PAPER

Feature Subset Selection for Ordered Logit Model via Tangent-Plane-Based Approximation

Mizuho NAGANUMA[†], **Yuichi TAKANO**[††a)], *Nonmembers*, *and* **Ryuhei MIYASHIRO**[†††], *Member*

SUMMARY This paper is concerned with a mixed-integer optimization (MIO) approach to selecting a subset of relevant features from among many candidates. For ordinal classification, a sequential logit model and an ordered logit model are often employed. For feature subset selection in the sequential logit model, Sato et al. [22] recently proposed a mixed-integer linear optimization (MILO) formulation. In their MILO formulation, a univariate nonlinear function contained in the sequential logit model was represented by a tangent-line-based approximation. We extend this MILO formulation toward the ordered logit model, which is more commonly used for ordinal classification than the sequential logit model is. Making use of tangent planes to approximate a bivariate nonlinear function involved in the ordered logit model, we derive an MILO formulation for feature subset selection in the ordered logit model. Our computational results verify that the proposed method is superior to the L1-regularized ordered logit model in terms of solution quality.

key words: *optimization, statistics, feature subset selection, ordered logit model*

1. Introduction

Statistical analysis of a large amount of diverse data is increasingly important because of advances in information-gathering technology. A central task in such analysis is selecting a subset of relevant features (or explanatory variables) from among many candidates for model construction. This feature subset selection aids in understanding causal relations between explanatory and response variables. Moreover, the predictive performance of statistical models can be improved by elimination of redundant features because adverse effects of overfitting are mitigated.

Various computational algorithms have been proposed for selecting feature subsets [9], [12], [15], [17]. These include the stepwise method [11], L1-regularized regression [27], and metaheuristics [31]. Many methods are categorized as belonging to a class of heuristic algorithms, which perform well even on large-scale datasets. However, these algorithms can sometimes terminate with solutions of low quality because the optimality of obtained solutions (e.g., in the least-squares sense) is not an objective.

In contrast with such heuristic algorithms, mixed-integer optimization (MIO) approaches have the potential to find the best subset of features with respect to a given criterion function. One of these approaches was first proposed in the 1970s [3], and recently they have received renewed attention due to advances in algorithms and hardware [8], [16]. The MIO approaches have recently been used effectively for linear regression [14], [19], [20], logistic regression [7], [23], support vector machine [18], classification tree [5], and various applications [29], [30].

This paper is focused on the classification of ordinal categorical data [1]. Typical examples of such data are credit ratings of financial instruments, Likert-type items on questionnaires, and academic grading of students. For ordinal classification, the sequential logit model [2], [28] (or the continuation-ratio logit model) is sometimes employed, and the ordered logit model [21] (also called the cumulative logit model or the proportional odds model) is more commonly used [1]. In the following, we describe the previous MIO approaches to feature subset selection for ordinal classification (see also Appendix A).

The sequential logit model involves a univariate nonlinear function, which makes it hard to use an MIO approach to feature subset selection. To resolve this issue, Tanaka and Nakagawa [26] devised a mixed-integer quadratic optimization (MIQO) formulation based on a quadratic approximation of the nonlinear function. Sato et al. [22] derived a mixed-integer linear optimization (MILO) formulation by applying a tangent-line-based approximation to the nonlinear function. They also showed that their MILO formulation offers better solution quality than the MIQO formulation.

In line with Sato et al. [22], we propose a computationally tractable MILO formulation for feature subset selection in the ordered logit model. We make use of tangent planes to approximate a bivariate nonlinear function involved in the ordered logit model. Using this approximation, we reduce the feature subset selection for the ordered logit model to an MILO problem, which can be handled using standard MIO software. We also develop a heuristic algorithm to select a limited number of tangent planes that work well for approximation.

The efficacy of our method is assessed through computational experiments on several datasets from the UCI Machine Learning Repository [10]. The computational results demonstrate that our MILO formulation provides a better subset of features than does the L1-regularized ordered logit model in terms of the in-sample log-likelihood and out-of-

Manuscript received May 23, 2018.
Manuscript revised November 29, 2018.
Manuscript publicized February 21, 2019.

[†]The author is with Graduate School of Informatics and Engineering, The University of Electro-Communications, Chofu-shi, 182–8585 Japan.
[††]The author is with Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba-shi, 305–8577 Japan.
[†††]The author is with Institute of Engineering, Tokyo University of Agriculture and Technology, Koganei-shi, 184–8588 Japan.
a) E-mail: ytakano@sk.tsukuba.ac.jp

DOI: 10.1587/transinf.2018EDP7188

sample predictive performances.

2. Ordered Logit Model

Suppose that we are given n samples, p features, and m ordinal classes. A vector $\boldsymbol{x}_i := (x_{i1}, x_{i2}, \dots, x_{ip})^\top$ is composed of p features for each sample $i = 1, 2, \dots, n$. Each sample is also given a class label

$$\delta_{ik} := \begin{cases} 1 & \text{if the } i\text{th sample belongs to the } k\text{th class,} \\ 0 & \text{otherwise} \end{cases}$$

for each ordinal class $k = 1, 2, \dots, m$.

