Journal of the Operations Research Society of Japan
Vol. 60, No. 1, January 2017, pp. 1–14
© The Operations Research Society of Japan

PIECEWISE-LINEAR APPROXIMATION FOR FEATURE SUBSET SELECTION IN A SEQUENTIAL LOGIT MODEL

Toshiki Sato
University of Tsukuba

Yuichi Takano
Senshu University

Ryuhei Miyashiro
Tokyo University of Agriculture and Technology

(Received October 7, 2015; Revised July 21, 2016)

Abstract This paper concerns a method of selecting a subset of features for a sequential logit model. Tanaka and Nakagawa (2014) proposed a mixed integer quadratic optimization formulation for solving the problem based on a quadratic approximation of the logistic loss function. However, since there is a significant gap between the logistic loss function and its quadratic approximation, their formulation may fail to find a good subset of features. To overcome this drawback, we apply a piecewise-linear approximation to the logistic loss function. Accordingly, we frame the feature subset selection problem of minimizing an information criterion as a mixed integer linear optimization problem. The computational results demonstrate that our piecewise-linear approximation approach found a better subset of features than the quadratic approximation approach.

Keywords: Optimization, statistics, feature subset selection, sequential logit model, piecewise-linear approximation, information criterion

1. Introduction

The analysis of ordinal categorical data [3, 24] is required in various areas including finance, econometrics, and bioinformatics. For instance, credit ratings of financial instruments are typical ordinal categorical data, and thus, many previous studies have analyzed such data by means of ordinal classification models (see, e.g., [1, 13, 29]). A sequential logit model [5, 18, 34], also known as the continuation ratio model [2, 15], is a commonly used ordinal classification model. It predicts an ordinal class label for each sample by successively applying separate logistic regression models. One can find various applications of sequential logit models: Kahn and Morimune [18] used this model to explain the duration of unemployment of workers; Weiler [36] investigated the choice behavior of potential attendees in higher education institutions; Fu and Wilmot [16] estimated dynamic travel demand caused by hurricane evacuation.

In order to enhance the reliability of these data analyses, it is critical to carefully choose a set of relevant features for model construction. Such a feature subset selection problem is of essential importance in statistics, data mining, artificial intelligence, and machine learning (see, e.g., [12, 17, 21, 25]). The mixed integer optimization (MIO) approach to feature subset selection has recently received a lot of attention as a result of algorithmic advances and hardware improvements (see, e.g., [10, 11, 22, 23, 26, 27]). In contrast to heuristic algorithms, e.g., stepwise regression [14], L_1-penalized regression [6, 20], and metaheuristic strategies [37], the MIO approach has the potential of providing an optimality guarantee for the selected set of features under a given goodness-of-fit measure.

Tanaka and Nakagawa [33] recently devised an MIO formulation for feature subset se-

lection in a sequential logit model. It is hard to exactly solve the feature subset selection problem for a sequential logit model, because its objective contains a nonlinear function called the logistic loss function. To resolve this issue, they employed a quadratic approximation of the logistic loss function. The resultant mixed integer quadratic optimization (MIQO) problem can be solved with standard mathematical optimization software; however, there is a significant gap between the logistic loss function and its quadratic approximation. As a result, the MIQO formulation may fail to find a good-quality solution to the original feature subset selection problem.

The purpose of this paper is to give a novel MIO formulation for feature subset selection in a sequential logit model. Sato et al. [30] used a piecewise-linear approximation in the feature subset selection problem for binary classification. In line with Sato et al. [30], we shall apply a piecewise-linear approximation to the sequential logit model for ordinal multi-class classification. Consequently, the problem is posed as a mixed integer linear optimization (MILO) problem. This approach is capable of approximating the logistic loss function more accurately than the quadratic approximation does. Moreover, our MILO formulation has the advantage of selecting a set of features with an optimality guarantee on the basis of an information criterion, such as the Akaike information criterion (AIC, [4]) or Bayesian information criterion (BIC, [32]).

The effectiveness of our MILO formulation is assessed through computational experiments on several datasets from the UCI Machine Learning Repository [7]. The computational results demonstrate that our piecewise-linear approximation approach found a better subset of features than the quadratic approximation approach did.

2. Sequential Logit Model

Let us suppose that we are given n samples of pairs, $(\boldsymbol{x}_i, y_i)$ for $i = 1, 2, \ldots, n$. Here, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ is a p-dimensional feature vector, and $y_i \in \{1, 2, \ldots, m+1\}$ is a ordinal class label to be predicted for each sample $i = 1, 2, \ldots, n$. In the sequential logit model for ordinal classification, we sequentially apply the following logistic regression models in order to predict a class label of each sample (see, e.g., [5, 18, 34]),

$$q_k(\boldsymbol{x}) = \Pr(y = k \mid y \geq k, \boldsymbol{x}) = \frac{1}{1 + \exp(-(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k))} \quad (k = 1, 2, \ldots, m), \tag{2.1}$$

where the intercept, b_k, and the p-dimensional coefficient vector, $\boldsymbol{w}_k = (w_{1k}, w_{2k}, \ldots, w_{pk})^\top$, are parameters to be estimated.

