BIAS-CORRECTED AIC FOR SELECTING VARIABLES IN MULTINOMIAL LOGISTIC REGRESSION MODELS

(Last Modified: January 23, 2012)

Hirokazu Yanagihara*[1], Ken-ichi Kamo**, Shinpei Imori* and Kenichi Satoh***

*Department of Mathematics, Graduate School of Science, Hiroshima University
1-3-1 Kagamiyama, Higashi-Hiroshima 739-8626, Japan

**Department of Liberal Arts and Sciences, Sapporo Medical University
South 1, West 17, Chuo-ku, Sapporo 060-8543, Japan

***Department of Environmetrics and Biometrics
Research Institute for Radiation Biology and Medicine, Hiroshima University

1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan

Abstract

In this paper, we consider the bias correction of Akaike's information criterion (AIC) for selecting variables in multinomial logistic regression models. For simplifying a formula of the bias-corrected AIC, we calculate the bias of the AIC to a risk function through the expectations of partial derivatives of the minus log-likelihood function. As a result, we can express the bias correction term of the bias-corrected AIC with only three matrices consisting of the second, third, and fourth derivatives of the minus log-likelihood function. By conducting numerical studies, we verify that the proposed bias-corrected AIC performs better than the crude AIC.

AMS 2010 *subject classifications*: Primary 62H12; Secondary 62J12.
Key words: AIC, Bias correction, Multinomial logistic model, MLE, Partial differential operator, Variable selection.

1. INTRODUCTION

A multinomial logistic regression model is a regression model that generalizes a logistic regression by allowing more than two discrete outcomes. When categories are unordered, the multinomial logistic model is one strategy often used. The multinomial logistic regression model has been introduced in many textbooks for applied statistical analysis (see e.g., Hosmer & Lemeshow, 2000, Chapter 8.1), and even now it is widely used in biometrics, econometrics, psychometrics, sociometrics, and many other fields of applications for the prediction of probabilities of different possible outcomes of categorically distributed response variables by a set of explanatory variables (e.g., Briz & Ward, 2009; Choi

[1]Corresponding author, *E-mail address*: yanagi@math.sci.hiroshima-u.ac.jp

& Alig, 2011; dell'Olio, Ibeas & Cecin, 2011). In addition, the multinomial logistic regression model can be easily fitted to real data by using the "vglm" function in "R" (R Development Core Team, 2011). Since we would like to specify the factors affecting the probabilities of response variables in the regression analysis, searching for the optimal subset of explanatory variables is important.

Akaike's information criterion (AIC) proposed by Akaike (1973, 1974) is widely used for selecting the best model among the candidate models (for details of statistical model selection, see e.g., Konishi, 1999; Burnham & Anderson, 2002; Konishi & Kitagawa, 2008). The model having the smallest AIC among the candidate models is regarded as the best model. In the multinomial logistic regression model, the subset of explanatory variables in the best model is the best subset. However, the AIC may perform poorly; that is, a model having too many parameters tends to be chosen as the best model when the sample size is small or the number of unknown parameters is large. Such a problem is often resolved by using a bias-corrected AIC (see e.g., Burnham & Anderson, 2002, Chapter 2.4). The AIC is an estimator of the risk function consisting of predictive Kullback-Leibler (K-L) information (Kullback & Leibler, 1951), which measures the discrepancy between the true model and the candidate model. The order of the bias of the AIC is $O(n^{-1})$ when the candidate model includes the true model, where n is the sample size. Although the AIC is an asymptotic unbiased estimator of the risk function, it has a nonnegligible bias to the risk function when the sample size is small or the number of unknown parameters is large. A bias-corrected AIC called CAIC in this paper improves the bias of AIC to $O(n^{-2})$ under the assumption that the candidate model includes the true model.

The CAIC in the logistic regression models was obtained by Yanagihara, Sekiguchi and Fujikoshi (2003). But the CAIC in multinomial regression models has not been derived yet, although the multinomial logistic regression model is widely used in many application fields. The CAIC can be obtained by removing the bias of the AIC to the risk function from the AIC with the use of a consistent estimator of the bias. The bias of the AIC to the risk function is then evaluated by moments of the maximum likelihood estimator (MLE) of unknown parameters. Since such moments are represented by the moments of response variables, calculating the moments of response variables is essential for evaluating the bias of the AIC in the ordinary bias correction method, which is used in Yanagihara, Sekiguchi and Fujikoshi (2003) and Kamo, Yanagihara and Satoh (2012), etc. However, in the case of multiple response variables, calculations and expressions of the moments of the MLE mediated by the moments of response variables become complicated. Hence, without directly calculating the moments of response variables, we derive the moments of the MLE by using expectations of the partial derivatives of the minus log-likelihood function. This different approach from the ordinary bias correction method leads to a simple expression of the bias correction term of the CAIC. In fact, the bias correction term of our CAIC is represented by only three matrices consisting of the second, third, and fourth derivatives of the minus log-likelihood function.

The present paper is organized as follows. In Section 2, we give a stochastic expansion of the MLE. In Section 3, the CAIC in multinomial logistic regression models is proposed. In Section 4, we verify that the proposed CAIC has better performance than the AIC by conducting numerical experiments. In Section 5, we conclude our discussion. Technical details are provided in the Appendix.

2. STOCHASTIC EXPANSION OF MLE

Suppose that the data consists of a sequence $\{\boldsymbol{y}_i, \boldsymbol{x}_i\}$, where $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are r-dimensional independent unordered discrete random vectors, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are k-dimensional vectors of known constants. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{kr})'$ be a kr-dimensional unknown regression coefficient vector that is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_r')'$, where $\boldsymbol{\beta}_j$ is a k-dimensional vector denoted by $\boldsymbol{\beta}_j = (\beta_{(j-1)k+1}, \ldots, \beta_{jk})'$. In the multinomial logistic regression model, we assume that $(y_{i0}, \boldsymbol{y}_i')' = (y_{i0}, y_{i1}, \ldots, y_{ir})'$ is distributed according to the multinomial distribution with the number of events n_i ($n_i = \sum_{j=0}^r y_{ij}$, $n = \sum_{i=1}^m n_i$) and the cell probability vector $(p_{i0}(\boldsymbol{\beta}), \boldsymbol{p}_i(\boldsymbol{\beta})')'$, given by

$$
\begin{aligned}
p_{i0}(\boldsymbol{\beta}) &= \frac{1}{1 + \sum_{j=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}, \\
\boldsymbol{p}_i(\boldsymbol{\beta}) = (p_{i1}(\boldsymbol{\beta}), \ldots, p_{ir}(\boldsymbol{\beta}))' &= \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1 + \sum_{j=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}, \ldots, \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_r)}{1 + \sum_{j=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)} \right)'.
\end{aligned}
\tag{1}
$$

The MLE of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function. By omitting the constant term, the log-likelihood function of the multinomial logistic regression model in (1) is expressed as

$$
\ell(\boldsymbol{\beta}) = \sum_{i=1}^m \left\{ (\boldsymbol{y}_i \otimes \boldsymbol{x}_i)'\boldsymbol{\beta} - n_i \log \left(1 + \sum_{j=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j) \right) \right\}. \tag{2}
$$

Hence, the MLE of $\boldsymbol{\beta}$ is given by

$$
\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}).
$$

To evaluate a bias of the AIC to the risk function, a stochastic expansion of $\hat{\boldsymbol{\beta}}$ is needed. The purpose of this section is to obtain the stochastic expansion $\hat{\boldsymbol{\beta}}$ up to the order $n^{-3/2}$. Two cases serve as a framework for asymptotic approximations:

Case (i): n_j's are fixed, and $m \to \infty$,

Case (ii): m is fixed, $n_j \to \infty$ and $\rho_j^{-1} = n/n_j = O(1)$ for each j.

