広島大学学位請求論文

Model Selection Criteria in Generalized Linear Models and their Extensions

(一般化線形モデル及びその拡張における モデル選択規準)

2014年

広島大学大学院理学研究科
数学専攻
伊森晋平

目　次

主 論 文

Model Selection Criteria in Generalized Linear Models and their Extensions

Shinpei Imori

Graduate School of Science
Hiroshima University
1-3-1 Kagamiyama
Higashi-Hiroshima 739-8526
Japan

Contents

Chapter 1

Introduction

In real data analysis, deciding the best model among a set of candidate models is an important problem. There have been a lot of literature to consider such model selection problems from the various standpoints. For example, a subset selection of explanatory variables in regression models in order to predict the future data is often considered. It is common for a model selection method to measure the goodness of fit of the model for the future data by the risk function based on the expected Kullback-Leibler (KL) information (Kullback & Leibler, 1951). For actual use, we must estimate the risk function, which depends on unknown parameters. The most famous estimator of the risk function is Akaike's information criterion (AIC) proposed by Akaike (1973, 1974). Since the AIC can be simply defined as -2 × "the maximum log-likelihood" $+2$ × "the number of parameters", the AIC is widely applied in chemometrics, engineering, econometrics, psychometrics, and many other fields for selecting appropriate models using a set of explanatory variables (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 addition, the order of the bias of the AIC to the risk function is $O(n^{-1})$, which indicates implicitly that the AIC sometimes has a non-negligible bias to the risk function when the sample size n is not so large. The AIC tends to underestimate the risk function and the bias of AIC is apt to increase with the number of parameters in the model. Potentially, the AIC has a tendency to choose the model that has more parameters than the true model as the best model Shibata (1980). Combined with these characteristics, the bias will cause a disadvantage whereby the model having the most parameters is easily chosen by the AIC among the candidate models as the best model. Such problem is often resolved by using a bias-corrected AIC (see e.g., Burnham & Anderson, 2002, Chapter 2.4). A number of authors have investigated bias-corrected AIC for various models. For example, Sugiura (1978) developed an unbiased estimator of the risk function in linear regression models, which is the UMVUE of the risk function reported by Davies, et al. (2006). Hurvich & Tsai (1989) formally adjusted the bias of the AIC (called AICc) in several models. In particular, the AICc is equivalent to Sugiura's bias-corrected AIC in the case of the linear regression model. Wong & Li (1998) extended Hurvich and Tsai's AICc to a wider model and verified that their AICc has a higher performance than the original AIC by conducting numerical studies.

Unfortunately, except for the linear regression model, the AICc does not completely reduce the bias of the AIC to $O(n^{-2})$. As mentioned previously, the goodness of fit of the model is measured by the risk function based on the expected KL information. Thus, obtaining a higher-order asymptotic unbiased estimator of the risk function will allow us to more accurately measure the goodness of fit of the model. This will further facilitate the reasonable selection of variables. From this viewpoint, Yanagihara, et al. (2003) and Kamo, et al. (2013) proposed the bias-corrected AIC's in the logistic model and the Poisson regression model, respectively, each of which reduce the bias of the AIC to $O(n^{-2})$ under the assumption that the candidate model includes the true model. We refer to the completely bias-corrected AIC to $O(n^{-2})$ as the corrected AIC (called CAIC). Frequently, the CAIC improves the performance of the original AIC dramatically. This strongly suggests the usefulness of the CAIC for real data analysis.

Nevertheless, the CAIC is rarely used in real data analysis because the CAIC has been derived only in a few models. Moreover, since the derivation of the bias is complicated, a great deal of practice is needed in order to carry out the calculation of the CAIC if a researcher wants to use the CAIC in a model in which the CAIC has not been derived. Therefore, the application of the CAIC to real data analysis is not penetrated, although the CAIC has better performance than the original AIC. If we can obtain the CAIC in a small amount of time, the CAIC will become a useful and user-friendly model selector.

In the former half of this paper, we attempt to expand the CAIC to two different models. One of the model considered is generalized linear model (GLM), which is a broad model class proposed by Nelder & Wedderburn (1972), and another is the multinomial logistic regression model, which is a generalization of the logistic regression model into multivariate data. Hence, the multinomial logistic regression model is regarded as a part of generalization of the GLM. The GLM can express a number of statistical models by changing the distribution and the link function, such as the normal linear regression model, the logistic regression model, and the probit model, which are currently commonly used in a number of applied fields (cf. Barnett & Nurmagambetov, 2010; Matas, et al., 2010; Sánchez-Carneo, et al., 2011; Teste & Lieffers, 2011). On the other hand, the multinomial logistic regression model is a regression model that generalizes a logistic regression by allowing more than two discrete response variables. 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 many 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, et al., 2011; dell'Olio et al., 2011).

Generally, 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. Although such moments should be calculated for each specified model, we emphasize that the moments do not remain in our formulation of the CAIC since the moments are represented by the moments of response variables. Practically speaking, the GLM and the multinomial logistic regression model can be easily fitted to real data using the "glm" and "vglm" function, respectively, in "R" (R Development Core Team, 2011) that is a free software environment for statistical computing and graphics. Therefore, the CAIC is confirmed useful in real data analysis since we can easily calculate the CAIC by using such software and the model class we considered herein is wide and can be easily fitted to real data.

In the latter of this paper, we consider the model selection problem in the longitudinal data analyzed in biomedical and epidemiological researches, it is often the case that the responses within individuals are dependent. The generalized estimating equation (GEE) approach was developed by Liang & Zeger (1986) for estimating regression coefficients in such correlated data; it is an expansion of the likelihood equation in the GLM. Using a GEE relaxes the assumption of joint distribution for the observations. We can use the GEE by only assuming a marginal distribution of each response and a working correlation structure, which is allowed to include an unknown parameter. Furthermore, under certain conditions, the GEE estimator is asymptotically normally distributed and consistent even when the working correlation structure has been misspecified (Liang & Zeger, 1986). However, some studies have noted that a misspecification of the working correlation structure may induce undesirable results. For instance, Crowder (Crowder, 1995) showed that a misspecification of the working correlation structure may ruin the asymptotic normality of the GEE estimator, since the parameter of the working correlation structure may not be minimized in the interior of the parameter space. Fitzmaurice (Fitzmaurice, 1995) showed that a GEE estimator is less efficient when an independent structure is assumed to the working correlation matrix. Thus, it is important to adequately determine the working correlation structure, although the primary use of the GEE approach is to estimate the regression parameter. Although we can estimate the correct correlation structure by using an unstructured correlation matrix, it is better not to use this as the working correlation matrix, since it may increase the variance of the GEE estimator unless the response has low dimensionality or the sample size is sufficiently large. Thus, we often wish to obtain a correct and lower-dimensional correlation structure.

Recently, a number of papers have considered the selection of a working correlation structure. Besides the AIC we introduced above, there are adequate criteria to select the best model. The Bayesian information criterion (BIC) proposed by Schwarz (1978) are often used to select the true model, due to the theoretical validity (Nishii, 1984; Shao, 1997). The BIC is defined by replacing the penalty term of the AIC, which is 2 × "the number of parameters", as "the logarithm of sample size" × "the number of parameters". The BIC and the GIC (Nishii, 1984), which is a generalization of the penalty term in the AIC and the BIC, can be used to select the true model since their selection probabilities of the true model goes to 1, which is called the consistency. However, we cannot use the information criteria such as the AIC, the BIC and the GIC since the

GEE approach does not assume the joint distribution of responses. Then, Pan (2001) considered using the quasi-likelihood instead of the likelihood and derived the quasi-likelihood under the independence model criterion (QIC), which is an AIC-type criterion. These criteria may be used to select a subset of explanatory variables rather than a working correlation structure. The correlation information criterion (CIC) (Hin & Wang, 2009) was derived from the penalty term of the QIC, and this improves the selection of the correlation structure. In addition, there have been some methods proposed for selecting the best working correlation structure. Pan & Connett (2002) attempted to select the working correlation structure that minimizes the mean squared prediction error estimated by a resampling method. Hin, et al. (2007) proposed a criterion based on a measurement between the true correlation and the candidate correlation structure. Chen & Lazar (2012) used an empirical likelihood approach to construct a model selection criterion. All of these works use different ways to measure the difference between two matrices. Although there are more studies that have considered the selection of the working correlation structure, little attention has been paid to the theoretical properties of these criteria.