In the ordered logit model, the following linear regression model is employed for ordinal classification:

$$\boldsymbol{w}^\top \boldsymbol{x}_i + e_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip} + e_i,$$

where $\boldsymbol{w} := (w_1, w_2, \dots, w_p)^\top$ is a coefficient vector to be estimated and e_i is random noise in the ith sample. To categorize each sample to one of m ordinal classes, we introduce thresholds as follows:

$$-\infty = b_0 < b_1 < \cdots < b_m = \infty, \tag{1}$$

where $\boldsymbol{b} := (b_1, b_2, \dots, b_{m-1})^\top$ is an $(m-1)$-dimensional vector to be estimated. Accordingly, the ith sample is assigned to the kth class when the following relationship is satisfied: $b_{k-1} < \boldsymbol{w}^\top \boldsymbol{x}_i + e_i \le b_k$.

We assume that the random noises are mutually independent and that each noise e_i follows the same logistic distribution, given by

$$\Pr(e_i \le \xi) = \frac{1}{1 + \exp(-\xi)}.$$

The probability of the ith sample belonging to the kth class is expressed as

$$\begin{aligned} q_{ik} &:= \Pr(b_{k-1} - \boldsymbol{w}^\top \boldsymbol{x}_i < e_i \le b_k - \boldsymbol{w}^\top \boldsymbol{x}_i) \\ &= \frac{1}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k)} - \frac{1}{1 + \exp(\boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1})}. \end{aligned}$$

Hence, the occurrence probability of the observed data (i.e., δ_{ik}) is written as

$$\prod_{i=1}^{n} \prod_{k=1}^{m} (q_{ik})^{\delta_{ik}}.$$

The log-likelihood function to be maximized quantifies the plausibility of $\boldsymbol{b}$ and $\boldsymbol{w}$ based on the occurrence probability as follows:

$$\begin{aligned} L(\boldsymbol{b}, \boldsymbol{w}) &:= \log \prod_{i=1}^{n} \prod_{k=1}^{m} (q_{ik})^{\delta_{ik}} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{m} \delta_{ik} f(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k, \boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1}), \end{aligned} \tag{2}$$

where $f(u, v)$ is the bivariate nonlinear function

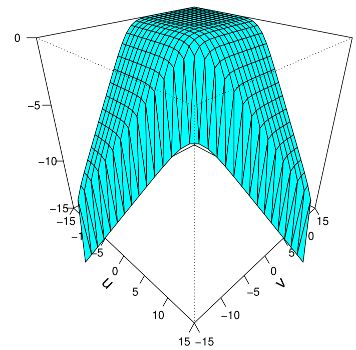


Fig. 1 Graph of $f(u, v)$.

$$f(u, v) := \log\left(\frac{1}{1 + \exp(u)} - \frac{1}{1 + \exp(v)}\right) \quad (u < v). \tag{3}$$

Figure 1 shows a graph of the bivariate nonlinear function (3). By straightforward calculation, the Hessian matrix of $f(u, v)$ is negative definite, so $f(u, v)$ is concave, as seen in Fig. 1.

3. Mixed-Integer Optimization Approach

This section presents an MILO formulation for feature subset selection in an ordered logit model.

3.1 Mixed-Integer Nonlinear Optimization Formulation

We focus on the problem of selecting a subset of features to maximize the log-likelihood function (2) subject to a constraint on the number of selected features. Before deriving our MILO formulation for this problem, we introduce a mixed-integer nonlinear optimization (MINLO) formulation in this subsection.

Let $\boldsymbol{z} := (z_1, z_2, \dots, z_p)^\top$ be a vector of binary decision variables for feature subset selection. That is,

$$z_j = \begin{cases} 1 & \text{if the } j\text{th feature is selected,} \\ 0 & \text{otherwise.} \end{cases}$$

Using $\boldsymbol{z}$, the feature subset selection can be formulated as an MINLO problem,

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} \sum_{k=1}^{m} \delta_{ik} f(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k, \boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1}) && (4) \\ \text{s. t.} \quad & b_{k-1} + \varepsilon \le b_k \quad (k = 2, 3, \dots, m-1), && (5) \\ & z_j = 0 \;\Rightarrow\; w_j = 0 \quad (j = 1, 2, \dots, p), && (6) \\ & \sum_{j=1}^{p} z_j = \theta, && (7) \\ & \boldsymbol{b} \in \mathbb{R}^{m-1}, \boldsymbol{w} \in \mathbb{R}^p, \boldsymbol{z} \in \{0, 1\}^p, && (8) \end{aligned}$$

where ε is a sufficiently small positive number and θ is a user-defined parameter specifying the number of selected features through constraint (7). Note that all the decision variables are listed in constraint (8). Constraint (5) enforces monotonicity on the thresholds from Eq. (1). If $z_j = 0$, then the jth feature is deleted because its coefficient must be zero by constraint (6). These logical implications can be represented by using a big-M method or a special-ordered-set constraint of type 1 (e.g., see [4]).