Although we only consider Case (i) in this paper, our formula can also be applied to Case (ii).

Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are members of an admissible set $\mathcal{F}$, i.e., $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{F}$. To expand the MLE, we consider the following regularity assumptions (see e.g., Fahrmeir & Kaufmann, 1985):

C.1 : $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B}$ is a convex and open set in $\mathbb{R}^k$,

C.2 : $(\boldsymbol{I}_r \otimes \boldsymbol{x}_i)'\boldsymbol{\beta} \in \Theta^0$, $i = 1, 2, \dots$, for all $\boldsymbol{\beta} \in \mathcal{B}$, where Θ^0 is the interior of the convex natural parameter space $\Theta \subset \mathbb{R}^r$,

C.3 : $^\exists m_0$ *s.t.* $\boldsymbol{X}'\boldsymbol{X}$ has the full rank for $m \geq m_0$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_m)'$.

Condition C.1 guarantees the uniqueness of the MLE if it exists. Condition C.2 is necessary to obtain the multinomial logistic regression model for all $\boldsymbol{\beta}$. Condition C.3 ensures that $\sum_{i=1}^m n_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i\boldsymbol{x}_i'$ is positive definite for all $\boldsymbol{\beta} \in \mathcal{B}$, $m \geq m_0$, where

$$\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \mathrm{diag}\left(\boldsymbol{p}_i(\boldsymbol{\beta})\right) - \boldsymbol{p}_i(\boldsymbol{\beta})\boldsymbol{p}_i(\boldsymbol{\beta})'. \tag{3}$$

Moreover, we prepare the following additional conditions to assure weak consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}$, which can be derived by slightly modifying the results in Fahrmeir and Kaufmann (1985):

C.4 : sequence $\{\boldsymbol{x}_i\}$ lies in $\mathcal{F}$ with $(\boldsymbol{I}_r \otimes \boldsymbol{x})'\boldsymbol{\beta} \in \Theta^0$, $\boldsymbol{\beta} \in \mathcal{B}$,

C.5 : $\liminf_{m\to\infty} \lambda(\sum_{i=1}^m n_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i\boldsymbol{x}_i'/n) > 0$, where $\lambda(\boldsymbol{A})$ indicates the smallest eigenvalue of symmetric matrix $\boldsymbol{A}$.

According to Corollary 1 in Fahrmeir and Kaufmann (1985), $\hat{\boldsymbol{\beta}}$ has weak consistency and asymptotic normality under these conditions. Furthermore, from C.5, $\sum_{i=1}^m n_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i\boldsymbol{x}_i' = O(n)$ is satisfied. Under the assumption that all conditions are satisfied, $\hat{\boldsymbol{\beta}}$ can be formally expanded as follows:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{b}_1 + \frac{1}{n}\boldsymbol{b}_2 + \frac{1}{n\sqrt{n}}\boldsymbol{b}_3 + O_p(n^{-2}), \tag{4}$$

where $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are kr-dimensional random vectors. The purpose of this section is achieved by specifying $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$.

Since the log-likelihood function $\ell(\boldsymbol{\beta})$ is a maximum at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the first derivative of $\ell(\boldsymbol{\beta})$ becomes $\boldsymbol{0}_{kr}$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, i.e.,

$$\left.\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^m \left\{(\boldsymbol{y}_i \otimes \boldsymbol{x}_i) - n_i(\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}) \otimes \boldsymbol{x}_i)\right\} = \boldsymbol{0}_{kr}, \tag{5}$$

where $\boldsymbol{0}_{kr}$ is a kr-dimensional vector of zeros. To expand equation (5), we prepare the following three matrices consisting of the second, third, and fourth derivatives of $-\ell(\boldsymbol{\beta})/n$:

$$\boldsymbol{G}_2(\boldsymbol{\beta}) = -\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}, \qquad \boldsymbol{G}_3(\boldsymbol{\beta}) = -\frac{1}{n}\left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta}),$$

$$\boldsymbol{G}_4(\boldsymbol{\beta}) = -\frac{1}{n}\left(\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta}).$$

The result of the first derivative of $-\ell(\boldsymbol{\beta})$ in (5) implies the following explicit forms of $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, and $\boldsymbol{G}_4(\boldsymbol{\beta})$ (details of the derivations are given in Appendix A.1):

$$\boldsymbol{G}_2(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \frac{\partial \boldsymbol{p}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \left(\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \right), \tag{6}$$

$$\boldsymbol{G}_3(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \frac{\partial}{\partial \boldsymbol{\beta}'} \right) \boldsymbol{p}_i(\boldsymbol{\beta}) \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \left\{ \boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \right\}, \tag{7}$$

$$\boldsymbol{G}_4(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \left(\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \otimes \frac{\partial}{\partial \boldsymbol{\beta}'} \right) \boldsymbol{p}_i(\boldsymbol{\beta}) \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \left\{ \boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \right\}, \tag{8}$$

where $\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta})$ and $\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta})$ are $kr \times (kr)^2$ and $(kr)^2 \times (kr)^2$ matrices, respectively, which are defined by

$$\begin{aligned}
\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) &= \sum_{a=1}^{r} p_{ia}(\boldsymbol{\beta}) \boldsymbol{e}_a' \otimes \boldsymbol{x}_i' \otimes (\boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta}))(\boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta}))' - \boldsymbol{p}_i(\boldsymbol{\beta})' \otimes \boldsymbol{x}_i' \otimes \boldsymbol{\Sigma}_i(\boldsymbol{\beta}), \\
\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) &= \sum_{a=1}^{r} p_{ia}(\boldsymbol{\beta})(\boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta}))(\boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta}))' \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes (\boldsymbol{e}_a \boldsymbol{e}_a' - \boldsymbol{e}_a \boldsymbol{p}_i(\boldsymbol{\beta})' - \boldsymbol{p}_i(\boldsymbol{\beta}) \boldsymbol{e}_a') \\
&\quad - \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes (\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) - \boldsymbol{p}_i(\boldsymbol{\beta}) \boldsymbol{p}_i(\boldsymbol{\beta})') \\
&\quad - \sum_{a,b}^{r} p_{ia} p_{ib}(\boldsymbol{\beta})(\boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta}))(\boldsymbol{e}_b - \boldsymbol{p}_i(\boldsymbol{\beta}))' \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes (\boldsymbol{e}_a \boldsymbol{e}_b' + \boldsymbol{e}_b \boldsymbol{e}_a').
\end{aligned} \tag{9}$$

Here, $\boldsymbol{e}_j$ is an r-dimensional jth coordinate unit vector whose jth element is 1 and others are 0, and the notation $\sum_{a_1,\ldots,a_j}^{r}$ means $\sum_{a_1=1}^{r} \cdots \sum_{a_j=1}^{r}$.