Hence, we propose a GIC-type criterion that can be used to select the true correlation structure. Furthermore, we attempt to determine sufficient conditions for the GIC-type criterion to be consistent. Since we do not assume a joint distribution, as discussed above, we need an alternative measurement. Thus, we consider to use a loss function instead of the likelihood. In this study, our criterion is constructed based on Stein's loss function (James & Stein, 1961), which is one of the famous loss function for matrices. Moreover, we can show the consistency property of our criterion.

The remainder of the paper is organized as follows: In Chapter 2, we propose the CAIC in the GLMs, which is based on the result of Imori, et al. (2014). In Chapter 3, we propose the CAIC in the multinomial logistic regression model, which is based on the result of Yanagihara, et al. (2012). In Chapter 4, we propose the criterion in order to select the true correlation structure and show the consistency of this criterion, which is based on the result of Imori (2014). Technical details are provided in the Appendix.

Chapter 2

Simple Formula for Calculating Bias-Corrected AIC in Generalized Linear Models

Chapter 2 is organized as follows: In Section 2.1, we consider a stochastic expansion of the maximum likelihood estimator (MLE) in the GLM. In Section 2.2, we propose a new information criterion by reducing the bias of the AIC in the GLMs to $O(n^{-2})$. In Section 2.3, we investigate the performance of the proposed CAIC through numerical simulations. Technical details are provided in Appendix A.1 and Appendix A.2.

2.1 Stochastic Expansion of the MLE in the GLM

The GLM considered herein is developed to allow us to fit regression models for the response variables that follow a very general distribution belonging to the exponential family, the probability density function of which is given as follows:

$$f(y;\theta,\phi) = \exp\left\{\frac{\theta y - a(\theta)}{\phi} + b(y,\phi)\right\}, \qquad (2.1.1)$$

where $a(\cdot)$ and $b(\cdot)$ are known functions, the unknown parameter θ is referred to as the natural location parameter, and ϕ is often referred to as the scale parameter. (For the details of the GLM, see, e.g., McCullagh & Nelder, 1989; Meyers, et al., 2002). In the present section, we assume that ϕ is known. The exponential family includes the normal, binomial, Poisson, geometric, negative binomial, exponential, gamma, and inverse normal distributions. Let the data consist of a sequence $\{(y_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$, where $y_1, \ldots, y_n$ are independent random variables referred to as response variables, and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are p-dimensional non-stochastic vectors referred to as explanatory variables. The expectation of the response y_i is related to the linear predictor $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ by a link function $h(\cdot)$, i.e., $h(\mathrm{E}[y_i]) = h(\mu(\theta_i)) = \eta_i$. For theoretical purposes, we define $u = (h \circ \mu)^{-1}$, i.e., $\theta_i = u(\eta_i)$. When $h = \mu^{-1}$, i.e., u is an identity function, we say that h is the natural link function. For example, the logistic regression model uses the natural link function. Finally, the candidate model is expressed as

$$y_i \overset{\text{Indep}}{\sim} f(y_i; \theta_i(\boldsymbol{\beta}), \phi),$$

where $f(\cdot)$ is given by (2.1.1). The p-dimensional unknown vector $\boldsymbol{\beta}$ can be estimated by the maximum likelihood method. The joint probability density function of $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is given by

$$f(\boldsymbol{y}; \boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i; \theta_i(\boldsymbol{\beta}), \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{\theta_i y_i - a(\theta_i)}{\phi} + b(y_i, \phi) \right\}.$$

Hence, the log-likelihood function of the GLM is expressed as

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \log f(\boldsymbol{y}; \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{\theta_i y_i - a(\theta_i)}{\phi} + b(y_i, \phi) \right\}.$$

Let $\hat{\boldsymbol{\beta}}$ be the MLE of $\boldsymbol{\beta}$. Here, $\hat{\boldsymbol{\beta}}$ is given as the solution of the following likelihood equation:

$$\frac{\partial \ell(\boldsymbol{\beta}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \sum_{i=1}^{n} (y_i - a_{i1}) c_{i1} \boldsymbol{x}_i = \frac{1}{\phi} \boldsymbol{X}' \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}_p,$$

where

$$a_{ij} = \frac{\partial^j a(\theta_i)}{\partial \theta_i^j}, \quad c_{ij} = \frac{\partial^j \theta_i}{\partial \eta_i^j},$$

$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, $\boldsymbol{\Delta} = \text{diag}\{c_{11}, \ldots, c_{n1}\}$, $\boldsymbol{\mu} = (a_{11}, \ldots, a_{n1})'$, and $\mathbf{0}_p$ is a p-dimensional vector of zeros. Note that $a(\theta)$ is a C^∞-class function and all of the orders of the moments of $\boldsymbol{y}$ exist in the interior Θ^0 of the natural parameter space Θ, and that a_{ij} is determined by the distribution of the model and c_{ij} is determined by the link function. In using some of the properties of the MLE, we have the following regularity assumptions (see, e.g., Fahrmeir & Kaufmann, 1985):

(A1) : $\boldsymbol{x}_i'\boldsymbol{\beta} \in h(\mathcal{M})$, $i = 1, \ldots, n$, for all $\boldsymbol{\beta} \in \mathcal{B}$,

(A2) : h is three times continuously differentiable,

(A3) : For all $\boldsymbol{x}_i \in \mathcal{F}$, $c_{i1} \neq 0$, $i = 1, \ldots, n$,

(A4) : $^\exists n_0$ $s.t.$ $\boldsymbol{X}'\boldsymbol{X}$ has full rank for $n \geq n_0$,

where $\mathcal{B}$ is an admissible open set in $\mathbb{R}^p$ for the parameter $\boldsymbol{\beta}$, $\mathcal{F}$ is a compact set for the regressors $\boldsymbol{x}_i$, and $\mathcal{M}$ denotes the image $\mu(\Theta^0)$. Condition (A1) is necessary in order to obtain the GLM for all $\boldsymbol{\beta}$. Condition (A2) is necessary in order to calculate the bias. Conditions (A3) and (A4) ensure that $\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{X}$ is positive definite for all $\boldsymbol{\beta} \in \mathcal{B}$, $n \geq n_0$, where

$$\boldsymbol{V} = \phi \ \text{diag}\{a_{12}, \ldots, a_{n2}\}.$$

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

(A5) : sequence $\{\boldsymbol{x}_i\}$ lies in $\mathcal{F}$ with $u(\boldsymbol{x}_i'\boldsymbol{\beta}) \in \Theta^0$, $\boldsymbol{\beta} \in \mathcal{B}$,

(A6) : $\liminf_{n\to\infty} \lambda_{\min}(\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{X}/n) > 0$,

(A7) : $^\exists c > 0$, n_1, $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) > c\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})$, $n \geq n_1$,

where $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ are the smallest and the largest eigenvalues of symmetric matrix $\boldsymbol{A}$, respectively. According to Theorem 5 in Fahrmeir & Kaufmann (1985), $\hat{\boldsymbol{\beta}}$ has strong consistency and asymptotic normality under these conditions. Furthermore, from (A6), $\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{V}\boldsymbol{\Delta}\boldsymbol{X} = O(n)$, with $n \to \infty$.