3.2 Tangent-Plane-Based Approximation

The MINLO formulation (4)–(8) in Sect. 3.1 is correct; however, the objective function (4) to be maximized is a concave but nonlinear function, which is difficult to handle directly by MIO software. For this reason, we approximate the bivariate nonlinear function (3) by tangent planes.

Let $\{(u_\ell, v_\ell, f(u_\ell, v_\ell)) \mid \ell = 1, 2, \ldots, h\}$ be a set of points of tangency for the function $f(u, v)$; we will explain how to choose these h points in Sect. 3.3. The associated tangent planes are expressed as

$$g_\ell(u, v) := f_u(u_\ell, v_\ell)(u - u_\ell) + f_v(u_\ell, v_\ell)(v - v_\ell) + f(u_\ell, v_\ell),$$

where f_u and f_v denote the partial derivatives of $f(u, v)$ with respect to u and v, respectively.

The graph of a concave function lies below any tangent plane of the function (except at the point of tangency, see also Fig. 2). Accordingly, $f(u, v)$ can be approximated by the pointwise minimum of a family of the h tangent planes as follows:

$$f(u, v) \approx G_h(u, v) := \min\{g_\ell(u, v) \mid \ell = 1, 2, \ldots, h\}.$$

For each (u, v), we have

$$\begin{aligned} &\min\{g_\ell(u, v) \mid \ell = 1, 2, \ldots, h\} \\ &= \max\{t \mid t \le g_\ell(u, v) \quad (\ell = 1, 2, \ldots, h)\}, \end{aligned}$$

where t is an auxiliary decision variable. Therefore, the tangent-plane-based approximation $G_h(u, v)$ is rewritten as

$$\begin{aligned} &G_h(u, v) \\ &= \max\{t \mid t \le f_u(u_\ell, v_\ell)(u - u_\ell) + f_v(u_\ell, v_\ell)(v - v_\ell) \\ &\qquad + f(u_\ell, v_\ell) \quad (\ell = 1, 2, \ldots, h)\}. \end{aligned} \tag{9}$$

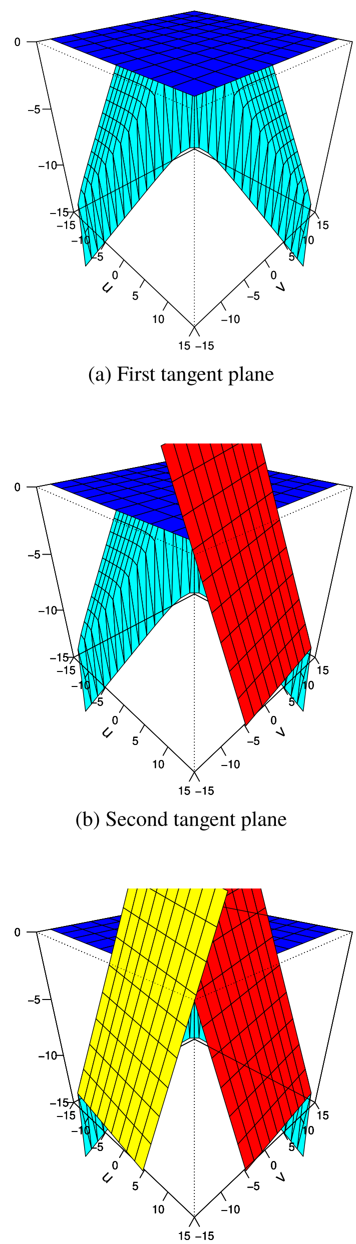
Let $\boldsymbol{T} := (t_{ik} \mid i = 1, 2, \ldots, n; k = 1, 2, \ldots, m)$ be a matrix of auxiliary decision variables for making a tangent-plane-based approximation

$$G_h(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k, \boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1}) \approx f(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k, \boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1})$$

in Eq. (4). By substituting $(u, v) = (\boldsymbol{w}^\top \boldsymbol{x}_i - b_k, \boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1})$ into Eq. (9), the MINLO problem (4)–(8) can be reduced to the following MILO problem:

$$\max \quad \sum_{i=1}^{n} \sum_{k=1}^{m} \delta_{ik} t_{ik} \tag{10}$$

$$\begin{aligned} \text{s. t.} \quad & t_{ik} \le f_u(u_\ell, v_\ell)(\boldsymbol{w}^\top \boldsymbol{x}_i - b_k - u_\ell) \\ & \qquad + f_v(u_\ell, v_\ell)(\boldsymbol{w}^\top \boldsymbol{x}_i - b_{k-1} - v_\ell) + f(u_\ell, v_\ell) \\ & (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m; \ell = 1, 2, \ldots, h), \end{aligned} \tag{11}$$