Applying a Taylor expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ to equation (5) yields

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{m} \{ (\boldsymbol{y}_i - n_i \boldsymbol{p}_i(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i \} \\
&\quad = \boldsymbol{G}_2(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{2} \boldsymbol{G}_3(\boldsymbol{\beta}) \left\{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right\} \\
&\qquad + \frac{1}{6} \left\{ \boldsymbol{I}_{kr} \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \right\} \boldsymbol{G}_4(\boldsymbol{\beta}) \left\{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right\} + O_p(n^{-2}).
\end{aligned} \tag{10}$$

Notice that the order of the left-hand side of equation (10) is $O_p(n^{-1/2})$. By comparing the $O_p(n^{-1/2})$, $O_p(n^{-1})$, and $O_p(n^{-3/2})$ terms after substituting (4) into (10), $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ in (4) are specified as

$$\begin{aligned}
\boldsymbol{b}_1 &= \frac{1}{\sqrt{n}} \boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \sum_{i=1}^{m} \{ (\boldsymbol{y}_i - n_i \boldsymbol{p}_i(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i \}, \\
\boldsymbol{b}_2 &= -\frac{1}{2} \boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1), \\
\boldsymbol{b}_3 &= -\frac{1}{2} \boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \left\{ \boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) + \frac{1}{3} (\boldsymbol{I}_{kr} \otimes \boldsymbol{b}_1') \boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \right\}.
\end{aligned} \tag{11}$$

We use the stochastic expansion of $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ to evaluate the bias of the AIC to the risk function.

3. MAIN RESULT

Let $\mathcal{L}(\boldsymbol{\beta})$ be a loss function defined by

$$\mathcal{L}(\boldsymbol{\beta}) = E[-2\ell(\boldsymbol{\beta})] = -2\sum_{i=1}^{m} n_i \left\{ (\boldsymbol{p}_i^* \otimes \boldsymbol{x}_i)'\boldsymbol{\beta} - \log\left(1 + \sum_{j=1}^{r} \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\right)\right\}, \quad (12)$$

where $\boldsymbol{p}_i^*$ is the cell probability vector of the true model. Then, the risk function consisting of the predictive K-L information is given by

$$R = E[\mathcal{L}(\hat{\boldsymbol{\beta}})]. \quad (13)$$

In this section, we propose a CAIC that improves the bias of the AIC to $O(n^{-2})$ under the assumption that the candidate model includes the true model. Notice that the crude AIC is defined by

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2kr. \quad (14)$$

Thus, it is sufficient to derive the bias of $-2\ell(\hat{\boldsymbol{\beta}})$ to R for evaluating the bias of the AIC. Also notice that $\boldsymbol{p}_i^* = \boldsymbol{p}(\boldsymbol{x}_i)$ holds when the candidate model includes the true model. Then, the bias of $-2\ell(\hat{\boldsymbol{\beta}})$ to R under the assumption that the candidate model includes the true model is expanded as

$$\begin{aligned} B &= R - E[-2\ell(\hat{\boldsymbol{\beta}})] \\ &= 2\sum_{i=1}^{m} E\left[\{(\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i\}' \hat{\boldsymbol{\beta}}\right] \\ &= 2\sqrt{n}E\left[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\hat{\boldsymbol{\beta}}\right] \\ &= 2\left\{\sqrt{n}E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{\beta}] + E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1] + \frac{1}{\sqrt{n}}E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2] + \frac{1}{n}E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_3]\right\} + O(n^{-2}), \end{aligned} \quad (15)$$

where matrices $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are given by (6), (7), and (8), respectively, and kr-dimensional random vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are given by (11). In many cases of practical interest, a moment of statistic can be expanded as a power series in n^{-1} (see e.g., Hall, 1992, p. 46). Hence, the order of the remainder term of (15) is shown by $O(n^{-2})$, not $O(n^{-3/2})$. Indeed, an $n^{-3/2}$ term of the stochastic expansion of $\sum_{i=1}^{m} \{(\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})) \otimes \boldsymbol{x}_i\}' \hat{\boldsymbol{\beta}}$ in the bias can be expressed as a fifth-order polynomial of elements of $\boldsymbol{b}_1$. Since $\boldsymbol{b}_1$ has an asymptotic normality, the expectation of an odd-order polynomial of elements of $\boldsymbol{b}_1$ becomes $O(n^{-1/2})$. Given this fact, the order of the remainder term of the expansion in (15) is $O(n^{-2})$.

From elementary linear algebra and the definition of $\boldsymbol{b}_2$ in (11), $\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})\boldsymbol{b}_2$ in (15) is expressed by the function of $\boldsymbol{b}_1$ as

$$\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2 = -\frac{1}{2}\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) = -\frac{1}{2}\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}. \quad (16)$$

Since the derivative is invariant to changes in the order of differentiation, we have

$$\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_2) = \boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_2 \otimes \boldsymbol{b}_1) = \boldsymbol{b}_2'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) = (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{b}_2.$$

It follows from the above equations that

$$\begin{aligned}\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) &= 2(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{b}_2 \\ &= -(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \\ &= -\mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\right\}.\end{aligned}$$

Thus, from the above result and the definition of $\boldsymbol{b}_3$ in (11), $\boldsymbol{b}_1'\boldsymbol{G}_4(\boldsymbol{\beta})\boldsymbol{b}_3$ in (15) is expressed by the function of $\boldsymbol{b}_1$ as

$$\begin{aligned}\boldsymbol{b}_1'\boldsymbol{G}_2\boldsymbol{b}_3 &= -\frac{1}{2}\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) - \frac{1}{6}\boldsymbol{b}_1'(\boldsymbol{I}_{kr} \otimes \boldsymbol{b}_1')\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \\ &= \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\right\} - \frac{1}{6}\mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}. \end{aligned} \tag{17}$$

Hence, equations (16) and (17) indicate that the expansion of B in (15) can be calculated until the fourth moment of $\boldsymbol{b}_1$.