As shown in Figure 1, a feature vector $\boldsymbol{x}$ is moved into class 1 with a probability $q_1(\boldsymbol{x})$. In the next step, it falls into class 2 with a probability $(1 - q_1(\boldsymbol{x}))q_2(\boldsymbol{x})$. In the similar manner, it reaches class k with a probability $(1 - q_1(\boldsymbol{x}))(1 - q_2(\boldsymbol{x})) \cdots (1 - q_{k-1}(\boldsymbol{x}))q_k(\boldsymbol{x})$.

Here, we define

$$\delta_k(y) = \begin{cases} 1 & \text{if } y = k, \\ 0 & \text{otherwise,} \end{cases} \qquad (k = 1, 2, \ldots, m). \tag{2.2}$$

It then follows that

$$1 - \sum_{s=1}^{k} \delta_s(y_i) = \begin{cases} 1 & \text{if } k < y_i, \\ 0 & \text{otherwise,} \end{cases} \qquad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m). \tag{2.3}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

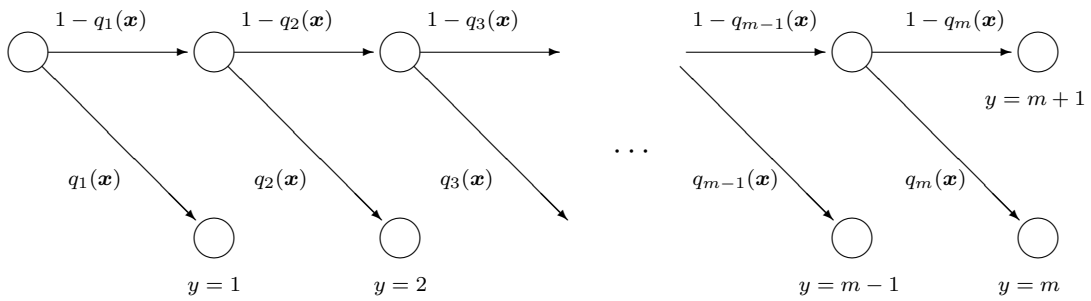

Figure 1: Diagram of sequential logit model

Therefore, the occurrence probability of a sample $(\boldsymbol{x}_i, y_i)$ is modeled as follows:

$$\prod_{k=1}^{m} \Big(1 - q_k(\boldsymbol{x}_i)\Big)^{1-\sum_{s=1}^{k} \delta_s(y_i)} \Big(q_k(\boldsymbol{x}_i)\Big)^{\delta_k(y_i)} \quad (i = 1, 2, \ldots, n). \tag{2.4}$$

We refer to model (2.4) as a forward sequential logit model because binary classification models (2.1) are applied in order from $k = 1$ to $k = m$. We can also consider the backward model that makes binary classification in the reverse order from $k = m$ to $k = 1$. It is known that these two models do not produce the same results (see [33]).

The maximum likelihood estimation method estimates the parameters, $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)^\top$ and $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m)$, so that the log likelihood function, $L(\boldsymbol{b}, \boldsymbol{W})$, is maximized:

$$\begin{aligned}
&L(\boldsymbol{b}, \boldsymbol{W}) = \log \prod_{i=1}^{n} \prod_{k=1}^{m} \Big(1 - q_k(\boldsymbol{x}_i)\Big)^{1-\sum_{s=1}^{k} \delta_s(y_i)} \Big(q_k(\boldsymbol{x}_i)\Big)^{\delta_k(y_i)} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{m} \left(\left(1 - \sum_{s=1}^{k} \delta_s(y_i)\right) \log(1 - q_k(\boldsymbol{x}_i)) + \delta_k(y_i) \log(q_k(\boldsymbol{x}_i)) \right) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{m} \left(\left(1 - \sum_{s=1}^{k} \delta_s(y_i)\right) \log \left(\frac{1}{1 + \exp(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k)} \right) \right. \\
&\quad \left. + \delta_k(y_i) \log \left(\frac{1}{1 + \exp(-(\boldsymbol{w}_k^\top \boldsymbol{x} + b_k))} \right) \right) \\
&= - \sum_{i=1}^{n} \left(\sum_{k=1}^{m} \left(1 - \sum_{s=1}^{k} \delta_s(y_i)\right) f(-(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)) + \sum_{k=1}^{m} \delta_k(y_i) f(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k) \right),
\end{aligned} \tag{2.5}$$

where

$$f(v) = \log(1 + \exp(-v)). \tag{2.6}$$

The function $f(v)$ is called the logistic loss function. This function is convex because its second derivative always has a positive value. Hence, maximizing the log likelihood function (2.5) is a convex optimization problem.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

From (2.2), (2.3) and (2.5), we obtain a compact formulation of the log likelihood function,

$$\begin{aligned} L(\boldsymbol{b}, \boldsymbol{W}) &= -\sum_{i=1}^{n} \left(\sum_{k=1}^{y_i-1} f(-(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)) + \sum_{k=y_i}^{y_i} f(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k) \right) \\ &= -\sum_{i=1}^{n} \sum_{k=1}^{m} |\psi_{ik}| f(\psi_{ik}(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)), \end{aligned}$$

where

$$\psi_{ik} = \begin{cases} -1 & \text{if } k < y_i, \\ 1 & \text{if } k = y_i, \\ 0 & \text{otherwise,} \end{cases} \qquad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m).$$

3. Mixed Integer Optimization Formulations for Feature Subset Selection

This section presents mixed integer optimization (MIO) formulations for feature subset selection in the sequential logit model.

