Feature subset selection for the multinomial logit model via mixed-integer optimization

Shunsuke Kamiya
Tokyo University of Agriculture and Technology

Ryuhei Miyashiro
Tokyo University of Agriculture and Technology

Yuichi Takano
University of Tsukuba

Abstract

This paper is concerned with a feature subset selection problem for the multinomial logit (MNL) model. There are several convex approximation algorithms for this problem, but to date the only exact algorithms are those for the binomial logit model. In this paper, we propose an exact algorithm to solve the problem for the MNL model. Our algorithm is based on a mixed-integer optimization approach with an outer approximation method. We prove the convergence properties of the algorithm for more general models including generalized linear models for multiclass classification. We also propose approximation of loss functions to accelerate the algorithm computationally. Numerical experiments demonstrate that our exact and approximation algorithms achieve better generalization performance than does an L_1-regularization method.

1 Introduction

Sparse estimation is an important task for multi-class classification. By reducing the complexity of models, we can obtain good predictors for unobserved data. In particular, a number of papers have been published on feature subset selection for the multinomial logit (MNL) model (McFadden, 1973), whose criterion is based on the likelihood of observed data.

The MNL model is used to predict not only which category is chosen by samples but also the probability of choosing each category; for example, it can be used to predict the purchase behavior of a customer in marketing (Guadagni and Little, 1983; Louviere and Woodworth, 1983) and the failure of a bank in financial modeling (Lau, 1987; Johnsen and Melicher, 1994). In practice, modelers often use maximum likelihood estimation to estimate the MNL model. However, because this method often causes over-fitting to observed data and the lack of interpretability of the model, such an estimator does not work well for unobserved data.

To avoid the aforementioned difficulties, it is worth considering the problem of feature subset selection. We can define this optimization problem as the maximization of likelihood under the cardinality constraint that the number of features that have nonzero coefficients for at least one category does not exceed a given number k. The complexity of the model is therefore bounded, and we expect to obtain a high-quality model that overcomes the two aforementioned shortcomings.

However, the feature subset selection problem is an NP-hard combinatorial optimization problem (Kohavi and John, 1997). To solve it approximately, several convex approximations have been proposed, including L_1-regularization and elastic net regularization. Some of these methods do not consider a group sparse structure (e.g., Krishnapuram et al. (2005); Friedman et al. (2010)), whereas others do (e.g., Simon et al. (2013); Vincent and Hansen (2014)). While all of these methods are relatively fast, their approximation accuracy is not particularly high.

In contrast to such approximation methods, there is renewed interest in solving the feature subset selection problem exactly via mixed-integer optimization (MIO). This is due to improved computational power and the development of high-performance MIO algorithms. There have been several studies on MIO approaches for linear regression models (Miyashiro and Takano, 2015; Bertsimas et al., 2016). Hastie et al. (2017) showed that the estimators obtained by the MIO approaches are more useful than those obtained by L_1-regularization when the signal-to-noise ratio is high.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

For binary classification, there have been similar studies involving the binomial logit model and a support vector machine (Sato et al., 2016; Bertsimas et al., 2017; Bertsimas and King, 2017). Similar to Hastie et al. (2017), Bertsimas and King (2017) have confirmed that their estimators have better generalization ability than do some heuristics such as L_1-regularization methods. However, it is difficult to solve the continuous relaxation problem for many multi-class classification models, including the binomial logit model, because of the nonlinearity of the loss functions to be minimized. It is therefore more difficult to construct efficient algorithms for these models than for the linear regression models, whose objective is defined as a convex quadratic function. In fact, Bertsimas and King (2017) have empirically observed that a naïve branch-and-bound method is useless for solving the problem within practical computational time for the binomial logit model.

Because of such difficulty, most previous studies modified the objective function to apply their methods to large data while maintaining the quality of their estimator. To accelerate the branch-and-bound method by removing the nonlinearity of continuous relaxations, a convex quadratic approximation (Tanaka and Nakagawa, 2014) and a piecewise linear approximation (Sato et al., 2016, 2017) have been proposed. It was also shown by Bertsimas et al. (2017) that the outer approximation method is applicable to larger data for these models with L_2-regularization. However, these studies only considered the binary or ordinal classification problems; to the best of our knowledge, no study has yet dealt with the feature subset selection problem of the MNL model via an MIO approach.

We extend a state-of-the-art method for the feature subset selection problem for the binomial logit model (Bertsimas et al., 2017) to the same problem but for the MNL model. Our method can generate a sparse solution as fast as existing heuristics under a certain regularization. Furthermore, we propose a new approximation of the loss function to accelerate our algorithm and confirm that the proposed approximation method also provides high-quality solutions.

In this paper, we make the following contributions to show the efficacy of our proposal.

- The work of Bertsimas et al. (2017) is extended to multi-class classification; for the feature subset selection problem with L_2-regularization, an MIO formulation and an algorithm based on the outer approximation method are proposed.
- Sufficient conditions for the convergence of our algorithm are described; the loss function of the MNL model is proven to satisfy these conditions.
- A novel approximation of the loss function is derived; this is shown to be advantageous over Tanaka and Nakagawa (2014) and Sato et al. (2016).
- Empirical properties of our method are shown; in particular, the method has both good generalization ability and fast convergence speed when the value of the L_2-regularization parameter is small.

2 Formulations

Let n, p, and m be the numbers of samples, features, and categories of data, respectively. Suppose that we are given $\boldsymbol{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in [m]^n$, where $[m] = \{1, 2, \ldots, m\}$. First, we consider sparse estimation of the MNL model. Specifically, parameters $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{W} = (w_{rj}) \in \mathbb{R}^{m\times p}$ are estimated so that the likelihood of observed data is maximized and nonzero elements in $\boldsymbol{W}$ are sparse. The likelihood is derived by the following softmax probability:

$$\mathrm{P}(Y = r \mid \boldsymbol{\eta}_{i\bullet}) = \frac{\exp(\eta_{ir})}{\sum_{s=1}^m \exp(\eta_{is})}, \tag{1}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^\top$, $\boldsymbol{\eta} = (\eta_{is})_{(i,s)\in[n]\times[m]} = \boldsymbol{X}\boldsymbol{W}^\top + \mathbf{1}\boldsymbol{b}^\top$, $\boldsymbol{\eta}_{i\bullet} = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{im}]^\top$, and $\boldsymbol{\eta}_{\bullet s} = [\eta_{1s}, \eta_{2s}, \ldots, \eta_{ns}]^\top$.