Since $\boldsymbol{b}_1$ consists of a centralized $\boldsymbol{y}_i$, we can directly calculate the expectations in (15) by centralized moments of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$. Then, all combinations of multivariate moments of $\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})$ are needed until the fourth-order. However, it is troublesome to calculate the third- and fourth-order multivariate moments of $\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})$, because we have to consider all combinations of the multivariate moments. For simplicity, the relations between the moments of $\boldsymbol{b}_1$ and the expectations of the derivatives of $-\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are used instead of calculating the multivariate moments of $\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})$. It is easy to obtain $E[\boldsymbol{b}_1] = \boldsymbol{0}_{kr}$ because $E[\boldsymbol{y}_i] = n_i\boldsymbol{p}_i(\boldsymbol{\beta})$. From the result of the first derivative of $\ell(\boldsymbol{\beta})$ in (5) and the definition of $\boldsymbol{b}_1$ in (11), we can see that

$$-\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}) = -\sqrt{n}\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1.$$

Notice that $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are constant matrices and $-E[\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}] = \boldsymbol{0}_{kr}$. By applying general formulas of expectations (A.7) in Appendix A.2 to the case of the multinomial logistic regression model, the following equations are obtained:

$$\begin{aligned}&n\boldsymbol{G}_2(\boldsymbol{\beta}) = n\boldsymbol{G}_2(\boldsymbol{\beta})E[\boldsymbol{b}_1\boldsymbol{b}_1']\boldsymbol{G}_2(\boldsymbol{\beta}), \\ &n\boldsymbol{G}_3(\boldsymbol{\beta}) = n\sqrt{n}\boldsymbol{G}_2(\boldsymbol{\beta})E[\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'](\boldsymbol{G}_2(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2(\boldsymbol{\beta})), \\ &n\boldsymbol{G}_4(\boldsymbol{\beta}) = n^2(\boldsymbol{G}_2(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2(\boldsymbol{\beta}))E[\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'](\boldsymbol{G}_2(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2(\boldsymbol{\beta})) \\ &\qquad\qquad - n^2\left\{(\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})(\boldsymbol{G}_2(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2(\boldsymbol{\beta})) + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta}))\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta}))'\right\},\end{aligned}$$

where $\mathrm{vec}(\boldsymbol{A})$ is an operator to transform a matrix to a vector by stacking the first to the last column of $\boldsymbol{A}$, i.e., $\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1', \ldots, \boldsymbol{a}_m')'$ when $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m)$ (see e.g., Harville, 1997, Chapter 16.2), and $\boldsymbol{K}_m$ is the $m^2 \times m^2$ vec-permutation matrix such that $\mathrm{vec}(\boldsymbol{B}) = \boldsymbol{K}_m \mathrm{vec}(\boldsymbol{B}')$ when $\boldsymbol{B}$ is an $m \times m$ matrix (see e.g., Harville, 1997, Chapter 16.3). These results lead us to the simple expression of moments of $\boldsymbol{b}_1$ as

$$E[\boldsymbol{b}_1\boldsymbol{b}_1'] = \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}, \tag{18}$$

$$E[\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'] = \frac{1}{\sqrt{n}}\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}), \tag{19}$$

$$\begin{aligned} E[\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'] = (\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}) \\ + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})' + O(n^{-1}). \end{aligned} \tag{20}$$

The result in (18) implies that

$$E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1] = E[\{\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1\boldsymbol{b}_1'\}] = \mathrm{tr}\{\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\} = kr. \tag{21}$$

Similarly, from (19) and (16), we have

$$\begin{aligned} E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2] &= -\frac{1}{2}E[\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})'(\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}] \\ &= -\frac{1}{2\sqrt{n}}\mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\}. \end{aligned} \tag{22}$$

Notice that $\boldsymbol{G}_3(\boldsymbol{\beta})\boldsymbol{K}_{kr} = \boldsymbol{G}_3(\boldsymbol{\beta})$ holds because the derivative is invariant to changes in the order of differentiation. By using this fact and equation (20), the expectation of the first part in (17) is given by

$$\begin{aligned} &E\left[\mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\right\}\right] \\ &\quad= \mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\} \\ &\qquad+ \mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}))'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})) + O(n^{-1}) \\ &\quad= 2\mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{G}_2^{-1}(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2^{-1}(\boldsymbol{\beta}))\right\} \\ &\qquad+ \mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}))'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})) + O(n^{-1}). \end{aligned} \tag{23}$$

Moreover, since the derivative is invariant to changes in the order of differentiation, we can see that $\boldsymbol{G}_4(\boldsymbol{\beta})\boldsymbol{K}_{kr} = \boldsymbol{G}_4(\boldsymbol{\beta})$ and

$$\mathrm{tr}\left\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\} = \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})'\boldsymbol{G}_4(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})).$$

By using the above relations and equation (20), the expectation of the second part in (17) is given by

$$\begin{aligned} &E\left[\mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}\right] \\ &\quad= \mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\} \\ &\qquad+ \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\boldsymbol{G}_4(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} + O(n^{-1}) \\ &\quad= 3\mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\} + O(n^{-1}). \end{aligned} \tag{24}$$

Hence, from equations (17), (23), and (24), we can see that

$$
\begin{aligned}
E\left[\boldsymbol{b}_1'\boldsymbol{G}_2\boldsymbol{b}_3\right] = \mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\otimes\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\}\\
+\frac{1}{2}\mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}))'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}))\\
-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\otimes\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\}+O(n^{-1}).
\end{aligned}
\tag{25}
$$

Consequently, by substituting $E[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{\beta}] = 0$, and equations (21), (22), and (25) into (15), the bias of $-2\ell(\hat{\boldsymbol{\beta}})$ to R is expanded as

$$
B = 2kr + \frac{1}{n}\{\alpha_1(\boldsymbol{\beta}) + \alpha_2(\boldsymbol{\beta}) - \alpha_3(\boldsymbol{\beta})\} + O(n^{-2}),
$$

where coefficients $\alpha_1(\boldsymbol{\beta})$, $\alpha_2(\boldsymbol{\beta})$, and $\alpha_3(\boldsymbol{\beta})$ are given by

$$
\begin{aligned}
\alpha_1(\boldsymbol{\beta}) &= \mathrm{tr}\left\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\otimes\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\},\\
\alpha_2(\boldsymbol{\beta}) &= \mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}))'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}((\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})),\\
\alpha_3(\boldsymbol{\beta}) &= \mathrm{tr}\left\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\otimes\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\right\}.
\end{aligned}
\tag{26}
$$

The CAIC can then be defined by adding an estimated B to $-2\ell(\hat{\boldsymbol{\beta}})$, i.e.,

$$
\mathrm{CAIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2kr + \frac{1}{n}\left\{\alpha_1(\hat{\boldsymbol{\beta}}) + \alpha_2(\hat{\boldsymbol{\beta}}) - \alpha_3(\hat{\boldsymbol{\beta}})\right\}. \tag{27}
$$

For an actual data analysis, an R-script for calculating the CAIC in (27) is provided in Appendix A.3. The CAIC improves the bias of the AIC to $O(n^{-2})$, although the order of the bias of the AIC is $O(n^{-1})$, i.e., the following equations are satisfied:

$$
R - E[\mathrm{AIC}] = O(n^{-1}), \quad R - E[\mathrm{CAIC}] = O(n^{-2}),
$$

where R is the risk function given by (13).

4. NUMERICAL STUDIES

In this section, we conduct numerical studies to show that the CAIC in (27) works better than the crude AIC in (14). To compare the performances of the AIC and the CAIC, the following two properties are considered:

(I) the selection probability: the frequency of the model chosen by minimizing the information criterion.

(II) the prediction error of the best model ($\mathrm{PE_B}$): the risk function of the best model chosen by the information criterion, which is defined by

$$
\mathrm{PE_B} = E[\mathcal{L}(\hat{\boldsymbol{\beta}}_{\mathrm{B}})],
$$

where $\mathcal{L}(\boldsymbol{\beta})$ is the loss function given by (12) and $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ is the MLE of $\boldsymbol{\beta}$ under the best model.

