.

Witten, D. M. and Tibshirani, R. (2011), Penalized classification using Fisher's linear discriminant, *J. Roy. Statist. Soc. Ser. B*, 73, 753–772.

Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *J. Roy. Statist. Soc. Ser. B*, 68, 49–67.

Zhang, C. (2010), Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, 38, 894–942.

Zhang, Y., Li, R., and Tsai, C. (2010), Regularization parameter selections via generalized information criterion, *J. Amer. Statist. Assoc.*, 105, 312–323.

Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2004), 1-norm support vector machines, *Adv. Neural Inf. Process. Syst.*, 16, 49–56.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. Roy. Statist. Soc. Ser. B*, 67, 301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007), On the "degrees of freedom" of the lasso, *Ann. Statist.*, 35, 2173–2192.

Received September 18, 2015
Revised December 4, 2015