(a) First tangent plane

(b) Second tangent plane

(c) Third tangent plane

Fig. 2 Process of adding tangent planes.

$$b_{k-1} + \varepsilon \le b_k \quad (k = 2, 3, \ldots, m - 1), \tag{12}$$

$$z_j = 0 \;\Rightarrow\; w_j = 0 \quad (j = 1, 2, \ldots, p), \tag{13}$$

$$\sum_{j=1}^{p} z_j = \theta, \tag{14}$$

$$\boldsymbol{b} \in \mathbb{R}^{m-1}, \boldsymbol{T} \in \mathbb{R}^{n \times m}, \boldsymbol{w} \in \mathbb{R}^p, \boldsymbol{z} \in \{0, 1\}^p, \tag{15}$$

where all the decision variables are listed in constraint (15).

3.3 Heuristic Algorithm for Selecting Tangent Planes

The accuracy of the approximation proposed in Sect. 3.2 is greatly affected by which points of tangency are selected and the number h of points. If appropriate points of tangency are selected, then the MILO problem (10)–(15) approaches the original MINLO problem (4)–(8) as $h \to \infty$. However, as h increases, the size of the MILO problem also grows larger, which increases the computational burden. Thus, limiting the necessary number of points of tangency is crucial to practical approximation. We develop a simple heuristic algorithm for determining a set of points of tangency that provide a good approximation.

Our algorithm starts with the initial tangent plane(s). Figure 2 (a) shows the graph of the nonlinear function $f(u, v)$ and its initial tangent plane $g_1(u, v)$ on the bounded domain $\mathcal{F} := [-15, +15] \times [-15, +15]$ as an example. Note that the graph of $f(u, v)$ meets the plane at the point of tangency $(u_1, v_1, f(u_1, v_1))$ with $(u_1, v_1) := (-15, 15)$. We next add tangent planes one at a time at the point of tangency $(u, v, f(u, v))$ such that the gap between $f(u, v)$ and its tangent-plane-based approximation $G_\ell(u, v)$ is largest. That is,

$$(u_{\ell+1}, v_{\ell+1}) \in \arg\max\{G_\ell(u, v) - f(u, v) \mid (u, v) \in \mathcal{F}\}.$$

Specifically, we examine a finite number of lattice points $(u, v) \in \mathcal{F}$ and select $(u_{\ell+1}, v_{\ell+1})$ such that the gap (i.e., $G_\ell(u, v) - f(u, v)$) is largest. In this manner, the second tangent plane is added as shown in Fig. 2 (b), and then the third tangent plane is added as shown in Fig. 2 (c). We repeat this procedure until the number of tangent planes is equal to h.

4. Computational Results

This section evaluates the computational performance of our method for selecting a subset of features in the ordered logit model.

4.1 Experimental Design

We downloaded eight datasets for ordinal classification from the UCI Machine Learning Repository [10]. Table 1 lists these instances. In the table, n is the number of samples, p is the number of candidate features, and m is the number of ordinal classes.

For all instances, each integer and real variable was standardized to have mean zero and standard deviation one.

Table 1 List of instances.

Abbreviation	n	p	m	Original dataset [10]
`Wine-R`	1599	11	6	Wine Quality (red wine)
`Wine-W`	4898	11	7	Wine Quality (white wine)
`Skill`	3338	18	7	SkillCraft1 Master Table Dataset
`Choice`	1474	21	3	Contraceptive Method Choice
`Tnns-W`	118	31	7	Tennis Major Tournament (Wimbledon-women)
`Tnns-M`	113	33	7	Tennis Major Tournament (Wimbledon-men)
`Stdnt-M`	395	40	18	Student Performance (mathematics)
`Stdnt-P`	649	40	17	Student Performance (Portuguese language)

Each categorical variable was transformed into a set of the appropriate number of dummy variables. Variables for which more than 10% of values are missing were eliminated; after that, samples that still have missing values were eliminated. In the `Tnns-W` and `Tnns-M` instances, the variables "Player 1" and "Player 2" were removed because they are not suitable for prediction purposes.

The performances of the following methods were compared by computational experiment:

MILO(h): our MILO formulation (10)–(15), where h is the number of tangent planes;
L1-Reg: L1-regularized ordered logit model.

All computations were performed on a Windows computer with an Intel Core i7-4790 CPU (3.60 GHz) and 16 GB memory. The MILO problems were solved using IBM ILOG CPLEX 12.8.0.0 [13], where the indicator function implemented in CPLEX was used to impose the constraint (13). The L1-regularized ordered logit model was estimated using the `ordinalNet` package [32] in R 3.4.2, and a set of features with nonzero coefficients was selected. Since this package produced results for a sequence of regularization parameter values, so the computation time for $\theta = 5$ was equal to that for $\theta = 10$.