Based on the above conditions, $\hat{\boldsymbol{\beta}}$ can be formally expanded as follows:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{b}_1 + \frac{1}{n}\boldsymbol{b}_2 + \frac{1}{n\sqrt{n}}\boldsymbol{b}_3 + O_p(n^{-2}). \tag{2.1.2}$$

Note that $\partial \ell(\hat{\boldsymbol{\beta}};\boldsymbol{y})/\partial\boldsymbol{\beta} = \mathbf{0}_p$. By applying a Taylor expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ to this equation, the likelihood equation is expanded as follows:

$$\begin{aligned}\mathbf{0}_p = &\frac{1}{\sqrt{n}}(\boldsymbol{g} + \boldsymbol{G}_2\boldsymbol{b}_1) + \frac{1}{n}\left\{\boldsymbol{G}_2\boldsymbol{b}_2 + \frac{1}{2}\boldsymbol{G}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)\right\} \\ &+ \frac{1}{n\sqrt{n}}\left\{\boldsymbol{G}_2\boldsymbol{b}_3 + \frac{1}{2}\boldsymbol{G}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) \right. \\ &\qquad \left. + \frac{1}{6}(\boldsymbol{I}_p \otimes \boldsymbol{b}_1')\boldsymbol{G}_4(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)\right\} + O_p(n^{-2}), \end{aligned} \tag{2.1.3}$$

where

$$\begin{aligned}\boldsymbol{g} &= \frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial\boldsymbol{\beta}} = \frac{1}{\sqrt{n}\phi}\sum_{i=1}^{n}(y_i - a_{i1})c_{i1}\boldsymbol{x}_i, \\ \boldsymbol{G}_2 &= \frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\beta};\boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = -\frac{1}{n\phi}\sum_{i=1}^{n}\{a_{i2}c_{i1}^2 - (y_i - a_{i1})c_{i2}\}\boldsymbol{x}_i\boldsymbol{x}_i', \\ \boldsymbol{G}_3 &= \frac{1}{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta};\boldsymbol{y}), \\ &= -\frac{1}{n\phi}\sum_{i=1}^{n}\{a_{i3}c_{i1}^3 + 3a_{i2}c_{i1}c_{i2} - (y_i - a_{i1})c_{i3}\}(\boldsymbol{x}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i'), \\ \boldsymbol{G}_4 &= \frac{1}{n}\left(\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta};\boldsymbol{y}) \\ &= -\frac{1}{n\phi}\sum_{i=1}^{n}\{a_{i4}c_{i1}^4 + 6a_{i3}c_{i1}^2c_{i2} + 3a_{i2}c_{i2}^2 + 4a_{i2}c_{i1}c_{i3} \\ &\qquad\qquad - (y_i - a_{i1})c_{i4}\}(\boldsymbol{x}_i\boldsymbol{x}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i').\end{aligned}$$

Let us define $\boldsymbol{Z}_j = \sqrt{n}(\boldsymbol{G}_j - \boldsymbol{M}_j)$, $j = 2, 3, 4$, where $\boldsymbol{M}_j = \mathrm{E}[\boldsymbol{G}_j]$, the explicit forms of which are

$$\begin{aligned}\boldsymbol{M}_2 &= -\frac{1}{n\phi}\sum_{i=1}^{n}a_{i2}c_{i1}^2\boldsymbol{x}_i\boldsymbol{x}_i', \\ \boldsymbol{M}_3 &= -\frac{1}{n\phi}\sum_{i=1}^{n}(a_{i3}c_{i1}^3 + 3a_{i2}c_{i1}c_{i2})(\boldsymbol{x}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i'), \\ \boldsymbol{M}_4 &= -\frac{1}{n\phi}\sum_{i=1}^{n}(a_{i4}c_{i1}^4 + 6a_{i3}c_{i1}^2c_{i2} + 3a_{i2}c_{i2}^2 + 4a_{i2}c_{i1}c_{i3})(\boldsymbol{x}_i\boldsymbol{x}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i').\end{aligned}$$

Thus, $\boldsymbol{Z}_j$, $j = 2, 3, 4$ can be expressed as

$$\begin{aligned}
\boldsymbol{Z}_2 &= \frac{1}{\sqrt{n}\phi} \sum_{i=1}^{n} (y_i - a_{i1}) c_{i2} \boldsymbol{x}_i \boldsymbol{x}_i', \\
\boldsymbol{Z}_3 &= \frac{1}{\sqrt{n}\phi} \sum_{i=1}^{n} (y_i - a_{i1}) c_{i3} (\boldsymbol{x}_i' \otimes \boldsymbol{x}_i \boldsymbol{x}_i'), \\
\boldsymbol{Z}_4 &= \frac{1}{\sqrt{n}\phi} \sum_{i=1}^{n} (y_i - a_{i1}) c_{i4} (\boldsymbol{x}_i \boldsymbol{x}_i' \otimes \boldsymbol{x}_i \boldsymbol{x}_i').
\end{aligned}$$

Based on the regularity assumptions, non-singularity of $\boldsymbol{M}_2$ is guaranteed. Furthermore, the regularity assumptions and conditions (A5), (A6) and (A7) ensure the asymptotic normality of $\boldsymbol{Z}_j$. Hence, we can rewrite (2.1.3) as

$$\begin{aligned}
\boldsymbol{0}_p = \; & \frac{1}{\sqrt{n}} (\boldsymbol{g} + \boldsymbol{M}_2 \boldsymbol{b}_1) + \frac{1}{n} \left\{ \boldsymbol{M}_2 \boldsymbol{b}_2 + \frac{1}{2} \boldsymbol{M}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) + \boldsymbol{Z}_2 \boldsymbol{b}_1 \right\} \\
& + \frac{1}{n\sqrt{n}} \left\{ \boldsymbol{M}_2 \boldsymbol{b}_3 + \frac{1}{2} \boldsymbol{M}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) \right. \\
& \qquad \left. + \frac{1}{6} (\boldsymbol{I}_p \otimes \boldsymbol{b}_1') \boldsymbol{M}_4 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) + \boldsymbol{Z}_2 \boldsymbol{b}_2 + \frac{1}{2} \boldsymbol{Z}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \right\} \\
& + O_p(n^{-2}).
\end{aligned} \tag{2.1.4}$$

Comparing the terms of the same order in both sides of (2.1.4), the explicit forms of $\boldsymbol{b}_1, \boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are obtained as follows:

$$\begin{aligned}
\boldsymbol{b}_1 &= -\boldsymbol{M}_2^{-1} \boldsymbol{g}, \\
\boldsymbol{b}_2 &= -\boldsymbol{M}_2^{-1} \left\{ \frac{1}{2} \boldsymbol{M}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) + \boldsymbol{Z}_2 \boldsymbol{b}_1 \right\}, \\
\boldsymbol{b}_3 &= -\boldsymbol{M}_2^{-1} \left\{ \frac{1}{2} \boldsymbol{M}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2 \otimes \boldsymbol{b}_1) + \frac{1}{6} (\boldsymbol{I}_p \otimes \boldsymbol{b}_1)' \boldsymbol{M}_4 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \right. \\
& \qquad \left. + \boldsymbol{Z}_2 \boldsymbol{b}_2 + \frac{1}{2} \boldsymbol{Z}_3 (\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \right\}.
\end{aligned}$$

2.2 Bias Correction of the AIC

The goodness of fit of the model is measured by the risk function based on the expected KL information, as follows:

$$Risk = \mathrm{E}_{\boldsymbol{y}}\mathrm{E}_{\boldsymbol{y}^*}[-2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y}^*)],$$

where $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$ is an n-dimensional random vector that is independent of $\boldsymbol{y}$ and has the same distribution as $\boldsymbol{y}$. At the beginning of this section, we derive the bias of $-2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y})$ to *Risk*. Under ordinary circumstances, calculation of the expectations of $\boldsymbol{y}$ under the specific distribution are needed in order to express the bias. However, based on the characteristics of the exponential family, we can obtain the bias without calculating the expectations of $\boldsymbol{y}$ under the specific distribution. The explicit form of the bias can be expressed by several derivatives of the log-likelihood function.