In particular, feature subset selection of the MNL model is defined as the selection of a feature subset that maximizes the likelihood under a constraint that the cardinality of the subset does not exceed given k. We formulate this as the following MIO problem:

$$\begin{aligned}
\underset{\boldsymbol{b},\boldsymbol{W},\boldsymbol{\eta},\boldsymbol{z}\in S_k^p}{\text{minimize}} \quad & \sum_{i=1}^n \ell^{\mathrm{MNL}}(y_i, \boldsymbol{\eta}_{i\bullet}) \\
\text{subject to} \quad & \boldsymbol{\eta}_{i\bullet} = \boldsymbol{W}\boldsymbol{x}_{i\bullet} + \boldsymbol{b}, && \forall i \in [n], \quad (2)\\
& z_j = 0 \;\Rightarrow\; \boldsymbol{w}_{\bullet j} = \mathbf{0}, && \forall j \in [p], \quad (3)\\
& \mathbf{1}^\top \boldsymbol{w}_{\bullet j} = 0, && \forall j \in [p], \quad (4)\\
& \mathbf{1}^\top \boldsymbol{b} = 0, && (5)
\end{aligned}$$

where $S_k^p = \{\boldsymbol{z} \in \{0,1\}^p \mid \mathbf{1}^\top \boldsymbol{z} \le k\}$, and ℓ^{MNL} is a loss function defined by the negative log-likelihood as

$$\begin{aligned}
\ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) &:= -\log \frac{\exp(\eta_y)}{\sum_{s=1}^m \exp(\eta_s)} \\
&= -\eta_y + \log \sum_{s=1}^m \exp(\eta_s); \qquad (6)
\end{aligned}$$

$z_j \in \{0, 1\}$ is defined as a binary variable that determines whether j-th feature is selected. For example, constraint (3) can be implemented as a linear inequality using the big-M method. This constraint expresses the group sparseness for features so that a common set

of variables is selected among all categories. In addition, the columns of $(\boldsymbol{b}, \boldsymbol{W})$ are centered through constraints (4) and (5) to ensure the uniqueness of optimal solutions (Simon et al., 2013).

Next, we consider the feature subset selection problem based on the maximum likelihood estimation with L_2-regularization. This problem can be formulated as follows (formulation (FS)):

$$\begin{aligned} \underset{\boldsymbol{b}, \boldsymbol{W}, \boldsymbol{\eta}, \boldsymbol{z} \in S_k^p}{\text{minimize}} \quad & \sum_{i=1}^{n} \ell(y_i, \boldsymbol{\eta}_{i\bullet}) + \frac{1}{2\gamma} \sum_{r=1}^{m} \|\boldsymbol{w}_{r\bullet}\|_2^2 \\ \text{subject to} \quad & (2), (3), \end{aligned}$$

where $\gamma > 0$ is a regularization parameter and $\ell : [m] \times \mathbb{R}^m \to [-\infty, +\infty]$ is a general loss function (e.g., the negative log-likelihood of generalized linear models (Nelder and Wedderburn, 1972)). We can remove the redundancy of coefficients by adding constraints (4) and (5) to problem (FS).

3 Outer Approximation Method

To construct an algorithm for solving (FS), we first define $(\mathrm{P}_{\boldsymbol{z}})$ as the (FS) with fixed $\boldsymbol{z} \in [0,1]^p$. Let us denote the optimal value of $(\mathrm{P}_{\boldsymbol{z}})$ by $c(\boldsymbol{z})$ as

$$c(\boldsymbol{z}) = \min_{\boldsymbol{b}, \boldsymbol{W}} \left\{ \sum_{i=1}^{n} \ell(y_i, \boldsymbol{W}\boldsymbol{x}_{i\bullet}^{\boldsymbol{z}} + \boldsymbol{b}) + \frac{1}{2\gamma} \sum_{r=1}^{m} \|\boldsymbol{w}_{r\bullet}\|_2^2 \right\},$$

where $\boldsymbol{x}_{i\bullet}^{\boldsymbol{z}} = [z_1 x_{i1},\, z_2 x_{i2},\, \ldots,\, z_p x_{ip}]^\top$. Note that constraints (4) and (5) do not affect the value of $c(\boldsymbol{z})$ when ℓ is ℓ^{MNL}.

It is clear that an optimal solution $\boldsymbol{z}^\star$ to (FS) is obtained by solving the following bilevel optimization problem:

$$\underset{\boldsymbol{z} \in S_k^p}{\text{minimize}} \quad c(\boldsymbol{z}). \tag{7}$$

In addition, optimal coefficients $(\boldsymbol{b}^\star, \boldsymbol{W}^\star)$ can be obtained by solving the continuous optimization problem $(\mathrm{P}_{\boldsymbol{z}^\star})$. Herein, we compute (7) by applying the outer approximation method (Bertsimas et al., 2017) to the problem (see Algorithm 1); Bonami et al. (2008) proposed this algorithm for general mixed-integer nonlinear optimization problems and reported that it is more useful than branch-and-bound methods for some instances.

The steps of Algorithm 1 are as follows. First, we express $c(\boldsymbol{z})$ as its first-order approximation at a given integer point $\boldsymbol{z}_0$. Because the minimization problem (7) then becomes a mixed-integer linear optimization problem, we can easily solve it and obtain a new integer point $\boldsymbol{z}_{t+1}$. Next, we add a new constraint of the first-order approximation at $\boldsymbol{z}_{t+1}$. These procedures are iterated until the true objective value $c(\boldsymbol{z}_t)$ becomes less than or equal to ζ_t, which is the optimal value of the problem after the linear constraints are added t times.