Table 1: Selection probability of the model and the prediction error of the best model

Case	m	Criterion	Selection Probability								$\mathrm{PE_B}$
			M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	
1	20	AIC	1.81	0.29	**74.84**	11.41	4.95	2.88	2.20	1.62	210.06
		CAIC	3.19	0.66	**79.92**	10.01	3.58	1.38	0.88	0.38	207.84
	50	AIC	0.01	0.00	**79.15**	11.16	4.77	2.21	1.56	1.14	511.32
		CAIC	0.01	0.00	**81.25**	10.71	4.27	1.88	1.17	0.71	511.04
2	20	AIC	**77.22**	10.92	4.86	2.69	1.56	1.11	0.77	0.87	202.42
		CAIC	**82.63**	10.06	3.81	2.07	0.76	0.38	0.19	0.10	200.41
	50	AIC	**79.21**	10.89	4.48	2.20	1.24	1.04	0.55	0.39	494.89
		CAIC	**80.99**	10.58	4.10	1.91	1.01	0.70	0.44	0.27	494.63

Note: The selection probability of the true model is marked in bold.

These two properties were evaluated by a Monte Carlo simulation with 10,000 iterations. The information criterion with the higher selection probability of the true model and the smaller prediction error of the best model is regarded as a high-performance model selector. In the basic concept of the AIC, a good model selection method is one that chooses the best model so that the prediction is improved. Hence, $\mathrm{PE_B}$ is a more important property than is the selection probability.

We prepared eight candidate models $M_1, \ldots, M_8$, with $m = 20$ and 50, $n_i = 5$ $(i = 1, \ldots, m)$ and $r = 2$. An $m \times 8$ matrix of explanatory variables $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)'$ was constructed as follows. The first column of $\boldsymbol{X}$ is $\mathbf{1}_m$, where $\mathbf{1}_m$ is an m-dimensional vector of ones, and the remaining seven columns of $\boldsymbol{X}$ were generated randomly from the binomial distribution $B(1, 0.5)$. Simulation data were generated from the multinomial distribution with the true cell probability consisting of $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime})'$. In this simulation study, we prepared two $\boldsymbol{\beta}^*$, as follows:

$$\begin{aligned}
&\text{Case 1}: \quad \boldsymbol{\beta}_1^* = (0, 0.2, -1.0, 0, 0, 0, 0, 0)', \quad \boldsymbol{\beta}_2^* = (-0.1, -0.4, 1.2, 0, 0, 0, 0, 0)', \\
&\text{Case 2}: \quad \boldsymbol{\beta}_1^* = (-0.5, 0, 0, 0, 0, 0, 0, 0)', \quad \boldsymbol{\beta}_2^* = (0.7, 0, 0, 0, 0, 0, 0, 0)'.
\end{aligned}$$

The matrix of explanatory variables in M_j consists of the first j columns of $\boldsymbol{X}$ $(j = 1, \ldots, 8)$. Thus, the true model in Case 1 is M_3, and the true model in Case 2 is M_1.

Table 1 shows the two properties (I) and (II). In the table, the selection probability of the true model is marked in bold. From this table, we can see that the selection probabilities and the prediction errors of the CAIC were improved in comparison with those of the AIC in all situations. We simulated several other models and obtained similar results.

5. CONCLUSION AND DISCUSSION

In this paper, we proposed the CAIC for selecting variables in the multinomial logistic regression model. The proposed CAIC improves the bias of the AIC to $O(n^{-2})$, although the order of the bias of the AIC is $O(n^{-1})$. By using relations between the moments of $\boldsymbol{b}_1$ and expectations of the derivatives of $-\ell(\boldsymbol{\beta})$ instead of directly calculating the moments of $\boldsymbol{y}_i$ to evaluate the moments of $\boldsymbol{b}_1$, a simple expression of the CAIC is developed. Indeed, the bias correction term of the proposed CAIC is represented by only three matrices $\boldsymbol{G}_2(\hat{\boldsymbol{\beta}})$, $\boldsymbol{G}_3(\hat{\boldsymbol{\beta}})$, and $\boldsymbol{G}_4(\hat{\boldsymbol{\beta}})$, which consist of the second, third, and fourth derivatives of $-\ell(\boldsymbol{\beta})$. Even though expressions of $\boldsymbol{G}_2(\hat{\boldsymbol{\beta}})$, $\boldsymbol{G}_3(\hat{\boldsymbol{\beta}})$, and $\boldsymbol{G}_4(\hat{\boldsymbol{\beta}})$ are not simple, we can derive the bias correction term of the CAIC from linear functions of $\boldsymbol{G}_2(\hat{\boldsymbol{\beta}})^{-1}$, $\boldsymbol{G}_3(\hat{\boldsymbol{\beta}})$, and $\boldsymbol{G}_4(\hat{\boldsymbol{\beta}})$. This is a desirable character of the CAIC.

In all situations of the simulation study, the CAIC improved the crude AIC in the sense of making a high selection probability of the true model and a small prediction error of the best model chosen by the information criterion. However, the improvements were smaller when the sample size was large. This is natural because the CAIC is proposed so that the bias of the AIC is corrected when the sample size is small. Needless to say, the AIC and the CAIC are asymptotical equivalents. Hence, the difference between two criteria becomes small when the sample size is increased. The sample sizes of our simulation were 100 and 250. Nevertheless, a clear difference exists in the performances of the CAIC and the AIC. This difference indicates that the CAIC is valuable even when the sample size is not so small. Consequently, we recommend using the CAIC instead of the AIC for selecting multinomial logistic regression models.

The simple expression of the proposed CAIC is based on the property that the second derivatives of $-\ell(\boldsymbol{\beta})$ do not depend on response variables. A generalized linear model (GLM) with a natural link and a known dispersion parameter, e.g., a logistic regression model or a Poisson regression model, will have this property. Then, we can simply express the bias-corrected AIC just like the proposed CAIC in (27) in the same way presented in Section 3. Namely, the bias-corrected AIC with constant second derivatives of the minus log-likelihood function may be stated by

$$\mathrm{CAIC} = \mathrm{AIC} + \gamma_1(\hat{\boldsymbol{\theta}}) + \gamma_2(\hat{\boldsymbol{\theta}}) - \gamma_3(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of unknown parameter $\boldsymbol{\theta}$, and coefficients $\gamma_1(\boldsymbol{\theta})$, $\gamma_2(\boldsymbol{\theta})$, and $\gamma_3(\boldsymbol{\theta})$ are given by

$$\begin{aligned}
\gamma_1(\boldsymbol{\theta}) &= \mathrm{tr}\left\{\boldsymbol{C}(\boldsymbol{\theta})'\boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{C}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})^{-1}\otimes\boldsymbol{H}(\boldsymbol{\theta})^{-1})\right\},\\
\gamma_2(\boldsymbol{\theta}) &= \mathrm{vec}((\boldsymbol{H}(\boldsymbol{\theta})^{-1}))'\boldsymbol{C}(\boldsymbol{\theta})'\boldsymbol{H}(\boldsymbol{\theta})^{-1}\boldsymbol{C}(\boldsymbol{\theta})\mathrm{vec}((\boldsymbol{H}(\boldsymbol{\theta})^{-1})),\\
\gamma_3(\boldsymbol{\theta}) &= \mathrm{tr}\left\{\boldsymbol{Q}(\boldsymbol{\theta})(\boldsymbol{H}(\boldsymbol{\theta})^{-1}\otimes\boldsymbol{H}(\boldsymbol{\theta})^{-1})\right\}.
\end{aligned}$$

Here, $\boldsymbol{H}(\boldsymbol{\theta})$, $\boldsymbol{C}(\boldsymbol{\theta})$, and $\boldsymbol{Q}(\boldsymbol{\theta})$ are matrices consisting of the second, third, and fourth derivatives, respectively, of the minus log-likelihood function and are defined by (A.5) in Appendix A.2.