4.2 Selection of Tangent Planes

We begin by reporting the tangent planes selected by our heuristic algorithm, where $(u_1, v_1) = (-10, 10)$ and $(u_2, v_2) = (-0.01, 0.01)$ were used for the initial tangent planes, and subsequent points (u_ℓ, v_ℓ) $(\ell = 3, 4, \ldots, h)$ were chosen from a set of 0.01-spaced lattice points in the domain $\mathcal{F} := [-15, +15] \times [-15, +15]$.

Figure 3 shows the largest gap between $f(u, v)$ and its tangent-plane-based approximation

$$\max\{G_h(u, v) - f(u, v) \mid (u, v) \in \mathcal{F}\}$$

as a function of the number h of tangent planes on a semilogarithmic graph. In the figure, we see that the largest gap narrows with the number of tangent planes used. In particular, the gap decreases sharply until the ninth tangent plane is added; after that, it decreases slowly as the number of tangent planes is increased.

Figure 4 shows the (u, v) coordinates of the points of

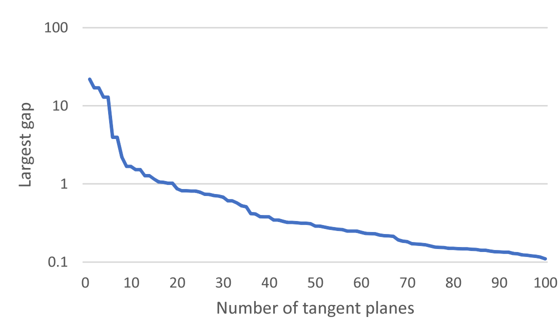
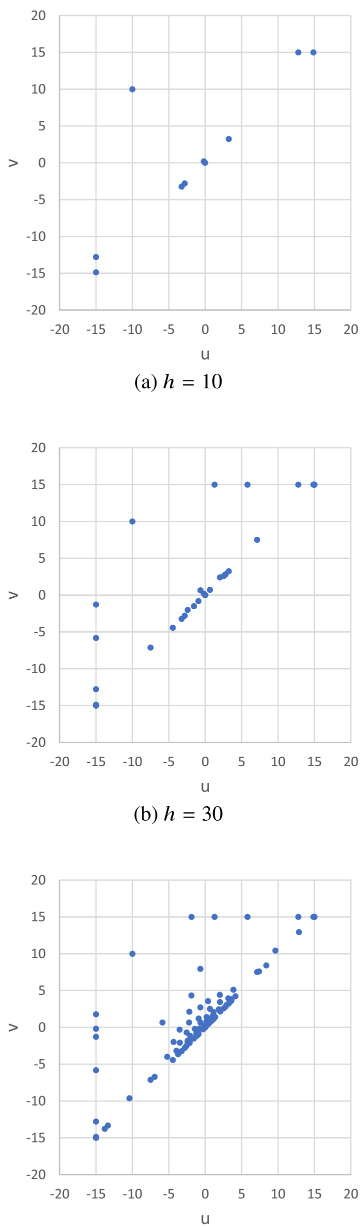


Fig. 3 Largest gap between $f(u, v)$ and its approximation.



(a) $h = 10$

(b) $h = 30$

(c) $h = 100$

Fig. 4 Points of tangency on uv plane.

tangency selected by our heuristic algorithm, where the number of tangent planes is $h \in \{10, 30, 100\}$. Note that Fig. 1 shows that $f(u, v)$ decreases indefinitely as (u, v) approaches the diagonal line $u = v$ and that $f(u, v)$ has a relatively large curvature around the origin $(u, v) = (0, 0)$. Consequently, many points were generated around the origin along the diagonal line on the uv coordinate plane in Fig. 4 (c).

4.3 Results of Feature Subset Selection

Tables 2 and 3 show the computational results of feature subset selection for the ordered logit model, where the number of selected features is $\theta = 5$ in Table 2 and $\theta = 10$ in Table 3. The column labeled "LogLike" shows the value of the log-likelihood (2), which was maximized using a selected subset of features; the largest log-likelihood values for each instance are indicated in bold. The column labeled "ObjVal" shows the optimal value of the objective function (10) (i.e., an approximate value of the maximum log-likelihood). The column labeled "Time (s)" shows the computation time in seconds.

Table 2 reveals that MILO(100) attained the largest log-likelihood values for all instances except `Wine-W`. We can also see that ObjVal approached LogLike as the number of tangent planes used in the MILO formulations increased. For example, in `Choice`, the LogLike value was −1460.81 and the ObjVal value was −876.86 for $h = 10$, while for $h = 100$ the values of LogLike and ObjVal were −1456.87 and −1432.43, respectively. In this case, the maximum log-likelihood was approximated to within about 2% by using 100 tangent planes.

The computation time of the MILO formulations increased greatly with the number of tangent planes. For instance, in `Choice` in Table 2, the MILO computation time was 15.76 s for $h = 10$ and 1021.32 s for $h = 100$. In contrast, the L1-Reg computations finished very quickly for all the instances; however, these yielded the worst log-likelihood value for most of the instances.