3.1. Mixed integer nonlinear optimization formulation

Similarly to the previous research [6, 19, 28, 35], we shall employ information criteria, e.g., the Akaike information criterion (AIC, [4]) and Bayesian information criterion (BIC, [32]), as a goodness-of-fit measure for the sequential logit model.

Let $S \subseteq \{1, 2, \ldots, p\}$ be a set of selected features. Accordingly, by setting the coefficients of other candidate features to zero, most information criteria can be expressed as follows:

$$-2 \max\{L(\boldsymbol{b}, \boldsymbol{W}) \mid w_{jk} = 0 \ (j \notin S; k = 1, 2, \ldots, m)\} + Fm(|S| + 1), \tag{3.1}$$

where F is a penalty for the number of selected features. For instance, $F = 2$ and $F = \log(n)$ correspond to the AIC and BIC, respectively.

Let $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)^\top$ be a vector of 0-1 decision variables; $z_j = 1$ if $j \in S$; $z_j = 0$, otherwise. The feature subset selection problem for minimizing the information criterion (3.1) of the sequential logit model can be formulated as a mixed integer nonlinear optimization (MINLO) problem,

$$\underset{\boldsymbol{b}, \boldsymbol{W}, \boldsymbol{z}}{\text{minimize}} \quad 2\sum_{i=1}^{n} \sum_{k=1}^{m} |\psi_{ik}| f(\psi_{ik}(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)) + Fm \left(\sum_{j=1}^{p} z_j + 1 \right) \tag{3.2}$$

$$\text{subject to} \quad z_j = 0 \Rightarrow w_{jk} = 0 \quad (j = 1, 2, \ldots, p; k = 1, 2, \ldots, m), \tag{3.3}$$

$$z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p). \tag{3.4}$$

The logical implications (3.3) can be represented by a special ordered set type one (SOS1) constraint [8, 9]. This constraint implies that not more than one element in the set can have a nonzero value, and it is supported by standard MIO software. Therefore, to incorporate the logical implications (3.3), it is only necessary to impose the SOS1 constraint on $\{1 - z_j, w_{jk}\}$ $(j = 1, 2, \ldots, p; k = 1, 2, \ldots, m)$. Indeed, if $z_j = 0$, then $1 - z_j$ has a nonzero value, and w_{jk} must be zero from the SOS1 constraints.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

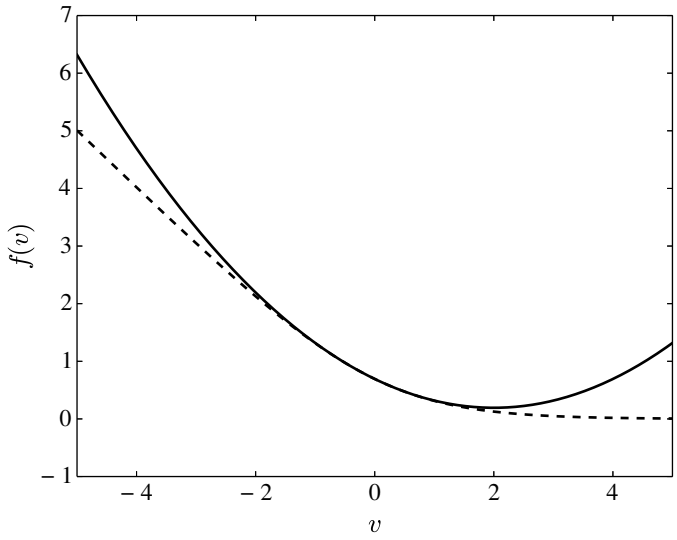

Figure 2: Logistic loss function and its quadratic approximation

3.2. Quadratic approximation

The objective function (3.2) to be minimized is a convex but nonlinear function, which may cause numerical instabilities in the computation. Moreover, most MIO software cannot handle such a nonlinear objective function. In view of these facts, Tanaka and Nakagawa [33] used a quadratic approximation of the logistic loss function.

The second-order Maclaurin series of the logistic loss function (2.6) is written as follows:

$$f(v) \approx \frac{v^2}{8} - \frac{v}{2} + \log 2. \tag{3.5}$$

This quadratic approximation of the logistic loss function reduces the MINLO problem (3.2)–(3.4) to a mixed integer quadratic optimization (MIQO) problem,

$$\begin{aligned} \underset{\boldsymbol{b},\boldsymbol{W},\boldsymbol{z}}{\text{minimize}} \quad & 2\sum_{i=1}^{n}\sum_{k=1}^{m} |\psi_{ik}| \left(\frac{\psi_{ik}^2(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)^2}{8} - \frac{\psi_{ik}(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k)}{2} + \log 2 \right) \\ & + Fm\left(\sum_{j=1}^{p} z_j + 1 \right) \qquad (3.6) \\ \text{subject to} \quad & z_j = 0 \Rightarrow w_{jk} = 0 \quad (j = 1, 2, \ldots, p; k = 1, 2, \ldots, m), \qquad (3.7) \\ & z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p). \qquad (3.8) \end{aligned}$$

Figure 2 shows the graphs of the logistic loss function (2.6) (dashed curve) and its quadratic approximation (3.5) (solid curve). We can see that the approximation error sharply increases with distance from $v = 0$. More importantly, the quadratic approximation function increases on the right side, while the logistic loss function monotonically decreases so that it reduces penalties on correctly classified samples. This means that the quadratic approximation imposes large penalties on correctly classified samples. Consequently, the MIQO problem (3.6)–(3.8) may fail to find a good subset of features.