Algorithm 1 Outer approximation method

Require: $(\boldsymbol{X}, \boldsymbol{y})$, $k \in \mathbb{Z}_+$, $\gamma > 0$, $\boldsymbol{z}_0 \in S_k^p$
  $\boldsymbol{z}_1 \leftarrow \boldsymbol{z}_0$, $\zeta_1 \leftarrow -\infty$, $t \leftarrow 1$
  while $\zeta_t < c(\boldsymbol{z}_t)$ **do**
    $\boldsymbol{z}_{t+1}, \zeta_{t+1} \leftarrow \operatorname{argmin}_{\boldsymbol{z}, \zeta} \{\zeta : \boldsymbol{z} \in S_k^p, \zeta \geq c(\boldsymbol{z}_i) + \nabla c(\boldsymbol{z}_i)^\top (\boldsymbol{z} - \boldsymbol{z}_i),\ i \in [t]\}$
    $t \leftarrow t + 1$
  end while
  return $\boldsymbol{z}_t$

To accelerate this algorithm, we use a warm-start technique. Specifically, running this algorithm in ascending order of $k \in \{0, 1, \ldots, p\}$, we can use the optimal solution obtained at the $(k-1)$-th execution as the initial solution $\boldsymbol{z}_0$ at the k-th iteration.

4 Theoretical Results

In this section, we discuss the convergence of our algorithm. To ensure that an optimal solution $\boldsymbol{z}^\star$ is obtained by Algorithm 1 in a finite number of iterations, the loss function ℓ is subject to certain assumptions. First, we claim that the algorithm converges under these assumptions for a general case. Next, we prove that the loss function defined by (6) satisfies these assumptions.

4.1 Results for General Loss Functions

First, we consider a general loss function $\ell : [m] \times \mathbb{R}^m \to [-\infty, +\infty]$ with categories $y_i \in [m]$ and $\boldsymbol{\eta}_{i\bullet} = (\boldsymbol{w}_{r\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r)_{r \in [m]}$. For this loss function, we also consider the following optimization problem (P):

$$\begin{aligned} \underset{(\boldsymbol{b}, \boldsymbol{W}) \in \mathbb{R}^{m \times (p+1)}, \boldsymbol{\eta}}{\text{minimize}} \quad & \sum_{i=1}^{n} \ell(y_i, \boldsymbol{\eta}_{i\bullet}) + \frac{1}{2\gamma} \sum_{r=1}^{m} \|\boldsymbol{w}_{r\bullet}\|_2^2 \\ \text{subject to} \quad & (2), \\ & \boldsymbol{\eta}_{i\bullet} \in \operatorname{dom}\, \ell(y_i, \bullet), \qquad \forall i \in [n], \end{aligned}$$

where $\operatorname{dom}\, \ell(y, \bullet) = \{\boldsymbol{\eta} \in \mathbb{R}^m \mid \ell(y, \boldsymbol{\eta}) < +\infty\}$ for all $y \in [m]$. To prove the convergence of Algorithm 1, we consider the following assumptions about ℓ and (P):

1. the loss function $\ell(y, \bullet)$ is proper convex for all $y \in [m]$, and (P) is bounded and has an interior feasible solution;

2. the Fenchel conjugate $\hat{\ell}(y, \bullet)$ is continuous on its effective domain and strictly convex for all $y \in [m]$;

3. dom $\hat{\ell}(y, \bullet)$ is nonempty, bounded and closed for all $y \in [m]$;

where $\hat{\ell}(y, \bullet) : \mathbb{R}^m \to [-\infty, +\infty]$ is the Fenchel conjugate of $\ell(y, \bullet)$ and is defined as

$$\hat{\ell}(y, \boldsymbol{\alpha}) = \sup_{\boldsymbol{\eta}} \{\boldsymbol{\alpha}^\top \boldsymbol{\eta} - \ell(y, \boldsymbol{\eta}) \mid \boldsymbol{\eta} \in \mathbb{R}^m\}.$$

We note that $\ell(y, \bullet)$ is proper convex if Assumption 2 holds; this is because the conjugate of a proper convex function is also proper convex.

For the analysis, we derive the dual expression for $c(\boldsymbol{z})$ by the following theorem:

Theorem 1. *The dual problem (D) of problem (P) is formulated as*

$$\begin{aligned} \underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}}{\text{maximize}} \quad & -\sum_{i=1}^{n} \hat{\ell}(y_i, \boldsymbol{\alpha}_{i\bullet}) - \frac{\gamma}{2} \sum_{r=1}^{m} \|\boldsymbol{X}^\top \boldsymbol{\alpha}_{\bullet r}\|_2^2 \\ \text{subject to} \quad & \mathbf{1}^\top \boldsymbol{\alpha}_{\bullet r} = 0, \qquad \forall r \in [m], \qquad (8) \\ & \boldsymbol{\alpha}_i \in \operatorname{dom} \hat{\ell}(y_i, \bullet), \qquad \forall i \in [n]. \qquad (9) \end{aligned}$$

Moreover, the strong duality holds when the loss function ℓ and (P) *satisfy Assumption 1.*

The proof is given in the supplement. The strong duality of problem $(\mathrm{P}_{\boldsymbol{z}})$ holds for arbitrary $\boldsymbol{z} \in [0, 1]^p$ under Assumption 1 by Theorem 1. Indeed, problem (P) can be formulated to transform $(\mathrm{P}_{\boldsymbol{z}})$ by $\boldsymbol{x}_{\bullet j} \coloneqq z_j \boldsymbol{x}_{\bullet j}$. We can thus redefine the nonlinear function c as the optimal value of the following maximization problem named $(\mathrm{D}_{\boldsymbol{z}})$:

$$\begin{aligned} c(\boldsymbol{z}) = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}}{\text{maximize}} \quad & -\sum_{i=1}^{n} \hat{\ell}(y_i, \boldsymbol{\alpha}_{i\bullet}) - \frac{\gamma}{2} \sum_{j=1}^{p} z_j \left\|\boldsymbol{\alpha}^\top \boldsymbol{x}_{\bullet j}\right\|_2^2 \qquad (10) \\ \text{subject to} \quad & (8), (9). \end{aligned}$$

By using this dual expression of $c(\boldsymbol{z})$, the following two lemmas are established. Their proofs are given in the supplement.