In Table 3, the largest log-likelihood value was provided by one of the MILO formulations for all instances except `Stdnt-P`. However, the differences in the log-likelihood between MILOs and L1-Reg in Table 3 were smaller than those in Table 2. Indeed, the log-likelihood values of MILO(10) were worse than those of L1-Reg for five out of eight instances. The main reason for this is that when many features need to be selected, more careful comparison is required. In fact, MILO(10) failed to find a subset of features of good quality due to the low accuracy of approximation based on a small number of tangent planes.

4.4 Out-of-Sample Predictive Performance

This subsection evaluates out-of-sample predictive performance of our method through two-fold cross-validation. A class label of each sample was predicted from the estimated probability of belonging to each ordinal class, and its accuracy (i.e., probability of correct answer) and root-mean-squared error (RMSE, based on difference between true and predicted labels) were calculated. We examined five instances: `Wine-R`, `Wine-W`, `Skill`, `Choice`, and `Stdnt-P` in the cross-validation. These instances were chosen because they contain enough samples to yield reliable results.

Tables 4 and 5 show the results of the cross-validation, where the number of tangent planes is $h = 100$, and the number of selected features is $\theta = 5$ (Table 4) or $\theta = 10$ (Table 5). The better accuracy/RMSE values between MILO(100) and L1-Reg are bold-faced in these tables. For $\theta = 5$ (Table 4), MILO(100) was better than L1-Reg in terms of accuracy and RMSE values for `Wine-R` and `Stdnt-P`; for $\theta = 10$ (Table 5), MILO(100) obtained better accuracy and RMSE

Table 2 Results of feature subset selection ($\theta = 5$).

Instance	n	p	m	Method	LogLike	ObjVal	Time (s)
`Wine-R`	1599	11	6	MILO(10)	**−1548.74**	−493.81	3.92
				MILO(30)	**−1548.74**	−1027.42	18.02
				MILO(100)	**−1548.74**	−1499.89	68.19
				L1-Reg	**−1548.74**	—	4.74
`Wine-W`	4898	11	7	MILO(10)	−5497.53	−1488.79	28.77
				MILO(30)	**−5497.41**	−3981.81	136.00
				MILO(100)	−5497.53	−5350.39	2214.23
				L1-Reg	−5502.91	—	16.85
`Skill`	3338	18	7	MILO(10)	−4548.52	−2104.42	52.60
				MILO(30)	−4550.57	−3908.59	262.77
				MILO(100)	**−4544.43**	−4430.16	2856.46
				L1-Reg	−4572.19	—	11.59
`Choice`	1474	21	3	MILO(10)	−1460.81	−876.86	15.76
				MILO(30)	−1457.78	−1330.67	119.23
				MILO(100)	**−1456.87**	−1432.43	1021.32
				L1-Reg	−1473.71	—	0.28
`Tnns-W`	118	31	7	MILO(10)	−144.91	−96.03	17.72
				MILO(30)	−146.04	−129.21	104.18
				MILO(100)	**−144.69**	−143.18	505.41
				L1-Reg	−149.84	—	1.68
`Tnns-M`	113	33	7	MILO(10)	**−131.42**	−90.72	11.40
				MILO(30)	**−131.42**	−117.11	71.46
				MILO(100)	**−131.42**	−130.25	201.80
				L1-Reg	−136.63	—	1.44
`Stdnt-M`	395	40	18	MILO(10)	−522.55	−246.09	18.50
				MILO(30)	−522.55	−424.19	118.79
				MILO(100)	**−522.43**	−509.82	387.37
				L1-Reg	−523.28	—	7.16
`Stdnt-P`	649	40	17	MILO(10)	−791.84	−319.72	43.23
				MILO(30)	−791.65	−602.21	151.12
				MILO(100)	**−789.83**	−768.88	795.01
				L1-Reg	−791.38	—	8.28

Table 3 Results of feature subset selection ($\theta = 10$).