3.3. Piecewise-linear approximation

In order to approximate the logistic loss function more accurately, we propose the use of a piecewise-linear approximation instead of a quadratic approximation.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

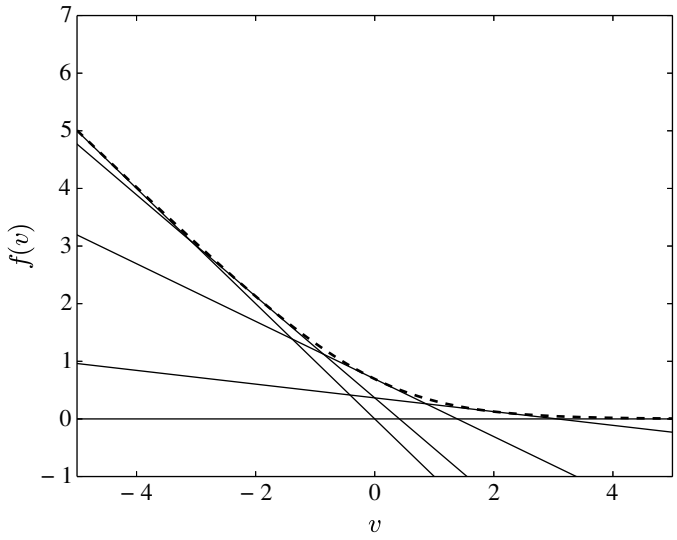

Figure 3: Logistic loss function and its tangent lines

By following Sato et al. [30], we make a piecewise-linear approximation of the logistic loss function. Let $V = \{v_1, v_2, \ldots, v_h\}$ be a set of h discrete points. Since the graph of a convex function lies above its tangent lines, the logistic loss function (2.6) can be approximated by the pointwise maximum of a family of tangent lines, that is,

$$\begin{aligned} f(v) &\approx \max\{f'(v_\ell)(v - v_\ell) + f(v_\ell) \mid \ell = 1, 2, \ldots, h\} \\ &= \min\{t \mid t \geq f'(v_\ell)(v - v_\ell) + f(v_\ell) \quad (\ell = 1, 2, \ldots, h)\}. \end{aligned}$$

Figure 3 shows the graph of the logistic loss function (2.6) (dashed curve) together with the tangent lines (solid lines) at $v_1 = -\infty, v_2 = -1.90, v_3 = 0, v_4 = 1.90,$ and $v_5 = \infty$. Also note that

$$\begin{aligned} f'(v_1)(v - v_1) + f(v_1) &= -v, \\ f'(v_5)(v - v_5) + f(v_5) &= 0. \end{aligned}$$

As shown in Figure 3, the pointwise maximum of the five tangent lines creates a piecewise-linear underestimator of the logistic loss function. It is clear that this approach approximates the logistic loss function more accurately than the quadratic approximation approach does.

By utilizing a piecewise-linear approximation of the logistic loss function, the feature subset selection problem for the sequential logit model can be posed as a mixed integer linear optimization (MILO) problem,

$$\begin{aligned} \underset{\boldsymbol{b},\boldsymbol{T},\boldsymbol{W},\boldsymbol{z}}{\text{minimize}} \quad & 2\sum_{i=1}^{n}\sum_{k=1}^{m} |\psi_{ik}| t_{ik} + Fm\left(\sum_{j=1}^{p} z_j + 1\right) && (3.9) \\ \text{subject to} \quad & t_{ik} \geq f'(v_\ell)(\psi_{ik}(\boldsymbol{w}_k^\top \boldsymbol{x}_i + b_k) - v_\ell) + f(v_\ell) \\ & \qquad (i = 1, 2, \ldots, n; k = 1, 2, \ldots, m; \ell = 1, 2, \ldots, h), && (3.10) \\ & z_j = 0 \Rightarrow w_{jk} = 0 \quad (j = 1, 2, \ldots, p; k = 1, 2, \ldots, m), && (3.11) \\ & z_j \in \{0, 1\} \quad (j = 1, 2, \ldots, p), && (3.12) \end{aligned}$$

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 1: List of instances

abbreviation	n	p	#class	original dataset [7]
`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)

where $\boldsymbol{T} = (t_{ik}; i = 1, 2, \ldots, n, k = 1, 2, \ldots, m)$ is a decision variable for calculating the value of piecewise-linear approximation function.

This MILO problem approaches the original MINLO problem (3.2)–(3.4) by increasing the number of tangent lines at appropriate points. Moreover, this MILO problem, as well as the MIQO problem (3.6)–(3.8), can be solved with standard mathematical optimization software.