The bias when we estimate *Risk* by $-2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y})$ is given as

$$\begin{aligned} B &= Risk - \mathrm{E}_{\boldsymbol{y}}[-2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y})] \\ &= \mathrm{E}_{\boldsymbol{y}}\mathrm{E}_{\boldsymbol{y}^*}[2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y}) - 2\ell(\hat{\boldsymbol{\beta}};\boldsymbol{y}^*)] \\ &= \frac{2}{\phi}\sum_{i=1}^{n}\mathrm{E}_{\boldsymbol{y}}[(y_i - a_{i1})\hat{\theta}_i]. \end{aligned} \tag{2.2.1}$$

By applying a Taylor expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ to $\hat{\theta}_i = (h \circ \mu)^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$, $\hat{\theta}_i$ is expanded as

$$\begin{aligned} \hat{\theta}_i = \theta_i &+ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\frac{\partial^2\theta_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &+ \frac{1}{6}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\left\{\left(\frac{\partial}{\partial\boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right)\theta_i\right\}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} \\ &+ O_p(n^{-2}). \end{aligned} \tag{2.2.2}$$

Substituting the stochastic expansion of $\hat{\boldsymbol{\beta}}$ in (2.1.2) into (2.2.2) yields the following:

$$\begin{aligned} \hat{\theta}_i = \theta_i &+ \frac{1}{\sqrt{n}}c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_1 + \frac{1}{n}\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_2 + \frac{1}{2}c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)^2\right\} \\ &+ \frac{1}{n\sqrt{n}}\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_3 + c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)(\boldsymbol{x}_i'\boldsymbol{b}_2) + \frac{1}{6}c_{i3}(\boldsymbol{x}_i'\boldsymbol{b}_1)^3\right\} + O_p(n^{-2}). \end{aligned} \tag{2.2.3}$$

By combining (2.2.1) and (2.2.3), we obtain

$$
\begin{aligned}
B = {} & \frac{2}{\phi}\sum_{i=1}^{n}\mathrm{E}[(y_i - a_{i1})\theta_i] + \frac{2}{\sqrt{n}\phi}\sum_{i=1}^{n}\mathrm{E}[(y_i - a_{i1})c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_1] \\
& + \frac{2}{n\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_2 + \frac{1}{2}c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)^2\right\}\right] \\
& + \frac{2}{n\sqrt{n}\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_3 + c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)(\boldsymbol{x}_i'\boldsymbol{b}_2) + \frac{1}{6}c_{i3}(\boldsymbol{x}_i'\boldsymbol{b}_1)^3\right\}\right] \\
& + O(n^{-2}).
\end{aligned}
\tag{2.2.4}
$$

Recall that $a_{i1} = \partial b(\theta_i)/\partial\theta_i = \mathrm{E}[y_i]$. This yields the first term of (2.2.4), as follows:

$$
\frac{2}{\phi}\sum_{i=1}^{n}\mathrm{E}[(y_i - a_{i1})\theta_i] = 0. \tag{2.2.5}
$$

Since $\mathrm{E}[\boldsymbol{g}\boldsymbol{g}'] = -\boldsymbol{M}_2$, the second term of (2.2.4) can be calculated as

$$
\frac{2}{\sqrt{n}\phi}\sum_{i=1}^{n}\mathrm{E}[(y_i - a_{i1})c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_1] = -2\mathrm{E}[\boldsymbol{g}'\boldsymbol{M}_2^{-1}\boldsymbol{g}] = 2p. \tag{2.2.6}
$$

The third term of (2.2.4) can be obtained as

$$
\begin{aligned}
& \frac{2}{n\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_2 + \frac{1}{2}c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)^2\right\}\right] \\
= {} & \frac{3}{n^2\phi}\sum_{i=1}^{n}a_{i3}c_{i1}^2c_{i2}u_{ii}^2 + \frac{1}{n^3\phi^2}\sum_{i,j}^{n}a_{i3}c_{i1}^3(a_{j3}c_{j1}^3 + 3a_{j2}c_{j1}c_{j2})u_{ij}^3 \\
& + O(n^{-2}),
\end{aligned}
\tag{2.2.7}
$$

where $\sum_{i,j}^{n}$ refers to $\sum_{i=1}^{n}\sum_{j=1}^{n}$, and u_{ij} is the (i,j)th element of the matrix $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{M}_2^{-1}\boldsymbol{X}'$, i.e.,

$$
u_{ij} = \boldsymbol{x}_i'\boldsymbol{M}_2^{-1}\boldsymbol{x}_j. \tag{2.2.8}
$$

Note that coefficient u_{ij} is determined by both the link function and the distribution of the model. The derivation of (2.2.7) is shown in Appendix A.1. Furthermore, the fourth term of (2.2.4) can be expanded as

$$\begin{aligned}
&\frac{2}{n\sqrt{n}\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i-a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_3+c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)(\boldsymbol{x}_i'\boldsymbol{b}_2)+\frac{1}{6}c_{i3}(\boldsymbol{x}_i'\boldsymbol{b}_1)^3\right\}\right]\\
&=-\frac{1}{n^2\phi}\sum_{i=1}^{n}(a_{i4}c_{i1}^4+6a_{i3}c_{i1}^2c_{i2}-a_{i2}c_{i2}^2)u_{ii}^2\\
&\quad-\frac{2}{n^3\phi^2}\sum_{i,j}^{n}\{(a_{i3}c_{i1}^3)(a_{j3}c_{j1}^3+3a_{j2}c_{j1}c_{j2})+2(a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}c_{j2})\}u_{ij}^3\\
&\quad-\frac{1}{n^3\phi^2}\sum_{i,j}^{n}\{(a_{i3}c_{i1}^3)(a_{j3}c_{j1}^3+3a_{j2}c_{j1}c_{j2})+4(a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}c_{j2})\}u_{ii}u_{ij}u_{jj}\\
&\quad+O(n^{-2}).
\end{aligned} \tag{2.2.9}$$

The detailed derivation of (2.2.9) is given in Appendix A.2.

Finally, by substituting (2.2.5), (2.2.6), (2.2.7), and (2.2.9) into (2.2.4), we obtain the asymptotic expansion of B up to order n^{-1} as

$$B=2p+\frac{1}{n}(w_1+w_2)+O(n^{-2}), \tag{2.2.10}$$

where

$$\begin{aligned}
w_1&=-\frac{1}{n\phi}\sum_{i=1}^{n}(a_{i4}c_{i1}^4+3a_{i3}c_{i1}^2c_{i2}-a_{i2}c_{i2}^2)u_{ii}^2,\\
w_2&=-\frac{1}{n^2\phi^2}\sum_{i,j}^{n}\big\{a_{i3}c_{i1}^3(a_{j3}c_{j1}^3+3a_{j2}c_{j1}c_{j2})\\
&\qquad\qquad+4(a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}c_{j2})\big\}(u_{ij}^3+u_{ii}u_{ij}u_{jj}).
\end{aligned} \tag{2.2.11}$$

By a simple calculation, we have $c_{i1}=1$ and $c_{i2}=0$ when the link function is natural. Thus, if the model has the natural link function, w_1

and w_2 became simple, as follows:

$$
\begin{aligned}
w_1 &= -\frac{1}{n\phi}\sum_{i=1}^{n} a_{i4}u_{ii}^2,\\
w_2 &= -\frac{1}{n^2\phi^2}\sum_{i,j}^{n} a_{i3}a_{j3}(u_{ij}^3 + u_{ii}u_{ij}u_{jj}).
\end{aligned}
$$

Equation (2.2.10) yields the following formula for the CAIC:

$$\mathrm{CAIC} = \mathrm{AIC} + \frac{1}{n}(\hat{w}_1 + \hat{w}_2),$$

where $\hat{w}_1$ and $\hat{w}_2$ are defined by replacing $\boldsymbol{\beta}$ in w_1 and w_2 with $\hat{\boldsymbol{\beta}}$. On the other hand, if h is not the natural link function, we have to use w_1 and w_2 in (2.2.11). Note that $\hat{w}_1$ and $\hat{w}_2$ depend only on several derivatives. Therefore, we can comfortably obtain coefficients $\hat{w}_1$ and $\hat{w}_2$ using formula manipulation software.