Instance	n	p	m	Method	LogLike	ObjVal	Time (s)
`Wine-R`	1599	11	6	MILO(10)	−1538.38	−486.94	2.79
				MILO(30)	−1538.38	−1012.38	18.02
				MILO(100)	**−1538.01**	−1489.75	41.50
				L1-Reg	**−1538.01**	—	4.74
`Wine-W`	4898	11	7	MILO(10)	**−5450.54**	−1478.68	14.20
				MILO(30)	−5450.58	−3928.87	55.22
				MILO(100)	−5450.58	−5305.49	1455.11
				L1-Reg	−5450.58	—	16.85
`Skill`	3338	18	7	MILO(10)	−4493.28	−2066.50	46.99
				MILO(30)	−4491.03	−3845.88	125.05
				MILO(100)	**−4490.59**	−4373.80	947.80
				L1-Reg	−4492.93	—	11.59
`Choice`	1474	21	3	MILO(10)	−1443.31	−867.96	56.30
				MILO(30)	**−1440.88**	−1311.46	345.56
				MILO(100)	**−1440.88**	−1416.80	4215.52
				L1-Reg	−1441.66	—	0.28
`Tnns-W`	118	31	7	MILO(10)	**−137.34**	−86.61	84.02
				MILO(30)	−138.67	−121.53	336.76
				MILO(100)	−139.04	−135.48	2984.95
				L1-Reg	−141.75	—	1.68
`Tnns-M`	113	33	7	MILO(10)	−127.65	−85.72	30.47
				MILO(30)	−127.79	−111.90	150.32
				MILO(100)	**−127.27**	−125.64	683.52
				L1-Reg	−130.41	—	1.44
`Stdnt-M`	395	40	18	MILO(10)	−518.95	−241.22	116.86
				MILO(30)	−518.09	−417.01	1840.31
				MILO(100)	**−517.66**	−504.87	5103.93
				L1-Reg	−518.48	—	7.16
`Stdnt-P`	649	40	17	MILO(10)	−782.16	−312.97	1497.83
				MILO(30)	−781.98	−594.53	4263.90
				MILO(100)	−782.45	−760.45	48464.35
				L1-Reg	**−781.94**	—	8.28

Table 4 Results of cross-validation ($\theta = 5$).

Instance	n	p	m	Accuracy (%) MILO(100)	Accuracy (%) L1-Reg	RMSE MILO(100)	RMSE L1-Reg
`Wine-R`	1599	11	6	**59.0**	58.6	**0.711**	0.718
`Wine-W`	4898	11	7	**51.9**	51.4	0.817	**0.812**
`Skill`	3338	18	7	**39.3**	38.8	1.048	**1.042**
`Choice`	1474	21	3	**44.9**	44.8	0.819	**0.813**
`Stdnt-P`	649	40	17	**48.5**	45.2	**1.152**	1.161

Table 5 Results of cross-validation ($\theta = 10$).

Instance	n	p	m	Accuracy (%) MILO(100)	Accuracy (%) L1-Reg	RMSE MILO(100)	RMSE L1-Reg
`Wine-R`	1599	11	6	59.3	**59.4**	0.708	**0.706**
`Wine-W`	4898	11	7	**52.2**	**52.2**	**0.815**	**0.815**
`Skill`	3338	18	7	**40.0**	39.8	**1.032**	1.034
`Choice`	1474	21	3	**46.7**	45.7	**0.805**	**0.805**
`Stdnt-P`	649	40	17	**46.9**	46.5	**1.156**	1.157

values in a majority of cases.

5. Conclusion

This paper dealt with the feature subset selection problem for the ordered logit model. We formulated it as an MILO problem by applying tangent-plane-based approximation to the bivariate nonlinear function. We also developed a heuristic algorithm to select a limited number of tangent planes suitable for approximation. The computational results confirmed that our method was effective in finding a subset of features of good quality, comparing with the L1-regularized ordered logit model.

Our MILO formulation has the potential to provide the best subset of features when sufficiently many tangent planes are used for approximation. However, proving the optimality (or approximation accuracy) of the obtained solutions can be computationally intensive in this approach. In contrast, heuristic approaches, represented here by L1-regularized regression, can complete the search process quickly at the cost of giving up potential optimality of the obtained solutions. For practical purposes, it is necessary to choose between the two approaches according to the intended use of feature subset selection.

A future direction of study is to extend our approximation framework to other logit models. We will also consider modifying our heuristic algorithm to improve the accuracy of the tangent-plane-based approximation. Additionally, MIO approaches to eliminating multicollinearity have been studied in recent years [6], [24], [25], so such methods could be incorporated in our MILO formulation to reduce adverse effects of multicollinearity on the ordered logit model.