Lemma 2. *Under Assumptions 2 and 3, c is continuously differentiable on $[0, 1]^p$, whereupon the differential is derived as*

$$\frac{\partial c(\boldsymbol{z})}{\partial z_j} = -\frac{\gamma}{2} \left\|\boldsymbol{\alpha}^\star(\boldsymbol{z})^\top \boldsymbol{x}_{\bullet j}\right\|_2^2, \quad \forall j \in [p], \qquad (11)$$

where $\boldsymbol{\alpha}^\star(\boldsymbol{z})$ is the optimal solution to $(\mathrm{D}_{\boldsymbol{z}})$.

Lemma 3. *The function c is a convex function on $[0, 1]^p$.*

From these lemmas, the following theorem is obtained with regard to the convergence of Algorithm 1.

Theorem 4. *Under Assumptions 1–3, Algorithm 1 converges to an optimal solution to (7) in a finite number of iterations.*

The proof is also given in the supplement.

Here we discuss the convergence properties mentioned in Bertsimas et al. (2017) and in the present paper. Because Bertsimas et al. (2017) considered only $m = 2$, our result is more general. We also note that they assumed only the convexity of ℓ. However, even when $m = 2$, Lemma 2 dictates that we cannot differentiate c if the conjugate $\hat{\ell}$ is not strictly convex; as such, our result is stricter than that of Bertsimas et al. (2017).

4.2 Results for Multinomial Logit Model

To show the convergence of Algorithm 1 for the MNL model, Assumptions 1–3 should be satisfied. We begin by deriving the conjugate $\hat{\ell}^{\mathrm{MNL}}$ and its effective domain by Proposition 3 in Lapin et al. (2018) as

$$\hat{\ell}^{\mathrm{MNL}}(y, \boldsymbol{\alpha}) = \begin{cases} \sum_{s \neq y} \alpha_s \log \alpha_s + \\ \quad (1 + \alpha_y) \log(1 + \alpha_y) & \text{if } \boldsymbol{\alpha} \in \mathcal{A}_y^{\mathrm{MNL}}, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\mathcal{A}_y^{\mathrm{MNL}} = \{\boldsymbol{\alpha} \in \mathbb{R}^m \mid \mathbf{1}^\top \boldsymbol{\alpha} = 0, \, \boldsymbol{\alpha}^{\setminus y} \in \Delta\},$$

where $\Delta = \{\boldsymbol{v} \mid \boldsymbol{v} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{v} \leq 1\}$ and $\boldsymbol{v}^{\setminus i} = [v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n]^\top$ for arbitrary $\boldsymbol{v} \in \mathbb{R}^n$. Herein, we use this definition to analyze properties and to provide problem $(\mathrm{D}_{\boldsymbol{z}})$. The following three propositions are established. Their proofs are given in the supplement.

Proposition 5. *The function $\ell^{\mathrm{MNL}}(y, \bullet)$ is proper convex for all $y \in [m]$, and problem* (P) *is bounded and has an interior feasible solution when ℓ is ℓ^{MNL}.*

Proposition 6. *The Fenchel conjugate $\hat{\ell}^{\mathrm{MNL}}(y, \bullet)$ is continuous on $\mathcal{A}_y^{\mathrm{MNL}}$ and strictly convex for all $y \in [m]$.*

Proposition 7. *The effective domain $\mathcal{A}_y^{\mathrm{MNL}}$ is nonempty, bounded, and closed for all $y \in [m]$.*

From Propositions 5–7, Assumptions 1–3 are satisfied for the MNL model and therefore the following corollary holds by Theorem 4.

Corollary 8. *Algorithm 1 converges to an optimal solution to (7) for the MNL model in a finite number of iterations.*

5 Approximations for Faster Computation

In Algorithm 1, the gradient of c defined by (11) must be calculated at each iteration. This means that the nonlinear optimization problem $(\mathrm{D}_{\boldsymbol{z}})$ must be solved many times, which is computationally expensive. In

this section, we suggest an efficient approximation for the nonlinear optimization problem to accelerate the calculation.

5.1 Methods for Binomial Logit Model

For simplicity, we begin by considering the binomial logit model, which is a special case of the MNL model with $m = 2$. We define the following function $f : \mathbb{R} \to \mathbb{R}$:

$$f(\eta) = \log(1 + \exp(\eta)).$$

When $m = 2$, the loss function ℓ^{BNL} is formulated as follows from the definition (6):

$$\ell^{\mathrm{BNL}}(y, \boldsymbol{\eta}) = f\left(u(\eta_2 - \eta_1)\right),$$

where $u = 1$ for $y = 1$, and $u = -1$ for $y = 2$.

The primal problem (P) can thus be reformulated as

$$\begin{aligned} \underset{\boldsymbol{b}, \boldsymbol{W}, \boldsymbol{\eta}}{\text{minimize}} \quad & \sum_{i=1}^{n} f(u_i(\eta_{i2} - \eta_{i1})) + \frac{1}{2\gamma} \sum_{r=1}^{2} \|\boldsymbol{w}_{r\bullet}\|_2^2 \\ \text{subject to} \quad & \boldsymbol{\eta}_i = \boldsymbol{W}\boldsymbol{x}_{i\bullet} + \boldsymbol{b}, \qquad \forall i \in [n], \end{aligned}$$

where $u_i = 1$ for $y_i = 1$, and $u_i = -1$ for $y_i = 2$. Similarly, we reformulate the dual problem (D) as

$$\begin{aligned} \underset{\boldsymbol{\alpha} \in \mathbb{R}^{n \times 2}}{\text{maximize}} \quad & -\sum_{i=1}^{n} \hat{f}(-\alpha_{iy_i}) - \frac{\gamma}{2} \sum_{r=1}^{2} \|\boldsymbol{X}^\top \boldsymbol{\alpha}_{\bullet r}\|_2^2 \\ \text{subject to} \quad & (8), (9), \end{aligned}$$

where the conjugate $\hat{f} : \mathbb{R} \to [-\infty, +\infty]$ is formulated as

$$\hat{f}(\alpha) = \begin{cases} \alpha \log \alpha + (1 - \alpha) \log(1 - \alpha) & \text{if } \alpha \in [0, 1], \\ +\infty & \text{otherwise.} \end{cases}$$

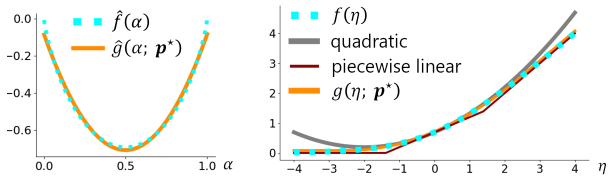
That is, if we approximate $\hat{f}$ by a convex quadratic function, subproblem ($\mathrm{D}_{\boldsymbol{z}}$) in Algorithm 1 becomes a convex quadratic optimization problem.