2.3 Numerical Studies

In this section, we conduct numerical studies to show that the CAIC is better than the original AIC. At the beginning of this section, we examine the numerical studies for the frequencies of the model and the prediction error of the best models selected by the criteria. We prepared the eight candidate models $M_1, \ldots, M_8$ with $n = 50$ and 100. First, we constructed an $n \times 8$ explanatory variable matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. The first column of $\boldsymbol{X}$ is $\mathbf{1}_n$, where $\mathbf{1}_n$ is an n-dimensional vector of ones, and the remaining seven columns of $\boldsymbol{X}$ were defined by realizations of independent dummy variables with binomial distribution $B(1, 0.4)$. In this simulation, we prepared two parameters $\boldsymbol{\beta}$, as follows:

$$\text{Case 1} : \boldsymbol{\beta} = (0.7, -0.7)', \quad \text{Case 2} : \boldsymbol{\beta} = (0.1, 0.1, 0.3, -0.8)'.$$

We assume the following nested setting in our numerical studies in order to simplify our simulation result. The explanatory variables matrix in the jth model M_j consists of the first j columns of $\boldsymbol{X}$, $j = 1, \ldots, 8$. Thus, in Case 1, the true model is the second model, and in Case 2, the true model is the fourth model. We simulated 2,000 realizations of $\boldsymbol{y} = (y_1, \ldots, y_n)'$

in the probit regression model, i.e., $y_i \overset{\text{Indep}}{\sim} B(1, p_i)$, where $p_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$, $i = 1, \ldots, n$.

Tables 2.1 and 2.2 list the following properties.

(1) Selection-probability (freq.): the frequency of the model chosen by minimizing the information criterion.

(2) Mean of the information criterion (mean): E[AIC] and E[CAIC], which is estimated as the average of the AIC and CAIC, respectively.

(3) Prediction error of the best model ($\mathrm{PE_B}$): the risk function of the model selected by the information criterion as the best model, which is defined as

$$\begin{aligned}\mathrm{PE_B} &= \frac{1}{2000}\sum_{k=1}^{2000} \mathrm{E}_{\boldsymbol{y}^*}[-2\ell(\hat{\boldsymbol{\beta}}_{\mathrm{B}_k}; \boldsymbol{y}^*)] \\ &= \frac{1}{2000}\sum_{k=1}^{2000}\sum_{i=1}^{n} -2\{p_i \log \hat{p}_i^{(k)} + (1-p_i)(1-\log \hat{p}_i^{(k)})\},\end{aligned}$$

where $\boldsymbol{y}^*$ is a future observation, $\hat{\boldsymbol{\beta}}_{\mathrm{B}_k}$ is the value of $\hat{\boldsymbol{\beta}}$ of the selected model at the kth iteration, and $\hat{p}_i^{(k)} = \Phi(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{B}_k})$.

The difference between the risk function and mean value of the information criterion should be small since the information criterion is an estimator of the risk function. The $\mathrm{PE_B}$ is an important property because it is equivalent to the expected KL information between the true model and the best model selected by the criteria.

From Tables 2.1 and 2.2, the model having the smallest risk (referred to as the principle best model) coincides with the true model in all situations. We can see that the selection-probabilities and prediction errors of the CAIC were improved in all situations in comparison with the AIC.

We simulated several other models and obtained similar results. Furthermore, the mean value of the CAIC is an improved estimator of the risk function when the differences between the risk function and mean value of the AIC is non-negligible. We can see the above result from Figures 2.1 and 2.2, which plot the risk function and mean value of the AIC and CAIC for Case 1 and 2, respectively.

Next, for the purpose of analyzing the GLM, we consider the data reported in Brown (1980), who discussed an experiment in which 53 prostate cancer patients underwent surgery to examine their lymph nodes for evidence of cancer. The response variable is the number of patients with nodal involvement, and there were five predictor variables: X Ray, Stage, Age, Acid, and Grade. We prepare all combinations of five variables as candidate models, i.e., $2^5 = 32$ models. First, we assume that the response variable y_i is distributed according to $B(1, p_i)$, $i = 1, \ldots, n$. For the link function, we prepare two functions: the logistic link function and the probit link function. In this analysis, we select the link functions and variables simultaneously. Table 2.3 shows the selection-probability of the model selected by minimizing the information criterion and the estimated prediction error of the best model selected by the information criterion. Note that we list only the selected models by the AIC or CAIC in Table 2.3.

Table 2.1: Selection-probability, mean value of AIC and CAIC, and prediction error in Case 1

n	Model		1	2	3	4	5	6	7	8	PE_{B}
50	risk function		64.67	**62.48**	63.77	65.08	66.71	68.54	70.58	72.89	-
	AIC	mean	65.03	**62.73**	63.66	64.60	65.46	66.27	67.10	67.84	66.49
		freq.	28.80	**44.45**	9.00	4.95	3.95	3.25	2.75	2.85	
	CAIC	mean	65.08	**62.96**	64.21	65.54	66.94	68.41	70.04	71.72	65.03
		freq.	33.20	**49.25**	8.10	3.90	2.60	1.50	0.95	0.50	
100	risk function		128.80	**123.11**	124.24	125.39	126.63	127.94	129.30	130.71	-
	AIC	mean	129.12	**123.44**	124.40	125.36	126.28	127.17	128.07	128.96	125.48
		freq.	9.55	**61.10**	11.55	5.40	4.25	3.45	2.20	2.50	
	CAIC	mean	129.15	**123.55**	124.65	125.80	126.94	128.11	129.36	130.64	125.01
		freq.	9.30	**62.35**	11.65	5.35	4.10	3.40	1.75	2.10	

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

Table 2.2: Selection-probability, mean value of AIC and CAIC, and prediction error in Case 2

n	Model		1	2	3	4	5	6	7	8	PE_{B}
50	risk function		70.03	70.90	71.09	**68.16**	69.72	71.54	73.39	75.56	-
	AIC	mean	70.15	71.18	71.28	**67.40**	68.26	69.02	69.87	70.64	72.21
		freq.	26.25	2.80	3.90	**40.40**	9.85	6.80	5.00	5.00	
	CAIC	mean	70.19	71.37	71.72	**68.28**	69.63	71.02	72.59	74.28	71.21
		freq.	33.65	3.25	4.40	**43.50**	7.70	4.25	2.10	1.15	
100	risk function		139.40	140.32	138.74	**131.55**	132.80	134.03	135.36	136.72	-
	AIC	mean	139.65	140.69	139.15	**131.32**	132.21	133.14	134.03	134.93	134.35
		freq.	7.55	0.45	1.85	**62.10**	12.85	6.25	4.85	4.10	
	CAIC	mean	139.67	140.78	139.36	**131.74**	132.84	134.03	135.24	135.52	134.08
		freq.	8.90	0.60	2.30	**65.60**	11.85	5.75	3.10	1.90	