Acknowledgments

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

References

[1] A. Agresti, Analysis of Ordinal Categorical Data, John Wiley & Sons, New York, 2010.
[2] T. Amemiya, "Qualitative response models: A survey," Journal of Economic Literature, vol.19, no.4, pp.1483–1536, 1981.
[3] T.S. Arthanari and Y. Dodge, Mathematical Programming in Statistics, John Wiley & Sons, New York, 1993.
[4] E.M.L. Beale and J.A. Tomlin, "Special facilities in a general mathematical programming system for nonconvex problems using ordered sets of variables," Proc. Fifth International Conference on Operational Research, pp.447–454, 1970.
[5] D. Bertsimas and J. Dunn, "Optimal classification trees," Machine Learning, vol.106, no.7, pp.1039–1082, 2017.
[6] D. Bertsimas and A. King, "OR forum—An algorithmic approach to linear regression," Operations Research, vol.64, no.1, pp.2–16, 2016.
[7] D. Bertsimas and A. King, "Logistic regression: From art to science," Statistical Science, vol.32, no.3, pp.367–384, 2017.
[8] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," The Annals of Statistics, vol.44, no.2, pp.813–852, 2016.
[9] A.L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artificial Intelligence, vol.97, no.1–2, pp.245–271, 1997.
[10] D. Dua and E.K. Taniskidou, "UCI machine learning repository," School of Information and Computer Science, University of California, http://archive.ics.uci.edu/ml, accessed May 4, 2018.
[11] M.A. Efroymson, "Multiple regression analysis," Mathematical Methods for Digital Computers, eds. A. Ralston and H.S. Wilf, pp.191–203, John Wiley & Sons, New York, 1960.
[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol.3, pp.1157–1182, 2003.
[13] IBM, "IBM ILOG CPLEX optimization studio," IBM, https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/, accessed May 4, 2018.
[14] K. Kimura and H. Waki, "Minimization of Akaike's information criterion in linear regression analysis via mixed integer nonlinear program," Optimization Methods and Software, vol.33, no.3, pp.633–649, 2018.
[15] R. Kohavi and G.H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol.97, no.1–2, pp.273–324, 1997.
[16] H. Konno and R. Yamamoto, "Choosing the best set of variables in regression analysis using integer programming," Journal of Global

Optimization, vol.44, no.2, pp.273–282, 2009.
[17] H. Liu and H. Motoda, eds., Computational Methods of Feature Selection, Chapman & Hall/CRC, Boca Raton, 2007.
[18] S. Maldonado, J. Pérez, R. Weber, and M. Labbé, "Feature selection for support vector machines via mixed integer linear programming," Information Sciences, vol.279, pp.163–175, 2014.
[19] R. Miyashiro and Y. Takano, "Subset selection by Mallows' C_p: A mixed integer programming approach," Expert Systems with Applications, vol.42, no.1, pp.325–331, 2015.
[20] R. Miyashiro and Y. Takano, "Mixed integer second-order cone programming formulations for variable selection in linear regression," European Journal of Operational Research, vol.247, no.3, pp.721–731, 2015.
[21] P. McCullagh, "Regression models for ordinal data," Journal of the Royal Statistical Society, Series B (Methodological), vol.42, no.2, pp.109–142, 1980.
[22] 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, vol.60, no.1, pp.1–14, 2017.
[23] T. Sato, Y. Takano, R. Miyashiro, and A. Yoshise, "Feature subset selection for logistic regression via mixed integer optimization," Computational Optimization and Applications, vol.64, no.3, pp.865–880, 2016.
[24] R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, and T. Matsui, "Best subset selection for eliminating multicollinearity," Journal of the Operations Research Society of Japan, vol.60, no.3, pp.321–336, 2017.
[25] R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, and T. Matsui, "Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor," Journal of Global Optimization, vol.73, no.2, pp.431–446, 2019.
[26] 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," Trans. Operations Research Society of Japan, vol.57, pp.92–111, 2014 (in Japanese).
[27] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society, Series B (Methodological), vol.58, no.1, pp.267–288, 1996.
[28] G. Tutz, "Sequential models in categorical regression," Computational Statistics & Data Analysis, vol.11, no.3, pp.275–295, 1991.
[29] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," Machine Learning, vol.102, no.3, pp.349–391, 2016.
[30] Z.T. Wilson and N.V. Sahinidis, "The ALAMO approach to machine learning," Computers & Chemical Engineering, vol.106, pp.785–795, 2017.
[31] S.C. Yusta, "Different metaheuristic strategies to solve the feature selection problem," Pattern Recognition Letters, vol.30, no.5, pp.525–534, 2009.
[32] M.J. Wurm, P.J. Rathouz, and B.M. Hanlon, "Regularized ordinal regression and the ordinalNet R package," arXiv preprint arXiv:1706.05003, June 2017.

Appendix A: List of Abbreviations for Mixed-Integer Optimization

MIO Mixed-Integer Optimization

MILO Mixed-Integer Linear Optimization
Formulation for sequential logit model [22]
Formulation for ordered logit model (10)–(15)

MIQO Mixed-Integer Quadratic Optimization
Formulation for sequential logit model [26]

MINLO Mixed-Integer Nonlinear Optimization
Formulation for ordered logit model (4)–(8)

Mizuho Naganuma received her Bachelor's degree in Information Science from Senshu University in 2018. She is currently a master's course student at the University of Electro-Communications. Her research interests are statistics and machine learning.

Yuichi Takano is an associate professor in the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. From the University of Tsukuba, he received his Bachelor's degree in Policy and Planning Sciences in 2005, Master's degree in Engineering in 2007, and Doctorate in Engineering in 2010. His primary research interests are mathematical optimization and its application to financial engineering and machine learning.

Ryuhei Miyashiro received the B.E., M.E., and Ph.D. degrees from the University of Tokyo. He is presently an associate professor at Tokyo University of Agriculture and Technology. His research interests include mathematical programming and combinatorial optimization. He is a member of ORSJ and IEICE.