To accomplish this, we construct a polynomial regression model, and minimize the residual sum of squares as

$$\underset{\boldsymbol{p} \in \mathbb{R}^3}{\text{minimize}} \int_0^1 \left(\hat{g}(\alpha; \boldsymbol{p}) - \hat{f}(\alpha)\right)^2 \mathrm{d}\alpha,$$

where $\hat{g}(\alpha; \boldsymbol{p}) = p_3\alpha^2 + p_2\alpha + p_1$. This optimization problem can be solved analytically, and the optimal solution is $\boldsymbol{p}^\star = (-1/12, -5/2, 5/2)^\top$. Figure 1(a) shows that this approximated function fits well to $\hat{f}(\alpha)$.

The approximation $\hat{\ell}^{\mathrm{BNL}}(y, \boldsymbol{\alpha}) = \hat{f}(-\alpha_y) \approx \hat{g}(-\alpha_y; \boldsymbol{p}^\star)$ is therefore effective in approximating the subproblem in Algorithm 1 as a convex quadratic optimization problem when $m = 2$. This approximation makes our algorithm more efficient because convex quadratic optimization problems can be solved much faster than can nonlinear optimization problems.



(a) $\hat{f}(\alpha)$ (b) $f(\eta)$

Figure 1: Comparison with Various Approximation Methods for the Loss Function of the Binomial Logit Model

5.2 Interpretation of Proposed Approximation

We next consider the relationship between our approximation and existing approximation methods. Tanaka and Nakagawa (2014) and Sato et al. (2016) approximated the function f, and so we also consider the primal problem here. By Theorem 1, the primal problem is derived by calculating a double conjugate from the conjugate $\hat{g}(\alpha; \boldsymbol{p})$.

Proposition 9. *Let $g(\eta; \boldsymbol{p})$ be the conjugate of $\hat{g}(\alpha; \boldsymbol{p})$, and assume $p_3 > 0$. The function $g(\eta; \boldsymbol{p})$ is given by*

$$g(\eta; \boldsymbol{p}) = \begin{cases} -p_1 & \text{if } \eta < p_2, \\ (\eta - p_2)^2/4p_3 - p_1 & \text{if } \eta \in [p_2,\ p_2 + 2p_3], \\ \eta - (p_1 + p_2 + p_3) & \text{otherwise.} \end{cases}$$

The proof is given in the supplement. In particular,

$$g(\eta; \boldsymbol{p}^\star) = \begin{cases} 1/12 & \text{if } \eta < -5/2, \\ (\eta + 5/2)^2/10 + 1/12 & \text{if } |\eta| \le 5/2, \\ \eta + 1/12 & \text{otherwise.} \end{cases}$$

Consequently, replacing $\hat{f}$ with the convex quadratic function $\hat{g}$ corresponds to the following approximation of f:

$$f(\eta) \approx g(\eta; \boldsymbol{p}).$$

This replacement reveals that our method has the advantages of the approximations proposed by Tanaka and Nakagawa (2014) and Sato et al. (2016). Sato et al. (2016) approximated the function f by a piecewise linear function [**piecewise linear** in Figure 1(b)]. Because the gradient $\mathrm{d}f/\mathrm{d}\eta$ changes little when $|\eta| \gg 0$, this approximation is good when $|\eta| \gg 0$. However, the approximation is poor around $\eta = 0$, where the

gradients change sharply. By contrast, Tanaka and Nakagawa (2014) used a Maclaurin expansion of f [**quadratic** in Figure 1(b)]. This approximation fits well around $\eta = 0$ but is quite poor when $|\eta| \gg 0$. Proposition 9 shows that our method approximates f by a quadratic function around $\eta = 0$ and by linear functions when $|\eta| \gg 0$ (Figure 1(b)). The above discussion implies that our method has the advantages of these two existing methods, thus we expect that our quadratic approximation is effective for the subproblem in Algorithm 1.

5.3 Methods for Multinomial Logit Models

Finally, we discuss how to apply our approximation to the MNL model. As mentioned above, the posterior probability of categories is formulated as (1). The following approximation for the probability was derived by Titsias (2016):

$$\begin{aligned}\mathrm{P}(Y = r \mid \boldsymbol{x}_{i\bullet}, \boldsymbol{W}, \boldsymbol{b}) &= \frac{\exp(\boldsymbol{w}_{r\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r)}{\sum_{s=1}^m \exp(\boldsymbol{w}_{s\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r)} \\ &\approx \prod_{s \neq r} \frac{\exp(\boldsymbol{w}_{r\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r)}{\exp(\boldsymbol{w}_{r\bullet}^\top \boldsymbol{x}_{i\bullet} + b_r) + \exp(\boldsymbol{w}_{s\bullet}^\top \boldsymbol{x}_{i\bullet} + b_s)}.\end{aligned}$$

Then, by using this approximation, we approximate the loss function ℓ^{MNL} to ℓ^{Titsias} as

$$\begin{aligned}\ell^{\mathrm{MNL}}(y, \boldsymbol{\eta}) &= -\eta_y + \log \sum_{s=1}^m \exp(\eta_s) \\ &\approx \sum_{s \neq y} \log\left[1 + \exp(\eta_s - \eta_y)\right] = \ell^{\mathrm{Titsias}}(y, \boldsymbol{\eta}). \quad (12)\end{aligned}$$

Titsias (2016) observed empirically that this approximation is helpful for the maximum likelihood estimation of the MNL model. We expect that his approximation is also useful for the feature subset selection problem and consider applying the approximation to the loss function.