4. Computational Results

This section compares the effectiveness of our piecewise-linear approximation approach with that of the quadratic approximation approach employed by Tanaka and Nakagawa [33].

We downloaded eight datasets for ordinal classification from the UCI Machine Learning Repository [7]. Table 1 lists these instances, where n and p are the number of samples and number of candidate features, respectively; and #class is the number of ordinal class labels, i.e., $m + 1$.

For all the instances, each integer and real variable was standardized so that its mean was zero and its standard deviation was one. Each categorical variable was transformed into dummy variable(s). Variables having missing values for samples of over 10% were eliminated. After that, samples including missing values were all 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 computational experiments compared the performances of the following methods:

Quad MIQO formulation (3.6)–(3.8) based on quadratic approximation,

PWL MILO formulation (3.9)–(3.12) based on piecewise-linear approximation using the following set of points for tangent lines, similarly to Sato et al. [30]:

$$V = \{0, \pm 0.44, \pm 0.89, \pm 1.37, \pm 1.90, \pm 2.63, \pm 3.55, \pm 5.16, \pm\infty\} \quad (|V| = 17).$$

All computations were performed on a Linux computer with an Intel Core i7-4820 CPU (3.70 GHz) and 32 GB memory. Gurobi Optimizer 6.0.0 (`http://www.gurobi.com`) was used to solve the MILO and MIQO problems. Here, the logical implications (3.7) and (3.11) were represented by the SOS type 1 function implemented in Gurobi Optimizer.

4.1. Results of AIC/BIC minimization

Tables 2–5 show the computational results of minimizing AIC/BIC in the forward/backward sequential logit models. The columns labeled "AIC" and "BIC" are the values of the corresponding information criteria calculated from the selected set of features. Note that the

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 2: AIC minimization in forward sequential logit model

instance	n	p	#class	method	AIC	objval	$\|S\|$	time (s)
Wine-R	1599	11	6	Quad	3057.5	4204.6	4	0.03
				PWL	**3028.4**	3013.2	10	428.05
Wine-W	4898	11	7	Quad	10859.6	14343.0	8	0.07
				PWL	**10726.7**	10671.2	9	1899.54
Skill	3338	18	7	Quad	9108.8	11289.1	9	5.13
				PWL	**9080.2**	8939.0	15	>10000
Choice	1474	21	3	Quad	2816.2	2850.2	9	7.07
				PWL	**2813.1**	2804.4	12	1632.37
Tnns-W	118	31	7	Quad	331.1	331.1	0	>10000
				PWL	**316.2**	315.4	4	>10000
Tnns-M	113	33	7	Quad	278.5	296.1	2	>10000
				PWL	**278.3**	277.6	4	>10000
Stdnt-M	395	40	18	Quad	1052.7	2251.2	1	>10000
				PWL	**946.2**	941.9	6	>10000
Stdnt-P	649	40	17	Quad	1709.9	3929.7	1	>10000
				PWL	**1653.6**	1645.2	7	>10000

Table 3: AIC minimization in backward sequential logit model

instance	n	p	#class	method	AIC	objval	$\|S\|$	time (s)
Wine-R	1599	11	6	Quad	3062.5	4073.5	5	0.04
				PWL	**3050.1**	3034.6	10	357.76
Wine-W	4898	11	7	Quad	10811.6	14697.4	7	0.05
				PWL	**10786.9**	10734.7	9	1785.95
Skill	3338	18	7	Quad	9024.5	11089.6	8	22.65
				PWL	**8961.2**	8925.9	10	>10000
Choice	1474	21	3	Quad	2829.4	2940.8	9	11.51
				PWL	**2826.1**	2816.7	11	3513.40
Tnns-W	118	31	7	Quad	331.1	447.4	0	253.13
				PWL	**327.0**	307.5	8	>10000
Tnns-M	113	33	7	Quad	307.4	419.8	0	9.53
				PWL	**280.5**	279.3	4	>10000
Stdnt-M	395	40	18	Quad	1065.5	2584.6	2	>10000
				PWL	**1044.1**	1038.1	2	>10000
Stdnt-P	649	40	17	Quad	1640.7	3532.1	2	>10000
				PWL	**1620.7**	1611.6	3	>10000

smaller of the AIC/BIC values between Quad and PWL are bold-faced for each instance. The column labeled "objval" is the value of the objective function, i.e., (3.6) and (3.9). The column labeled "$|S|$" is the number of selected features, and the column labeled "time (s)" is computation time in seconds. The computation for solving the MILO/MIQO problems was terminated if it did not finish by itself after 10000 seconds. In this case, the tables show the result of the best solution obtained within 10000 seconds.