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

Figure 2.1: Risk and average value of AIC and CAIC in Case 1

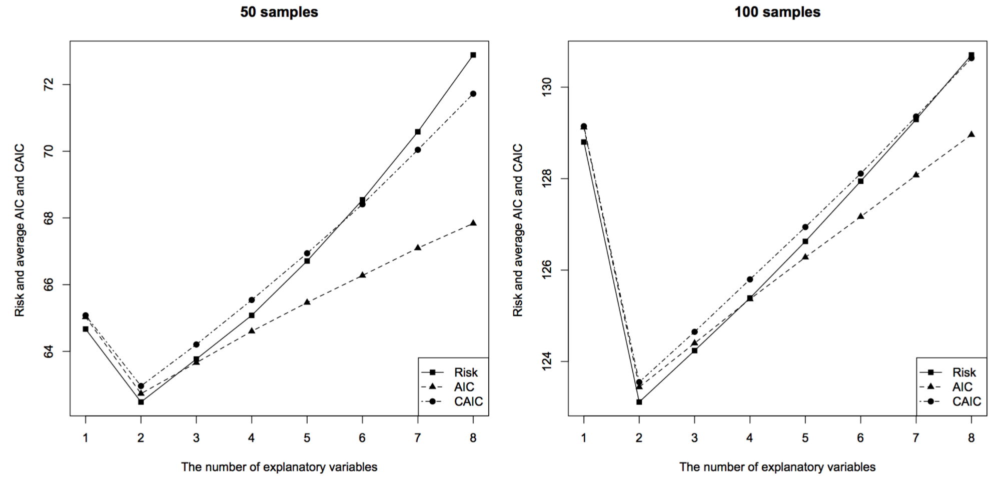

Figure 2.2: Risk and average value of AIC and CAIC in Case 2

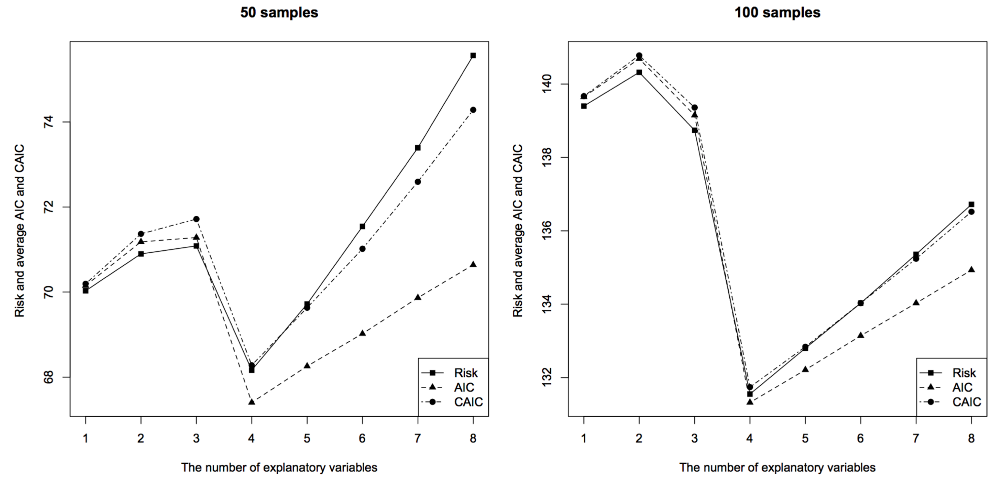

Table 2.3: Selection-probability and estimated prediction error

Selected model	AIC			CAIC		
	Logistic	Probit	Total	Logistic	Probit	Total
X Ray	0	0	0	0	5	5
X Ray, Acid	0	3	3	0	3	3
X Ray, Age	0	1	1	0	1	1
X Ray, Grade	0	1	1	1	3	4
X Ray, Stage	25	5	30	24	38	62
X Ray, Grade, Acid	0	4	4	0	1	1
X Ray, Stage, Acid	5	22	27	0	15	15
X Ray, Stage, Age	0	4	4	0	6	6
X Ray, Stage, Grade	0	2	2	0	2	2
X Ray, Grade, Age, Acid	2	0	2	1	0	1
X Ray, Stage, Age, Acid	0	17	17	0	0	0
X Ray, Stage, Grade, Acid	1	6	7	0	0	0
X Ray, Stage, Grade, Age	0	1	1	0	0	0
X Ray, Stage, Grade, Age, Acid	0	1	1	0	0	0
$\widehat{\mathrm{PE}_{\mathrm{B}}}$	62.97			62.70		

We divide the data into calibration sample data and validation sample data. The sample sizes of the calibration sample and the validation sample were 43 and 10, respectively. The best subset of the variables and the link function were selected by criteria derived from the calibration sample. The selection-probabilities were evaluated from only the calibration sample. The prediction errors were estimated as follows. Let $\boldsymbol{d}_j = (d_{1j}, \dots, d_{nj})'$ be an n-dimensional vector expressing a pattern to leave out 10 data at the jth iteration, $j = 1, \dots, 100$, i.e., $d_{ij} = 1$ or 0 and $\sum_{i=1}^{n} d_{ij} = 10$, which are generated by using "sample" of the "R" software. Moreover, we let $\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_j]}$ denote $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_j]}$ of $\boldsymbol{\beta}$ of the best model evaluated from the calibration sample, where $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_j]}$ is given as

$$\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_j]} = \underset{\boldsymbol{\beta}}{\mathrm{argmax}} \sum_{i=1}^{53} (1 - d_{ij}) \log f(y_i; \boldsymbol{\beta}).$$

Finally, the estimated PE_{B} is given as

$$\widehat{\mathrm{PE}}_{\mathrm{B}} = \frac{43}{100} \sum_{j=1}^{100} \frac{1}{10} \sum_{i=1}^{53} d_{ij} \{-2 \log f(y_i; \hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_j]})\}.$$

Table 2.3 indicates that the models selected by the AIC were spread over a wider area than those of the CAIC, although the model most selected by the AIC is the same as that selected by the CAIC. In particular, the selection probability of the model most selected by the CAIC is much higher than that selected by the AIC. The estimated prediction error of the CAIC was smaller than that of the AIC. Thus, the CAIC is thought to have improved the accuracy of the original AIC.

Consequently, from Tables 2.1, 2.2, and 2.3, we recommend the use of the CAIC rather than the AIC for selecting variables in the GLMs.

Chapter 3

Bias-Corrected AIC for Selecting Variables in Multinomial Logistic Regression Models

Chapter 3 is organized as follows. In Section 3.1, we give a stochastic expansion of the MLE. In Section 3.2, the CAIC in the multinomial logistic regression models is proposed. In Section 3.3, we verify that the proposed CAIC has better performance than the AIC by conducting numerical experiments. Technical details are provided in Appendix A.3 and Appendix A.4.