APPENDIX

A.1. EXPLICIT FORMS OF $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, AND $\boldsymbol{G}_4(\boldsymbol{\beta})$

In this subsection, for simplicity, we write $\boldsymbol{\Sigma}_i(\boldsymbol{\beta})$, $\boldsymbol{p}_i(\boldsymbol{\beta})$, and $p_{ij}(\boldsymbol{\beta})$ as $\boldsymbol{\Sigma}_i$, $\boldsymbol{p}_i$, and p_{ij}, respectively.

Notice that

$$\frac{\partial \boldsymbol{p}_i}{\partial \boldsymbol{\beta}_j'} = p_{ij}\left(\boldsymbol{e}_j - \boldsymbol{p}_i\right)\boldsymbol{x}_i', \quad (j = 1, \ldots, r),$$

where $\boldsymbol{e}_j$ is the jth coordinate unit vector, which is used in equation (9). This result and equation (3) imply that

$$\frac{\partial \boldsymbol{p}_i}{\partial \boldsymbol{\beta}'} = \left(p_{i1}\left(\boldsymbol{e}_1 - \boldsymbol{p}_i\right)\boldsymbol{x}_i', \ldots, p_{ir}\left(\boldsymbol{e}_r - \boldsymbol{p}_i\right)\boldsymbol{x}_i'\right) = \boldsymbol{\Sigma}_i \otimes \boldsymbol{x}_i'.$$

Substituting the above result into the definition of $\boldsymbol{G}_2(\boldsymbol{\beta})$ yields equation (6). Furthermore, from the definitions of $\boldsymbol{G}_3(\boldsymbol{\beta})$ and $\boldsymbol{G}_4(\boldsymbol{\beta})$, we can see that $\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta})$ and $\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta})$ in (7) and (8), respectively, satisfy

$$\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \boldsymbol{\Sigma}_i, \quad \boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \otimes \boldsymbol{\Sigma}_i.$$

Notice that the (a, b)th element of $\boldsymbol{\Sigma}_i$ is $p_{ia}\delta_{ab} - p_{ia}p_{ib}$, where δ_{ab} is the Kronecker delta, i.e., $\delta_{aa} = 1$ and $\delta_{ab} = 0$ for $a \neq b$. This equation leads us to other expressions of $\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta})$ and $\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta})$, as follows:

$$\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) = \sum_{a,b}^{r} \frac{\partial}{\partial \boldsymbol{\beta}'}(p_{ia}\delta_{ab} - p_{ia}p_{ib}) \otimes \boldsymbol{e}_a\boldsymbol{e}_b', \quad \boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) = \sum_{a,b}^{r} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}(p_{ia}\delta_{ab} - p_{ia}p_{ib}) \otimes \boldsymbol{e}_a\boldsymbol{e}_b'. \quad \text{(A.1)}$$

Derivatives of p_{ia} are calculated as

$$\begin{aligned}
\frac{\partial p_{ia}}{\partial \boldsymbol{\beta}} &= p_{ia}(\boldsymbol{e}_a - \boldsymbol{p}_i) \otimes \boldsymbol{x}_i, \\
\frac{\partial^2 p_{ia}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= p_{ia}(\boldsymbol{e}_a - \boldsymbol{p}_i)(\boldsymbol{e}_a - \boldsymbol{p}_i)' \otimes \boldsymbol{x}_i\boldsymbol{x}_i - p_{ia}\boldsymbol{\Sigma}_i \otimes \boldsymbol{x}_i\boldsymbol{x}_i' \\
&= p_{ia}\left\{(\boldsymbol{e}_a - \boldsymbol{p}_i)(\boldsymbol{e}_a - \boldsymbol{p}_i)' - \boldsymbol{\Sigma}_i\right\} \otimes \boldsymbol{x}_i\boldsymbol{x}_i, \\
\frac{\partial^2 p_{ia}p_{ib}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= p_{ib}\frac{\partial^2 p_{ia}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} + \frac{\partial p_{ib}}{\partial \boldsymbol{\beta}}\frac{\partial p_{ia}}{\partial \boldsymbol{\beta}'} + \frac{\partial p_{ia}}{\partial \boldsymbol{\beta}}\frac{\partial p_{ib}}{\partial \boldsymbol{\beta}'} + p_{ia}\frac{\partial^2 p_{ib}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \\
&= p_{ia}p_{ib}\left\{(\boldsymbol{e}_a + \boldsymbol{e}_b - 2\boldsymbol{p}_i)(\boldsymbol{e}_a + \boldsymbol{e}_b - 2\boldsymbol{p}_i)' - 2\boldsymbol{\Sigma}_i\right\} \otimes \boldsymbol{x}_i\boldsymbol{x}_i'.
\end{aligned}$$

By substituting the above derivatives into (A.1), we have

$$\begin{aligned}
\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) &= \sum_{a,b}^{r} \left\{\delta_{ab}p_{ia}(\boldsymbol{e}_a - \boldsymbol{p}_i)' - p_{ia}p_{ib}(\boldsymbol{e}_a - \boldsymbol{p}_i)' - p_{ia}p_{ib}(\boldsymbol{e}_b - \boldsymbol{p}_i)'\right\} \otimes \boldsymbol{x}_i' \otimes \boldsymbol{e}_a\boldsymbol{e}_b' \\
&= \sum_{a,b}^{r} p_{ia}\left\{(\delta_{ab} - p_{ib})(\boldsymbol{e}_a - \boldsymbol{p}_i)' - p_{ib}(\boldsymbol{e}_b - \boldsymbol{p}_i)'\right\} \otimes \boldsymbol{x}_i' \otimes \boldsymbol{e}_a\boldsymbol{e}_b' \\
&= \sum_{a=1}^{r} p_{ia}(\boldsymbol{e}_a \otimes \boldsymbol{x}_i)' \otimes \left\{(\boldsymbol{e}_a - \boldsymbol{p}_i)(\boldsymbol{e}_a - \boldsymbol{p}_i)'\right\} - (\boldsymbol{p}_i \otimes \boldsymbol{x}_i)' \otimes \boldsymbol{\Sigma}_i,
\end{aligned}$$