Tables 2 and 3 show the results of AIC minimization in the forward and backward

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 4: BIC minimization in forward sequential logit model

instance	n	p	#class	method	BIC	objval	$\|S\|$	time (s)
`Wine-R`	1599	11	6	Quad	**3191.9**	4339.1	4	0.05
				PWL	**3191.9**	3175.3	4	331.58
`Wine-W`	4898	11	7	Quad	11127.4	14543.6	3	0.05
				PWL	**11077.0**	11019.4	4	6901.04
`Skill`	3338	18	7	Quad	9434.7	11518.1	3	1.41
				PWL	**9333.3**	9291.3	6	>10000
`Choice`	1474	21	3	Quad	**2903.4**	2932.4	5	3.00
				PWL	**2903.4**	2894.5	5	1372.28
`Tnns-W`	118	31	7	Quad	**347.7**	347.7	0	99.67
				PWL	**347.7**	345.9	0	833.61
`Tnns-M`	113	33	7	Quad	**323.7**	323.7	0	96.83
				PWL	**323.7**	321.3	0	514.01
`Stdnt-M`	395	40	18	Quad	1187.9	2386.4	1	220.77
				PWL	**1181.7**	1175.9	2	>10000
`Stdnt-P`	649	40	17	Quad	**1853.2**	4073.0	1	725.73
				PWL	**1853.2**	1835.3	1	>10000

Table 5: BIC minimization in backward sequential logit model

instance	n	p	#class	method	BIC	objval	$\|S\|$	time (s)
`Wine-R`	1599	11	6	Quad	3257.3	4197.6	2	0.02
				PWL	**3206.5**	3190.6	4	296.37
`Wine-W`	4898	11	7	Quad	11151.9	14907.6	3	0.03
				PWL	**11143.2**	11083.8	4	8233.98
`Skill`	3338	18	7	Quad	9425.3	11315.8	4	3.47
				PWL	**9275.4**	9241.3	6	>10000
`Choice`	1474	21	3	Quad	2917.3	3009.5	4	1.86
				PWL	**2917.0**	2907.5	5	1601.25
`Tnns-W`	118	31	7	Quad	**347.7**	464.0	0	0.10
				PWL	**347.7**	344.1	0	3205.70
`Tnns-M`	113	33	7	Quad	**323.7**	436.1	0	0.09
				PWL	**323.7**	319.9	0	9767.70
`Stdnt-M`	395	40	18	Quad	**1207.0**	2740.4	1	421.56
				PWL	**1207.0**	1199.8	1	>10000
`Stdnt-P`	649	40	17	Quad	**1822.4**	3717.1	1	>10000
				PWL	**1822.4**	1809.3	1	>10000

sequential logit models. These tables reveal that AIC values of PWL were smaller than those of Quad for all the instances. PWL provided better-quality solutions than Quad did, while the computation time of PWL was longer than that of Quad because the problem size of PWL is dependent on the number of samples (see the constraints (3.10)). However, the computation time of 10000 seconds is often allowable in reality when the purpose of feature subset selection is to analyze stored data. In addition, we should notice that the number of features selected by PWL sometimes differed greatly from that selected by Quad. For

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 6: Candidate features in `Choice` instance

abbreviation	attribute information	
`WifeAge`	wife's age	numerical
`WifeEdc1`	wife's education	1: low
`WifeEdc2`		2
`WifeEdc3`		3
`WifeEdc4`		4: high
`HsbndEdc1`	husband's education	1: low
`HsbndEdc2`		2
`HsbndEdc3`		3
`HsbndEdc4`		4: high
`NmbrChld`	number of children ever born	numerical
`WifeRlgn`	wife's religion	0: non-Islam, 1: Islam
`WifeWrkng`	wife's now working?	0: yes, 1: no
`HsbndOccp1`	husband's occupation	1: low
`HsbndOccp2`		2
`HsbndOccp3`		3
`HsbndOccp4`		4: high
`StndLvng1`	standard-of-living index	1: low
`StndLvng2`		2
`StndLvng3`		3
`StndLvng4`		4: high
`MdExpsr`	media exposure	0: good, 1: not good

instance, Quad and PWL respectively selected 4 and 10 features for `Wine-R` in Table 2.

We can see from Tables 2 and 3 that PWL approximated the logistic loss function very accurately, whereas Quad caused a major gap between AIC and objval. Additionally, since the objective function of PWL is an underestimator relative to AIC, its optimal value serves as a lower bound of the smallest AIC. In the case of `Wine-R` in Table 2, AIC and objval of Quad were 3057.5 and 4204.6, whereas those of PWL were 3028.4 and 3013.2. It follows that PWL proved that the smallest AIC value necessarily fell between 3013.2 and 3028.4. This optimality guarantee is the most notable characteristic of our piecewise-linear approximation approach, and it cannot be shared by the quadratic approximation approach.

Tables 4 and 5 show the results of BIC minimization in the forward and backward sequential logit models. We should recall that the penalty, F, for the number of selected features in BIC (i.e., $F = \log(n)$) is larger than that in AIC (i.e., $F = 2$). Hence, BIC minimization selected a small number of features, and thus, Quad and PWL often yielded the same subset of features for each instance in Tables 4 and 5. Meanwhile, when these two methods provided different subsets of features, PWL always found the better one.