Proposition 10. *For the loss function ℓ^{Titsias}, the Fenchel conjugate $\hat{\ell}^{\mathrm{Titsias}}$ and its effective domain are derived as*

$$\hat{\ell}^{\mathrm{Titsias}}(y, \boldsymbol{\alpha}) = \begin{cases} \sum_{s \neq y}[\alpha_s \log \alpha_s + \\ \quad (1 - \alpha_s)\log(1 - \alpha_s)] & \text{if } \boldsymbol{\alpha} \in \mathcal{A}_y^{\mathrm{Titsias}}, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\mathcal{A}_y^{\mathrm{Titsias}} = \{\boldsymbol{\alpha} \in \mathbb{R}^m \mid \mathbf{1}^\top \boldsymbol{\alpha} = 0,\ \mathbf{0} \leq \boldsymbol{\alpha}^{\setminus y} \leq \mathbf{1}\}.$$

The proof is given in the supplement.

From this result, our approximation is applicable to the conjugate function for all $\boldsymbol{\alpha} \in \mathrm{dom}\ \hat{\ell}^{\mathrm{Titsias}}(y, \bullet)$ as

$$\begin{aligned}\hat{\ell}^{\mathrm{Titsias}}(y, \boldsymbol{\alpha}) &= \sum_{s \neq y} \hat{f}(\alpha_s) \\ &\approx \sum_{s \neq y} \hat{g}(\alpha_s; \boldsymbol{p}) = \hat{\ell}^{\mathrm{quad}}(y, \boldsymbol{\alpha}), \quad (13)\end{aligned}$$

where $\hat{\ell}^{\mathrm{quad}}$ is the approximated function. We then obtain the following result for the convergence of Algorithm 1.

Theorem 11. *Let the conjugate function $\hat{\ell}$ be $\hat{\ell}^{\mathrm{quad}}$. An optimal solution to* (FS) *can then be obtained in a finite number of iterations by Algorithm 1 when $p_3 > 0$.*

The proof is also given in the supplement. Consequently, we can approximate subproblem ($\mathrm{D}_{\boldsymbol{z}}$) by a convex quadratic optimization problem for the MNL model. From the discussion in Section 5.2 and empirical results by Titsias (2016), this is expected to be a high-quality approximation to the original model.

6 Numerical Experiments

To demonstrate effectiveness of our methods, we perform experiments with synthetic and real data. All experiments were conducted on a 64-bit machine with Intel Xeon 2.10 GHz processors, 8 cores, and 128 GB main memory. We implemented our methods in Julia 0.6.4 (Lubin and Dunning, 2015; Bezanson et al., 2017). Problem ($\mathrm{D}_{\boldsymbol{z}}$) was solved by IPOPT 0.4.0 (Wächter and Biegler, 2006), and the approximated problem of ($\mathrm{D}_{\boldsymbol{z}}$) and the mixed-integer linear optimization problem in Algorithm 1 were solved by Gurobi 8.0.0 (Gurobi Optimization, 2018). The proposed methods were compared with the L_1-regularized classification using the `GLMNet` package (Friedman et al., 2018). All outer approximation methods were implemented as a single tree search (Bonami et al., 2008) for computational efficiency. To implement these, we used the so-called lazy constraint callback of Gurobi.

6.1 Synthetic Data

The synthetic data were generated based on Simon et al. (2013). The numbers of samples, features, and categories of data are denoted by n, p, and m, respectively. A design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ was generated from the Gaussian distribution of $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\sigma_{ij} = \rho$ for $i \neq j$ and $\sigma_{ij} = 1$ otherwise. The matrix $\boldsymbol{W}^{\mathrm{oracle}} \in \mathbb{R}^{m \times p}$ is an oracle coefficient matrix; we set $\boldsymbol{w}_{\bullet j}^{\mathrm{oracle}} = \mathbf{0}$ if $j \notin \{\lfloor p(k-1)/k^{\mathrm{oracle}} \rfloor + 1 \mid k \in [k^{\mathrm{oracle}}]\}$, where k^{oracle} is the cardinality of the oracle feature subset, and each nonzero coefficient was sampled from $\mathrm{N}(0, 10/m^2)$. A chosen category was given as $y_i = \min\{r \in [m] \mid t_i \leq \sum_{s=1}^r \mathrm{P}(Y = s \mid \boldsymbol{x}_{i\bullet}, \boldsymbol{W}^{\mathrm{oracle}}, \mathbf{0})\}$, where t_i was drawn from a uniform distribution $\mathrm{U}(0, 1)$.

First, we confirm the relationship between computational time and incumbent objective values. As stated previously, Algorithm 1 tries to update its incumbent solution in each iteration. If the algorithm generates a good solution early on, a practical strategy is to terminate the algorithm at that time.

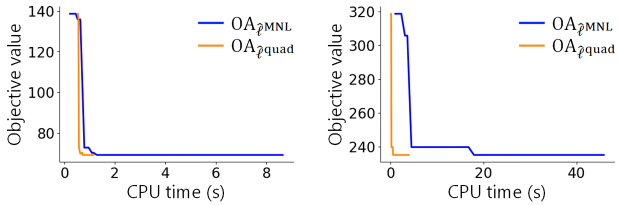
The proposed methods were compared with the L_1-

regularization method and a naïve outer approximation method for solving (FS) based on Bertsimas and King (2017). The latter method is easily applied to the MNL model. The L_1-regularization method was performed for all $\lambda \in \{0.001, 0.002, \ldots, 1.000\}$, and a feature subset that has k^{oracle} nonzero elements was selected. We ran our methods with $k = k^{\text{oracle}}$.

In the following Figures 2–8, $\mathbf{OA}_{\hat{\ell}^{\text{MNL}}}$ and $\mathbf{OA}_{\hat{\ell}^{\text{quad}}}$ correspond to Algorithm 1 with the exact loss (6) and that with the approximated loss (13), respectively. Similarly, ***L*$_1$-regularization** and **naïve OA** correspond to the L_1-regularization method and the extended method of Bertsimas and King (2017), respectively.