and

$$
\begin{aligned}
\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) &= \sum_{a,b}^{r} p_{ia}\Big(\delta_{ab}\left\{(\boldsymbol{e}_a-\boldsymbol{p}_i)(\boldsymbol{e}_a-\boldsymbol{p}_i)'-\boldsymbol{\Sigma}_i\right\} \\
&\qquad\qquad - p_{ib}\left\{(\boldsymbol{e}_a+\boldsymbol{e}_b-2\boldsymbol{p}_i)(\boldsymbol{e}_a+\boldsymbol{e}_b-2\boldsymbol{p}_i)'-2\boldsymbol{\Sigma}_i\right\}\Big)\otimes \boldsymbol{x}_i\boldsymbol{x}_i'\otimes \boldsymbol{e}_a\boldsymbol{e}_b' \\
&= \sum_{a=1}^{r} p_{ia}(\boldsymbol{e}_a-\boldsymbol{p}_i)(\boldsymbol{e}_a-\boldsymbol{p}_i)'\otimes \boldsymbol{x}_i\boldsymbol{x}_i'\otimes(\boldsymbol{e}_a\boldsymbol{e}_a'-\boldsymbol{e}_a\boldsymbol{p}_i'-\boldsymbol{p}_i\boldsymbol{e}_a') \\
&\quad - \boldsymbol{\Sigma}_i\otimes \boldsymbol{x}_i\boldsymbol{x}_i'\otimes(\boldsymbol{\Sigma}_i-\boldsymbol{p}_i\boldsymbol{p}_i') - \sum_{a,b}^{r} p_{ia}p_{ib}(\boldsymbol{e}_a-\boldsymbol{p}_i)(\boldsymbol{e}_b-\boldsymbol{p}_i)'\otimes \boldsymbol{x}_i\boldsymbol{x}_i'\otimes(\boldsymbol{e}_a\boldsymbol{e}_b'+\boldsymbol{e}_b\boldsymbol{e}_a').
\end{aligned}
$$

The above two equations indicate that explicit forms of $\boldsymbol{G}_3(\boldsymbol{\beta})$ and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are given in (7) and (8), respectively.

A.2. EXPECTATIONS OF DERIVATIVES OF THE MINUS LOG-LIKELIHOOD FUNCTION

In this subsection, we derive general formulas of the expectations of derivatives of the minus log-likelihood function. Let $f(\boldsymbol{u}|\boldsymbol{\theta})$ be a joint probability density function of $\boldsymbol{u}$ specified by q-dimensional parameter vector $\boldsymbol{\theta}$, and $L(\boldsymbol{\theta})$ be a minus log-likelihood function defined by $L(\boldsymbol{\theta}) = -\log f(\boldsymbol{u}|\boldsymbol{\theta})$. Suppose that

$$
\dot{f}_{a_1\cdots a_j} = \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} f(\boldsymbol{u}|\boldsymbol{\theta}), \quad \dot{L}_{a_1\cdots a_j} = \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} L(\boldsymbol{\theta}).
$$

By carrying out tedious calculations, we have

$$
\begin{aligned}
&\dot{L}_a = -\frac{\dot{f}_a}{f}, \qquad \dot{L}_{ab} = \dot{L}_a\dot{L}_b - \frac{\dot{f}_{ab}}{f}, \quad \dot{L}_{abc} = -\dot{L}_a\dot{L}_b\dot{L}_c + \sum_{[3]} \dot{L}_a\dot{L}_{bc} - \frac{\dot{f}_{abc}}{f}, \\
&\dot{L}_{abcd} = \dot{L}_a\dot{L}_b\dot{L}_c\dot{L}_d - \sum_{[6]} \dot{L}_a\dot{L}_b\dot{L}_{cd} + \sum_{[3]} \dot{L}_{ab}\dot{L}_{cd} + \sum_{[4]} \dot{L}_a\dot{L}_{bcd} - \frac{\dot{f}_{abcd}}{f},
\end{aligned}
\tag{A.2}
$$

where we simplify $f(\boldsymbol{u}|\boldsymbol{\theta})$ as f, and $\sum_{[j]}$ is the summation of a total of j terms of different combinations, e.g., $\sum_{[3]} \dot{L}_{ab}\dot{L}_{cd} = \dot{L}_{ab}\dot{L}_{cd} + \dot{L}_{ac}\dot{L}_{bd} + \dot{L}_{ad}\dot{L}_{bc}$. It follows from $\int f d\boldsymbol{u} = 1$ that

$$
E\left[\frac{\dot{f}_{a_1\cdots a_j}}{f}\right] = \int \frac{\dot{f}_{a_1\cdots a_j}}{f} f d\boldsymbol{u} = \int \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} f d\boldsymbol{u} = \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} \int f d\boldsymbol{u} = 0. \tag{A.3}
$$

The above equation can be satisfied when $\boldsymbol{u}$ is continuous. Even when $\boldsymbol{u}$ is discrete, we can obtain the same result by replacing the integration with a summation. Equations (A.2) and (A.3) imply that

$$
\begin{aligned}
&E[\dot{L}_a] = 0, \quad E[\dot{L}_{ab}] = E[\dot{L}_a\dot{L}_b], \quad E[\dot{L}_{abc}] = -E[\dot{L}_a\dot{L}_b\dot{L}_c] + \sum_{[3]} E[\dot{L}_a\dot{L}_{bc}], \\
&E[\dot{L}_{abcd}] = E[\dot{L}_a\dot{L}_b\dot{L}_c\dot{L}_d] - \sum_{[6]} E[\dot{L}_a\dot{L}_b\dot{L}_{cd}] + \sum_{[3]} E[\dot{L}_{ab}\dot{L}_{cd}] + \sum_{[4]} E[\dot{L}_a\dot{L}_{bcd}].
\end{aligned}
\tag{A.4}
$$

Let us consider a vector of the first derivatives, and matrices of the second, third, and fourth derivatives, which are defined as

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\theta}) &= -\frac{\partial}{\partial \boldsymbol{\theta}}\ell(\boldsymbol{\theta}), & \boldsymbol{H}(\boldsymbol{\theta}) &= -\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\ell(\boldsymbol{\theta}), \\ \boldsymbol{C}(\boldsymbol{\theta}) &= -\left(\frac{\partial}{\partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\ell(\boldsymbol{\theta}), & \boldsymbol{Q}(\boldsymbol{\theta}) &= -\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\ell(\boldsymbol{\theta}). \end{aligned} \tag{A.5}$$