4.2. Analysis of selected features

This subsection analyzes the selected features in the `Choice` instance because the results of that instance are relatively easy to interpret. The samples in this instance are married and non-pregnant women, and its objective is to predict the current contraceptive method choice, i.e., no-use ($y_i = 1$), short-term methods ($y_i = 2$) or long-term methods ($y_i = 3$), of a woman. Table 6 shows candidate features in the `Choice` instance, where `WifeAge` and `NmbrChld` are integer variables, and all the other features are dummy variables.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 7: Estimated coefficients by AIC minimization in forward sequential logit model

	Quad		PWL	
	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$	$\boldsymbol{w}_1$	$\boldsymbol{w}_2$
(intercept)	−1.78	3.16	−2.22	2.02
`WifeAge`	0.08	−0.07	0.09	−0.07
`WifeEdc1`	0.46	0.26	0.44	0.23
`WifeEdc2`	—		—	
`WifeEdc3`	0.70	0.07	0.57	−0.01
`WifeEdc4`	—		0.74	1.36
`HsbndEdc1`	—		0.43	0.51
`HsbndEdc2`	−0.49	−0.64	—	
`HsbndEdc3`	−1.23	−0.99	−0.70	−0.34
`HsbndEdc4`	0.48	−2.63	0.35	−2.82
`NmbrChld`	−0.34	0.02	−0.35	0.02
`WifeRlgn`	—		—	
`WifeWrkng`	—		—	
`HsbndOccp1`	—		—	
`HsbndOccp2`	−0.08	−0.56	—	
`HsbndOccp3`	—		0.22	0.49
`HsbndOccp4`	—		−0.04	0.65
`StndLvng1`	—		—	
`StndLvng2`	0.57	0.45	0.48	0.47
`StndLvng3`	—		—	
`StndLvng4`	—		—	
`MdExpsr`	—		−0.30	0.03

Table 7 shows the estimated coefficients of features selected by AIC minimization in the forward sequential logit models. Note that these models first removed samples of "no-use" based on the coefficient vector $\boldsymbol{w}_1$ and then chose those of "short-term methods" from the remainder on the basis of $\boldsymbol{w}_2$. Although Quad and PWL selected different subset of features, they obtained similar estimates of coefficients for the features selected in common by them. From the results, we can describe the following observations:

- contraceptive methods are less likely to be used as a wife gets older.
- long-term methods are likely to be used when husband's education is high.
- contraceptive methods are more likely to be used as the number of children grows.
- short-term methods are likely to be used when husband's occupational status is high.
- no-use and short-term methods are likely to be selected when the standard of living is slightly lower.

These are largely plausible explanations for contraceptive method choice.

4.3. Evaluation of out-of-sample predictive performance

This subsection evaluates out-of-sample predictive performance of our method through three-fold cross-validation. A predicted class label was defined as $\hat{y}_i \in \arg\max_k(\prod_{s=1}^{k-1}(1 - q_s(\boldsymbol{x}_i)))q_k(\boldsymbol{x}_i)$ for each sample $i = 1, 2, \ldots, n$, and its root mean squared error (RMSE, i.e., $\sqrt{(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)/n}$) and accuracy (i.e., percentage of correct answers) were used as measures of predictive performance.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Table 8: Three-fold cross-validation of AIC minimization

					RMSE		Accuracy (%)	
instance	n	p	#class	model	Quad	PWL	Quad	PWL
`Wine-R`	1599	11	6	forward	0.729	**0.719**	58.29	**58.85**
				backward	0.743	**0.735**	**57.54**	**57.54**
`Wine-W`	4898	11	7	forward	0.812	**0.811**	53.08	**53.61**
				backward	0.821	**0.819**	**52.88**	52.86
`Skill`	3338	18	7	forward	**1.068**	1.077	**39.78**	39.36
				backward	1.077	**1.068**	39.60	**40.20**
`Choice`	1474	21	3	forward	0.900	**0.891**	49.83	**50.78**
				backward	0.872	**0.869**	51.66	**53.02**

Table 8 shows the results of three-fold cross-validation in AIC minimization, where the better of the RMSE/accuracy values between Quad and PWL are bold-faced. We employed the four instances (i.e., `Wine-R`, `Wine-W`, `Skill` and `Choice`) because they contain a sufficient number of samples. It is clear from the results that PWL obtained better RMSE/accuracy values more frequently than Quad did.

5. Conclusions

This paper dealt with the feature subset selection problem for a sequential logit model. We formulated it as a mixed integer linear optimization (MILO) problem by applying a piecewise-linear approximation to the logistic loss functions. The computational results confirmed that our formulation has a clear advantage over the mixed integer quadratic optimization (MIQO) formulation proposed in the previous study [33].

In contrast to the MIQO formulation, the approximation accuracy of the logistic loss function can be controlled by the number of tangent lines in our MILO formulation. Furthermore, after the MILO problem is solved, it provides an optimality guarantee of the selected features on the basis of information criteria. To the best of our knowledge, this paper is the first to compute a subset of features with an optimality guarantee for a sequential logit model.

A future direction of study will be to extend our piecewise-linear approximation approach to other logit models. However, this will be a difficult task because it is imperative to approximate a multivariate objective function. Another direction of future research is to analyze actual data by means of our feature subset selection method. For instance, Sato et al. [31] investigated consumers' store choice behavior by applying feature subset selection based on mixed integer optimization. Since proper feature subset selection is essential for data analysis, our approach has a clear advantage over heuristic algorithms.