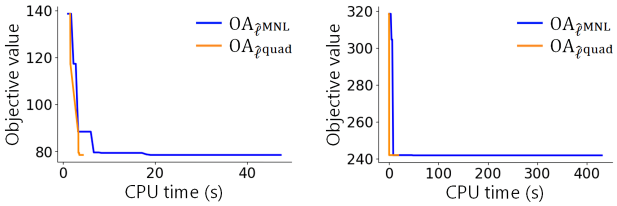
Figures 2 and 3 show results for $n = 200, p \in \{30, 50\}, m \in \{2, 5\}, k^{\text{oracle}} = 5$, and $\rho = 0.2$. Clearly, either an optimal solution or at least a solution whose objective value is close to optimal was obtained within seconds, thereby making early termination a good strategy for large instances. In addition, these figures show that the optimal solution can be obtained from the approximated model; moreover, it was found much faster than from the exact model.

Next, we investigated how the hyperparameter γ affects the computational time and the feature subset selection ability. Figures 4 and 5 show results for $n = 50, p = 20, m = 5, k^{\text{oracle}} = 5$, and $\rho = 0.2$. The results were averaged over 20 repetitions. When γ was small, Algorithm 1 with the exact loss (6) and that with the approximated loss (13) ran as fast as did the L_1-regularization method; this is because Algorithm 1 required relatively a few iterations. For all γ, our approximated method was approximately 10 times faster than our exact one. Figure 5 shows that the generalization abilities of our methods were the best when $\gamma = 10^{-1}$ and were greater than or equal to that of the L_1-regularization method. Also, Figure 4 shows that our exact method gave a high-quality solution within approximately 10 s whereas the extended method of Bertsimas and King (2017) took approximately 10^3 s to obtain the same solution. Additionally, the extended method of Bertsimas and King (2017)



(a) $m = 2$ (b) $m = 5$

Figure 2: Incumbent Objective Value ($n = 200, p = 30$)



(a) $m = 2$ (b) $m = 5$

Figure 3: Incumbent Objective Value ($n = 200, p = 50$)

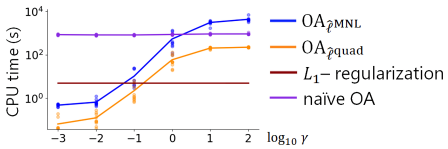


Figure 4: Computation Time ($n = 50$, $p = 20$, $m = 5$, $k^{\text{oracle}} = 5$, $\rho = 0.2$)

took almost the same computational time regardless of γ. That is, our methods are more appropriate for problem (FS) than is their method.

An analogous experiment was conducted for $n = 200, p = 30, m = 5, k^{\text{oracle}} = 5$, and $\rho = 0.2$. The computation of the proposed methods was terminated deliberately after 600 s. As shown in Figure 6, the solutions obtained with our methods have better generalization ability than do those of the L_1-regularization method when $\gamma \geq 10^{-1}$. These high-quality solutions were obtained even though these algorithms did not satisfy the termination condition $\zeta_t \geq c(\boldsymbol{z}_t)$ when $\gamma \geq 10^0$.

Finally, we confirm the relationship between the number of categories m and the feature subset selection ability. Figures 7 and 8 show results for $n = 200, p = 50, k^{\text{oracle}} = 8, \rho = 0.2$ and for $n = 200, p = 50, k^{\text{oracle}} = 8, \rho = 0.6$, respectively. The results were averaged over 30 repetitions. The computation of the proposed methods was terminated deliberately after 600 s. Except for the two cases of $(m, \rho) = (7, 0.2)$ and $(8, 0.2)$, our proposed methods performed better than did the L_1-regularization method. However, our approximated method performed poorer than did

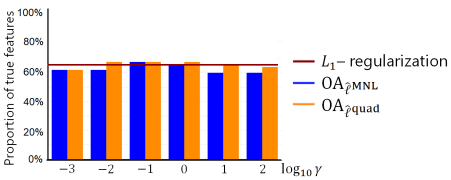


Figure 5: Proportion of True Features ($n = 50$, $p = 20$, $m = 5$, $k^{\text{oracle}} = 5$, $\rho = 0.2$)

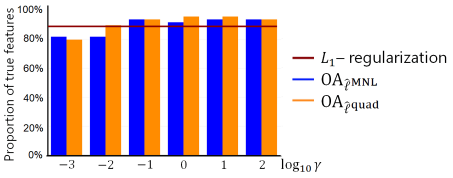


Figure 6: Proportion of True Features ($n = 200$, $p = 30$, $m = 5$, $k^{\text{oracle}} = 5$, $\rho = 0.2$)

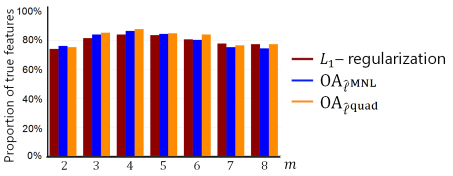


Figure 7: Proportion of True Features ($n = 200$, $p = 50$, $k^{\text{oracle}} = 8$, $\rho = 0.2$)

the exact one when $(m, \rho) = (8, 0.6)$. This is because approximation (12) fits well when m is relatively small. When $\rho = 0.6$, the efficiency of our methods was clearer than when $\rho = 0.2$. Overall, the proposed methods are thus more useful than is the L_1-regularization method, especially in the regime of strong correlation.

6.2 Real Data

Finally, we assessed the generalization ability of our methods with instances in the UCI Machine Learning Repository (Lichman, 2013) (Table 1). The ratio of training and test samples was 8:2, respectively, and we used a stratified sampling based on categories $\boldsymbol{y}$. We chose the best cardinality $k^\star$ in $\{0, 1, \ldots, p\}$ and the best L_2-regularization parameter $\gamma^\star$ in $\{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$ by means of 5-fold cross validation. Coefficients were estimated by likelihood maximization for all methods. That is, we solved $(\mathrm{P}_{\boldsymbol{z}})$ with $\boldsymbol{z}$ obtained based on the result of feature subset selection; we set $\gamma = 10^4$ at this time. The computation of the proposed methods was terminated deliberately after 180 s.

Table 2 lists the average correct classification rate (**CCR**) and the average likelihood value (**Likelihood**) for each method. For all instances, the prediction accuracies of our methods $\mathbf{OA}_{\hat{\ell}^{\mathrm{MNL}}}$ and $\mathbf{OA}_{\hat{\ell}^{\mathrm{quad}}}$ are greater than or equal to that of the L_1-regularization method L_1**-reg**. Moreover, for the Vehicle S. and Flags instances, our methods obtained sparser solutions than did the L_1-regularization method.