3.1 Stochastic Expansion of MLE

Suppose that the data consists of a sequence $\{(\boldsymbol{y}_i, \boldsymbol{x}_i); i = 1, \ldots, n\}$, 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

$$p_{i0}(\boldsymbol{\beta}) = \frac{1}{1+\sum_{j=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}, \quad \boldsymbol{p}_i(\boldsymbol{\beta}) = (p_{i1}(\boldsymbol{\beta}), \dots, p_{ir}(\boldsymbol{\beta}))', \quad (3.1.1)$$

where

$$p_{ij}(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)}{1+\sum_{k=1}^r \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}, \quad j = 1, \dots, r.$$

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 (3.1.1) is expressed as

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

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

(B1) : $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B}$ is a convex and open set in $\mathbb{R}^k$,

(B2) : $(\boldsymbol{I}_r \otimes \boldsymbol{x}_i)'\boldsymbol{\beta} \in \Theta^0$, $i = 1, 2, \dots,$ for all $\boldsymbol{\beta} \in \mathcal{B}$, where Θ^0 is the interior of the convex natural parameter space $\Theta \subset \mathbb{R}^r$,

(B3) : $^\exists m_0$ *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 (B1) guarantees the uniqueness of the MLE if it exists. Condition (B2) is necessary to obtain the multinomial logistic regression model for all $\boldsymbol{\beta}$. Condition (B3) 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

$$\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \mathrm{diag}\{\boldsymbol{p}_i(\boldsymbol{\beta})\} - \boldsymbol{p}_i(\boldsymbol{\beta})\boldsymbol{p}_i(\boldsymbol{\beta})'. \tag{3.1.2}$$

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 & Kaufmann (1985):

(B4) : sequence $\{\boldsymbol{x}_i\}$ lies in $\mathcal{F}$ with $(\boldsymbol{I}_r \otimes \boldsymbol{x}_i)'\boldsymbol{\beta} \in \Theta^0$, $\boldsymbol{\beta} \in \mathcal{B}$,

(B5) : $\liminf_{m\to\infty} \lambda(\sum_{i=1}^{m} n_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i'/n) > 0$, where $\lambda(\boldsymbol{A})$ indicates the smallest eigenvalue of symmetric matrix $\boldsymbol{A}$.

According to Corollary 1 in Fahrmeir & Kaufmann (1985), $\hat{\boldsymbol{\beta}}$ has weak consistency and asymptotic normality under these conditions. Furthermore, from (B5), $\sum_{i=1}^{m} n_i \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' = O(n)$ is satisfied. Under the assumption that all conditions are satisfied, $\hat{\boldsymbol{\beta}}$ can be formally expanded as follows:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{1}{\sqrt{n}}\boldsymbol{b}_1 + \frac{1}{n}\boldsymbol{b}_2 + \frac{1}{n\sqrt{n}}\boldsymbol{b}_3 + O_p(n^{-2}), \tag{3.1.3}$$

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.,

$$\left.\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{m} [(\boldsymbol{y}_i \otimes \boldsymbol{x}_i) - n_i\{\boldsymbol{p}_i(\hat{\boldsymbol{\beta}}) \otimes \boldsymbol{x}_i\}] = \boldsymbol{0}_{kr}, \tag{3.1.4}$$

where $\boldsymbol{0}_{kr}$ is a kr-dimensional vector of zeros. To expand equation (3.1.4), we prepare the following three matrices consisting of the second, third, and fourth derivatives of $-\ell(\boldsymbol{\beta})/n$:

$$\boldsymbol{G}_2(\boldsymbol{\beta}) = -\frac{1}{n}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}, \quad \boldsymbol{G}_3(\boldsymbol{\beta}) = -\frac{1}{n}\left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta}),$$
$$\boldsymbol{G}_4(\boldsymbol{\beta}) = -\frac{1}{n}\left(\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)\ell(\boldsymbol{\beta}).$$

The result of the first derivative of $\ell(\boldsymbol{\beta})$ in (3.1.4) 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.3):

$$\boldsymbol{G}_2(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \frac{\partial \boldsymbol{p}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \{ \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \}, \tag{3.1.5}$$

$$\boldsymbol{G}_3(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \left(\frac{\partial}{\partial \boldsymbol{\beta}'} \otimes \frac{\partial}{\partial \boldsymbol{\beta}'} \right) \boldsymbol{p}_i(\boldsymbol{\beta}) \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \{ \boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \}, \tag{3.1.6}$$

$$\boldsymbol{G}_4(\boldsymbol{\beta}) = \sum_{i=1}^{m} \rho_i \left\{ \left(\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \otimes \frac{\partial}{\partial \boldsymbol{\beta}'} \right) \boldsymbol{p}_i(\boldsymbol{\beta}) \right\} \otimes \boldsymbol{x}_i = \sum_{i=1}^{m} \rho_i \{ \boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \}, \tag{3.1.7}$$

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

$$\begin{aligned}
\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) &= \sum_{a=1}^{r} p_{ia}(\boldsymbol{\beta}) \boldsymbol{e}_a' \otimes \boldsymbol{x}_i' \otimes \boldsymbol{q}_{i,a}(\boldsymbol{\beta}) \boldsymbol{q}_{i,a}(\boldsymbol{\beta})' - \boldsymbol{p}_i(\boldsymbol{\beta})' \otimes \boldsymbol{x}_i' \otimes \boldsymbol{\Sigma}_i(\boldsymbol{\beta}), \\
\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) &= \sum_{a=1}^{r} p_{ia}(\boldsymbol{\beta}) \boldsymbol{q}_{i,a}(\boldsymbol{\beta}) \boldsymbol{q}_{i,a}(\boldsymbol{\beta})' \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes \{ \boldsymbol{q}_{i,a}(\boldsymbol{\beta}) \boldsymbol{q}_{i,a}(\boldsymbol{\beta})' - \boldsymbol{p}_i(\boldsymbol{\beta}) \boldsymbol{p}_i(\boldsymbol{\beta})' \} \\
&\quad - \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes \{ \boldsymbol{\Sigma}_i(\boldsymbol{\beta}) - \boldsymbol{p}_i(\boldsymbol{\beta}) \boldsymbol{p}_i(\boldsymbol{\beta})' \} \\
&\quad - \sum_{a,b}^{r} p_{ia}(\boldsymbol{\beta}) p_{ib}(\boldsymbol{\beta}) \boldsymbol{q}_{i,a}(\boldsymbol{\beta}) \boldsymbol{q}_{i,b}(\boldsymbol{\beta})' \otimes \boldsymbol{x}_i \boldsymbol{x}_i' \otimes (\boldsymbol{e}_a \boldsymbol{e}_b' + \boldsymbol{e}_b \boldsymbol{e}_a').
\end{aligned} \tag{3.1.8}$$

Here, $\boldsymbol{e}_j$ is an r-dimensional jth coordinate unit vector whose jth element is 1 and others are 0, $\boldsymbol{q}_{i,a}(\boldsymbol{\beta}) = \boldsymbol{e}_a - \boldsymbol{p}_i(\boldsymbol{\beta})$, 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 (3.1.4) yields

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{m} [\{ \boldsymbol{y}_i - n_i \boldsymbol{p}_i(\boldsymbol{\beta}) \} \otimes \boldsymbol{x}_i] \\
&= \boldsymbol{G}_2(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{2} \boldsymbol{G}_3(\boldsymbol{\beta}) \{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \} \\
&\quad + \frac{1}{6} \{ \boldsymbol{I}_{kr} \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \} \boldsymbol{G}_4(\boldsymbol{\beta}) \{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \} + O_p(n^{-2}).
\end{aligned} \tag{3.1.9}$$

Notice that the order of the left-hand side of equation (3.1.9) 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 (3.1.3) into (3.1.9), $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ in (3.1.3) are specified as

$$\begin{aligned}
\boldsymbol{b}_1 &= \frac{1}{\sqrt{n}}\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\sum_{i=1}^{m}[\{\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})\}\otimes\boldsymbol{x}_i],\\
\boldsymbol{b}_2 &= -\frac{1}{2}\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\otimes\boldsymbol{b}_1),\\
\boldsymbol{b}_3 &= -\frac{1}{2}\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\Big\{\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\otimes\boldsymbol{b}_2+\boldsymbol{b}_2\otimes\boldsymbol{b}_1)\\
&\qquad\qquad + \frac{1}{3}(\boldsymbol{I}_{kr}\otimes\boldsymbol{b}_1')\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\otimes\boldsymbol{b}_1)\Big\}.
\end{aligned} \tag{3.1.10}$$

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. The stochastic expansion is regarded as a special case of the general stochastic expansion of MLE, e.g., in McCullagh & Cox (1986).