Acknowledgments

This work was partially supported by Grants-in-Aid for Scientific Research by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] A. Afonso, P. Gomes, and P. Rother: Ordered response models for sovereign debt ratings. *Applied Economics Letters,* **16** (2009), 769–773.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

[2] A. Agresti: *Categorical Data Analysis* (Wiley, New York, 1990).

[3] A. Agresti: *Analysis of Ordinal Categorical Data, Second Edition* (John Wiley & Sons, New York, 2010).

[4] H. Akaike: A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* **19** (1974), 716–723.

[5] T. Amemiya: Qualitative response models: A survey. *Journal of Economic Literature,* **19** (1981), 1483–1536.

[6] K.J. Archer and A.A.A. Williams: L_1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. *Statistics in Medicine,* **31** (2012), 1464–1474.

[7] K. Bache and M. Lichman: UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science (2013).

[8] E.M.L. Beale: Two transportation problems. In G. Kreweras and G. Morlat (eds.): *Proceedings of the Third International Conference on Operational Research* (Dunod, Paris and English Universities Press, London, 1963), 780–788.

[9] E.M.L. Beale and J.A. Tomlin: Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. In J. Lawrence (ed.): *Proceedings of the Fifth International Conference on Operational Research* (Tavistock Publications, London, 1970), 447–454.

[10] D. Bertsimas and A. King: OR forum—An algorithmic approach to linear regression. *Operations Research,* **64** (2016), 2–16.

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

[12] A.L. Blum and P. Langley: Selection of relevant features and examples in machine learning. *Artificial Intelligence,* **97** (1997), 245–271.

[13] L.H. Ederington: Classification models and bond ratings. *Financial Review,* **20** (1985), 237–262.

[14] M.A. Efroymson: Multiple regression analysis. In A. Ralston and H.S. Wilf (eds.): *Mathematical Methods for Digital Computers* (Wiley, New York, 1960), 191–203.

[15] S.E. Fienberg: *The Analysis of Cross-Classified Categorical Data* (The MIT Press, Cambridge, 1980).

[16] H. Fu and C. Wilmot: Sequential logit dynamic travel demand model for hurricane evacuation. *Transportation Research Record: Journal of the Transportation Research Board,* **1882** (2004), 19–26.

[17] I. Guyon and A. Elisseeff: An introduction to variable and feature selection. *The Journal of Machine Learning Research,* **3** (2003), 1157–1182.

[18] L.M. Kahn and K. Morimune: Unions and employment stability: A sequential logit approach. *International Economic Review,* **20** (1979), 217–235.

[19] L.N. Kazembe: A semiparametric sequential ordinal model with applications to analyse first birth intervals. *Austrian Journal of Statistics,* **38** (2009), 83–99.

[20] H.T. Kiiveri: A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics,* **9** (2008), 195.

[21] R. Kohavi and G.H. John: Wrappers for feature subset selection. *Artificial Intelligence,* **97** (1997), 273–324.

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

[22] H. Konno and Y. Takaya: Multi-step methods for choosing the best set of variables in regression analysis. *Computational Optimization and Applications,* **46** (2010), 417–426.

[23] H. Konno and R. Yamamoto: Choosing the best set of variables in regression analysis using integer programming. *Journal of Global Optimization,* **44** (2009), 273–282.

[24] I. Liu and A. Agresti: The analysis of ordered categorical data: An overview and a survey of recent developments. *Test,* **14** (2005), 1–73.

[25] A. Miller: *Subset Selection in Regression, Second Edition* (CRC Press, Boca Raton, 2002).

[26] R. Miyashiro and Y. Takano: Subset selection by Mallows' C_p: A mixed integer programming approach. *Expert Systems with Applications,* **42** (2015), 325–331.

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

[28] D. Nagakura and M. Kobayashi: Testing the sequential logit model against the nested logit model. *Japanese Economic Review,* **60** (2009), 345–361.

[29] W.P. Poon, M. Firth, and H.G. Fung: A multivariate analysis of the determinants of Moody's bank financial strength ratings. *Journal of International Financial Markets, Institutions and Money,* **9** (1999), 267–283.

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

[31] T. Sato, Y. Takano, and T. Nakahara: Using mixed integer optimisation to select variables for a store choice model. *International Journal of Knowledge Engineering and Soft Data Paradigms,* **5** (2016), 123–134.

[32] G. Schwarz: Estimating the dimension of a model. *The Annals of Statistics,* **6** (1978), 461–464.

[33] 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** (2014), 92–111.

[34] G. Tutz: Sequential models in categorical regression. *Computational Statistics & Data Analysis,* **11** (1991), 275–295.

[35] G. Tutz and A. Groll: Binary and ordinal random effects models including variable selection. Technical Report 97, Department of Statistics, University of Munich (2010).

[36] W.C. Weiler: A sequential logit model of the access effects of higher education institutions. *Economics of Education Review,* **5** (1986), 49–55.

[37] S.C. Yusta: Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters,* **30** (2009), 525–534.

Toshiki Sato
Doctoral Program in Social Systems and Management
Graduate School of Systems and Information Engineering
University of Tsukuba
1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8573, Japan
E-mail: tsato@sk.tsukuba.ac.jp

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.