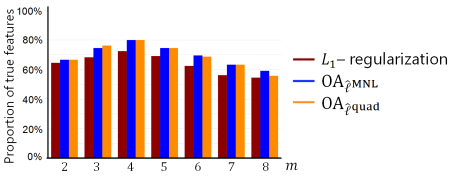


Figure 8: Proportion of True Features ($n = 200$, $p = 50$, $k^{\text{oracle}} = 8$, $\rho = 0.6$)

Table 1: Real Data

Data	n	p	m
Iris	150	4	3
Glass Identification (Glass I.)	214	9	6
Zoo	101	16	7
Vehicle Silhouettes (Vehicle S.)	846	18	4
Image Segmentation (Image S.)	210	19	7
Cardiotocography (Card.)	2126	31	3
Flags	194	66	8

Table 2: Results for Real Data (boldface numbers denote the best values)

Data	Method	CCR	Likelihood	$k^\star$
Iris	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	0.966	0.942	2
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	0.966	0.942	2
	L_1-reg.	0.966	0.873	**1**
Glass I.	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	0.628	0.394	2
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	**0.641**	0.371	2
	L_1-reg.	0.628	0.394	2
Zoo	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	0.714	0.519	2
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	**0.762**	0.344	2
	L_1-reg.	0.714	0.519	2
Vehicle S.	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	0.718	0.566	8
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	**0.741**	0.596	**7**
	L_1-reg.	0.688	0.521	11
Image S.	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	**0.905**	0.800	4
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	0.857	0.692	**3**
	L_1-reg.	0.857	0.676	**3**
Card.	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	0.984	0.894	4
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	0.984	0.894	4
	L_1-reg.	0.984	0.894	4
Flags	$\mathrm{OA}_{\hat{\ell}^{\mathrm{MNL}}}$	**0.641**	0.363	**3**
	$\mathrm{OA}_{\hat{\ell}^{\mathrm{quad}}}$	**0.641**	0.363	**3**
	L_1-reg.	0.590	0.350	4

7 Conclusions

In this paper, we proposed an MIO formulation and an outer approximation algorithm for the feature subset selection problem with L_2-regularization. The convergence of the proposed algorithm was proved for general loss functions, including the loss function for the MNL model. Moreover, to accelerate our algorithm, an approximation method for subproblem $(\mathrm{D}_{\boldsymbol{z}})$ was also proposed. By means of numerical experiments, we showed that the generalization ability of the proposed algorithm is greater than or equal to that of an L_1-regularization method for synthetic and real data.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers JP17K01246 and JP17K12983.

References

D. Bertsimas, J. Pauphilet, and B. V. Parys. Sparse classification and phase transitions: A discrete optimization perspective. arXiv preprint arXiv:1710.01352, 2017.

D. Bertsimas and A. King. Logistic regression: From art to science. Statistical Science, 32(3):367–384, 2017.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. Annals of Statistics, 44(2):813–852, 2016.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. SIAM Review 59(1):65–98, 2017.

P. Bonami, L. Biegler, A. Conn, G. Cornuéjols, I. Grossmann, C. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Wächter. An algorithmic framework for convex mixed integer nonlinear programs. Discrete Optimization 5(2):186–204, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. GLMNet: Lasso and elastic-net regularized generalized linear models. R package version 2.0-16, 2018.

P. M. Guadagni and J. D. C. Little. A logit model of brand choice calibrated on scanner data. Marketing Science, 2(3):203–238, 1983.

Gurobi Optimization, Inc. Gurobi Optimizer Reference Manual [http://www.gurobi.com/], 2018. Last accessed: 2018-09-05.

T. Hastie, R. Tibshirani, R. J. Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692, 2017.

T. Johnsen and R. W. Melicher. Predicting corporate bankruptcy and financial distress: Information value added by multinomial logit models. Journal of Economics and Business, 46(4):269–286, 1994.

R. Kohavi and G. H. John. Wrappers for feature subset selection. Artificial Intelligence, 97(1–2):273–324, 1997.

B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(6):957-968, 2005.

M. Lapin, M. Hein, and B. Schiele. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(7):1533–1554, 2018.

A. H.-L. Lau. A five-state financial distress prediction model. Journal of Accounting Research, 25(1):127-138, 1987.

M. Lichman. UCI machine learning repository [http://archive.ics.uci.edu/ml/], 2013. Last accessed: 2018-09-05.

J. Louviere and G. Woodworth. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. Journal of Marketing Research, 20(4):350–367, 1983.

M. Lubin and I. Dunning. Computing in operations research using julia. INFORMS Journal on Computing 27(2):238–248, 2015.

D. McFadden. Conditional logit analysis of qualitative choice behavior. In Frontiers of Econometrics, ed. P. Zarembka, New York: Academic Press, 1973.

R. Miyashiro and Y. Takano. Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research, 247(3):721–731, 2015.

J. Nelder and R. Wedderburn. Generalized linear models. Journal of the Royal Statistical Society, Series A, 135(3):370–384, 1972.

T. Sato, Y. Takano, and R. Miyashiro. Piecewise-linear approximation for feature subset selection in a sequential logit model. Journal of the Operations Research Society of Japan, 60(1):1–14, 2017.

T. Sato, Y. Takano, R. Miyashiro, and A. Yoshise. Feature subset selection for logistic regression via mixed integer optimization. Computational Optimization and Applications, 64(3):865–880, 2016.

N. Simon, J. Friedman, and T. Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. arXiv preprint arXiv:1311.6529, 2013.

K. Tanaka and H. Nakagawa. A method of corporate credit rating classification based on support vector machine and its validation in comparison of sequential logit model. Transactions of the Operations Research of Japan, 57:92–111, 2014.

M. K. Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities. In NIPS 2016, pages 4161–4169, 2016.

M. Vincent and N. R. Hansen. Sparse group lasso and high dimensional multinomial classification. Computational Statistics and Data Analysis, 71:771–786, 2014.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathematical Programming 106(1):25–57, 2006.