From the expectations in (A.4), we obtain $E[\boldsymbol{H}(\boldsymbol{\theta})]$, $E[\boldsymbol{C}(\boldsymbol{\theta})]$, and $E[\boldsymbol{Q}(\boldsymbol{\theta})]$ as

$$\begin{aligned} E[\boldsymbol{H}(\boldsymbol{\theta})] &= E[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \\ E[\boldsymbol{C}(\boldsymbol{\theta})] &= -E[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] \\ &\quad + E[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{H}(\boldsymbol{\theta})] + E[\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{g}(\boldsymbol{\theta})'] + E[\boldsymbol{g}(\boldsymbol{\theta})\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'], \\ E[\boldsymbol{Q}(\boldsymbol{\theta})] &= E[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] \\ &\quad - (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)E[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{H}(\boldsymbol{\theta})](\boldsymbol{I}_{q^2} + \boldsymbol{K}_q) \\ &\quad - E[\mathrm{vec}(\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})')\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'] - E[\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})')'] \\ &\quad + (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)E[\{\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{H}(\boldsymbol{\theta})\}] + E[\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'] \\ &\quad + E[\{\boldsymbol{g}(\boldsymbol{\theta}) \otimes \boldsymbol{C}(\boldsymbol{\theta})\}](\boldsymbol{I}_{q^2} + \boldsymbol{K}_q) + (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)E[\{\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{C}(\boldsymbol{\theta})'\}]. \end{aligned} \tag{A.6}$$

Recall that $E[\boldsymbol{g}(\boldsymbol{\theta})] = \boldsymbol{0}_q$ holds. Furthermore, we note that $\boldsymbol{C}(\boldsymbol{\theta})$ and $\boldsymbol{Q}(\boldsymbol{\theta})$ are constant when $\boldsymbol{H}(\boldsymbol{\theta})$ is constant. Hence, when $\boldsymbol{H}(\boldsymbol{\theta})$ is constant, $\boldsymbol{H}(\boldsymbol{\theta})$, $\boldsymbol{C}(\boldsymbol{\theta})$, and $\boldsymbol{Q}(\boldsymbol{\theta})$ become simpler, as follows:

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\theta}) &= E[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \\ \boldsymbol{C}(\boldsymbol{\theta}) &= -E[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \\ \boldsymbol{Q}(\boldsymbol{\theta}) &= E[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] - (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)\{\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{H}(\boldsymbol{\theta})\} - \mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'. \end{aligned} \tag{A.7}$$

A.3. R-SCRIPT FOR CALCULATING THE CAIC

In this subsection, we provide the R-script for calculating the CAIC in (27). In the script, a variable `Y` corresponds to the $m \times (r+1)$ matrix whose $(a, b+1)$th element is y_{ab}, and a variable `X` corresponds to the $m \times k$ matrix $\boldsymbol{X}$ whose first column is $\boldsymbol{1}_m$. When we fit the multinomial logisitic regression model to the data, it is only necessary to carry out the command `vglm(Y~X,,family=multinomial)`.

```
library(VGAM)
# X: Explanatory variables
# Y: Response variables
CAIC.MLRM.f <- function(X,Y){
        vec.func <- function(A){
        x <- dim(A)[1]
```

```
        y <- dim(A)[2]
        Res <- NULL
        for (i in 1:y){
                temp <- t(t(A[,i]))
                Res <- rbind(Res,temp)}
        Res}
        ei.vec <- function(p,i){
        e <- matrix(c(rep(0,p)),ncol=1)
        e[i] <- 1
        e}

n <- sum(Y)
k <- dim(X)[2]
rho <- apply(Y,1,sum)/n

if(dim(X)[2]==1){est <- vglm(Y~1,family=multinomial)}
if(dim(X)[2]>1){est <- vglm(Y~X[,-1],family=multinomial)}

AIC <- AIC(est)
P <- fitted(est)

G_2 <- matrix(0,(k*r),(k*r))
G_3 <- matrix(0,k*r,(k*r)^2)
G_4 <- matrix(0,(k*r)^2,(k*r)^2)
for (i in 1:m){
        x_i <- X[i,]
        xx_i <- x_i%*%t(x_i)
        p_i <- P[i,-1]
        pp_i <- p_i%*%t(p_i)
        S_i <- diag(p_i)-pp_i
        G_2 <- G_2+rho[i]*(S_i%x%xx_i)
        D3 <- (-1)*(t(p_i%x%x_i)%x%S_i)
        D4 <- (-1)*(S_i%x%xx_i%x%(S_i-pp_i))
        for (j in 1:r){
                e_j <- ei.vec(r,j)
```

```
                e_jj <- e_j%*%t(e_j)
                r_j <- e_j-p_i
                r_jj <- r_j%*%t(r_j)
                pe_j <- p_i%*%t(e_j)
                ep_j <- t(pe_j)
                D3 <- D3+p_i[j]*(t(e_j%x%x_i)%x%r_jj)
                D4 <- D4+p_i[j]*(r_jj%x%xx_i%x%(e_jj-ep_j-pe_j))
                        for (s in 1:r){
                        e_s <- ei.vec(r,s)
                        e_js <- e_j%*%t(e_s)
                        e_sj <- t(e_js)
                        r_s <- e_s-p_i
                        r_js <- r_j%*%t(r_s)
                        D4 <- D4-p_i[j]*p_i[s]*(r_js%x%xx_i%x%(e_js+e_sj))}}
        G_3 <- G_3+rho[i]*D3%x%xx_i
        G_4 <- G_4+rho[i]*D4%x%xx_i}

invG_2 <- solve(G_2)
vinvG_2 <- vec.func(invG_2)
a1 <- sum(diag(G_3%*%(invG_2%x%invG_2)%*%t(G_3)%*%invG_2))
a2 <- t(vinvG_2)%*%t(G_3)%*%invG_2%*%G_3%*%vinvG_2
a3 <- sum(diag(G_4%*%(invG_2%x%invG_2)))

CAIC <- AIC+(1/n)*(a1+a2-a3)
CAIC}
```


ACKNOWLEDGMENTS

This research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Challenging Exploratory Research, #22650058, 2010-2012, and Grant-in-Aid for Young Scientists (B), #23700337, 2011–2014.

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov and F. Csáki), 267–281, Akadémiai Kiadó, Budapest.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716–723.

[3] Briz, T. & Ward, R. W. (2009). Consumer awareness of organic products in Spain: An application of multinomial logit models. *Food Policy*, **34**, 295–304.

[4] Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach* (2nd ed.). Springer-Verlag, New York.

[5] Choi, S.-W., Sohngen, B. & Alig, R. (2011). An assessment of the influence of bioenergy and marketed land amenity values on land uses in the Midwestern US. *Ecol. Econ.*, **70**, 713–720.

[6] dell'Olio, L., Ibeas, A. & Cecin, P. (2011). The quality of service desired by public transport users. *Transport Policy*, **18**, 217–227.

[7] Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**, 342–368.

[8] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

[9] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag, New York.

[10] Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, Inc., New York.

[11] Kamo, K., Yanagihara, H. & Satoh, K. (2012). Bias-corrected AIC for selecting variables in Poisson regression models. *Comm. Statist. Theory Methods* (in press).

[12] Konishi, S. (1999). Statistical model evaluation and information criteria. In *Multivariate Analysis, Design of Experiments, and Survey Sampling* (ed. S. Ghosh), 369–399, Marcel Dekker, New York.

[13] Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Science+Business Media, LLC, New York.

[14] Kullback, S. & Leibler, R. (1951). On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

[15] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[16] Yanagihara, H., Sekiguchi, R. & Fujikoshi, Y. (2003). Bias correction of AIC in logistic regression models. *J. Statist. Plann. Inference*, **115**, 349–360.