3.2 Main Result

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

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \mathrm{E}[-2\ell(\boldsymbol{\beta})]\\
&= -2\sum_{i=1}^{m} n_i\left[(\boldsymbol{p}_i^*\otimes\boldsymbol{x}_i)'\boldsymbol{\beta} - \log\left\{1+\sum_{j=1}^{r}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\right\}\right],
\end{aligned} \tag{3.2.1}$$

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

$$Risk = \mathrm{E}[\mathcal{L}(\hat{\boldsymbol{\beta}})]. \tag{3.2.2}$$

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

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2kr. \tag{3.2.3}$$

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

$$\begin{aligned} B &= Risk - \mathrm{E}[-2\ell(\hat{\boldsymbol{\beta}})] \\ &= 2\sum_{i=1}^{m} \mathrm{E}[[\{\boldsymbol{y}_i - n_i\boldsymbol{p}_i(\boldsymbol{\beta})\} \otimes \boldsymbol{x}_i]'\hat{\boldsymbol{\beta}}] \\ &= 2\sqrt{n}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\hat{\boldsymbol{\beta}}] \\ &= 2\sqrt{n}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{\beta}] + 2\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1] + \frac{2}{\sqrt{n}}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2] \\ &\quad + \frac{2}{n}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_3] + O(n^{-2}), \end{aligned} \tag{3.2.4}$$

where matrices $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are given by (3.1.5), (3.1.6), and (3.1.7), respectively, and kr-dimensional random vectors $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and $\boldsymbol{b}_3$ are given by (3.1.10). 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 (3.2.4) 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 (3.2.4) is $O(n^{-2})$.

From elementary linear algebra and the definition of $\boldsymbol{b}_2$ in (3.1.10), $\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2$ in (3.2.4) is expressed by the function of $\boldsymbol{b}_1$ as

$$\begin{aligned} \boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2 &= -\frac{1}{2}\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1) \\ &= -\frac{1}{2}\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}. \end{aligned} \tag{3.2.5}$$

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

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

It follows from the above equations and the definition of $\boldsymbol{b}_2$ in (3.1.10) that

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

Thus, from the above result and the definition of $\boldsymbol{b}_3$ in (3.1.10), $\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_3$ in (3.2.4) is expressed by the function of $\boldsymbol{b}_1$ as

$$\begin{aligned}\boldsymbol{b}_1'\boldsymbol{G}_2\boldsymbol{b}_3 &= -\frac{1}{2}\boldsymbol{b}_1'\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\otimes\boldsymbol{b}_2+\boldsymbol{b}_2\otimes\boldsymbol{b}_1)\\ &\quad -\frac{1}{6}\boldsymbol{b}_1'(\boldsymbol{I}_{kr}\otimes\boldsymbol{b}_1')\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\otimes\boldsymbol{b}_1)\\ &= \frac{1}{2}\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1'\otimes\boldsymbol{b}_1\boldsymbol{b}_1')\}\\ &\quad -\frac{1}{6}\mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1'\otimes\boldsymbol{b}_1\boldsymbol{b}_1')\}.\end{aligned}\tag{3.2.6}$$

Hence, equations (3.2.5) and (3.2.6) indicate that the expansion of B in (3.2.4) 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 (3.2.4) by centralized moments of $\boldsymbol{y}_1,\dots,\boldsymbol{y}_m$. 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 $\mathrm{E}[\boldsymbol{b}_1] = \boldsymbol{0}_{kr}$ because $\mathrm{E}[\boldsymbol{y}_i] = n_i\boldsymbol{p}_i(\boldsymbol{\beta})$. From the result of the first derivative of $\ell(\boldsymbol{\beta})$ in (3.1.4) and the definition of $\boldsymbol{b}_1$ in (3.1.10), we can see that

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

Notice that $\mathrm{E}[\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}] = \mathbf{0}_{kr}$ holds and $\boldsymbol{G}_2(\boldsymbol{\beta})$, $\boldsymbol{G}_3(\boldsymbol{\beta})$, and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are constant matrices. By applying general formulas of expectations (A.4.5) in Appendix A.4 to the case of the multinomial logistic regression model, the following equations are obtained:

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

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., $\mathrm{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1', \dots, \boldsymbol{a}_m')'$ when $\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_m)$ (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

$$\mathrm{E}[\boldsymbol{b}_1\boldsymbol{b}_1'] = \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}, \tag{3.2.7}$$

$$\mathrm{E}[\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'] = \frac{1}{\sqrt{n}}\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}, \tag{3.2.8}$$

$$\begin{aligned}\mathrm{E}[\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1'] &= (\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}\\
&\qquad + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})' + O(n^{-1}).\end{aligned} \tag{3.2.9}$$

The result in (3.2.7) implies that

$$\begin{aligned}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1] &= \mathrm{E}[\mathrm{tr}\{\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_1\boldsymbol{b}_1'\}]\\
&= \mathrm{tr}\{\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}\\
&= kr.\end{aligned} \tag{3.2.10}$$

Similarly, from (3.2.8) and (3.2.5), we have

$$\begin{aligned}&\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{b}_2]\\
&= -\frac{1}{2}\mathrm{E}[\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})'(\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}]\\
&= -\frac{1}{2\sqrt{n}}\mathrm{tr}[\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}].\end{aligned} \tag{3.2.11}$$

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 (3.2.9), the expectation of the first part in (3.2.6) is given by

$$\begin{aligned}
&\mathrm{E}[\mathrm{tr}\{\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}] \\
&= \mathrm{tr}[\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})(\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}] \\
&\quad + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}) \\
&\quad + O(n^{-1}) \\
&= 2\mathrm{tr}[\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\{\boldsymbol{G}_2^{-1}(\boldsymbol{\beta}) \otimes \boldsymbol{G}_2^{-1}(\boldsymbol{\beta})\}] \\
&\quad + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}) \qquad (3.2.12) \\
&\quad + O(n^{-1}).
\end{aligned}$$

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

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

By using the above relations and equation (3.2.9), the expectation of the second part in (3.2.6) is given by

$$\begin{aligned}
&\mathrm{E}[\mathrm{tr}\{\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{b}_1\boldsymbol{b}_1' \otimes \boldsymbol{b}_1\boldsymbol{b}_1')\}] \\
&= \mathrm{tr}[\boldsymbol{G}_4(\boldsymbol{\beta})(\boldsymbol{I}_{k^2r^2} + \boldsymbol{K}_{kr})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}] \\
&\quad + \mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})\boldsymbol{G}_4(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}) + O(n^{-1}) \\
&= 3\mathrm{tr}[\boldsymbol{G}_4(\boldsymbol{\beta})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}] + O(n^{-1}). \qquad (3.2.13)
\end{aligned}$$

Hence, from equations (3.2.6), (3.2.12), and (3.2.13), we can see that

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2\boldsymbol{b}_3] \\
&= \mathrm{tr}[\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}] \\
&\quad + \frac{1}{2}\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1})'\boldsymbol{G}_3(\boldsymbol{\beta})'\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\boldsymbol{G}_3(\boldsymbol{\beta})\mathrm{vec}(\boldsymbol{G}_2(\boldsymbol{\beta})^{-1}) \qquad (3.2.14) \\
&\quad - \frac{1}{2}\mathrm{tr}[\boldsymbol{G}_4(\boldsymbol{\beta})\{\boldsymbol{G}_2(\boldsymbol{\beta})^{-1} \otimes \boldsymbol{G}_2(\boldsymbol{\beta})^{-1}\}] + O(n^{-1}).
\end{aligned}$$

Consequently, by substituting $\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{G}_2(\boldsymbol{\beta})\boldsymbol{\beta}] = 0$, and equations (3.2.10), (3.2.11), and (3.2.14) into (3.2.4), the bias of $-2\ell(\hat{\boldsymbol{\beta}})$ to *Risk* 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

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

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

$$\mathrm{CAIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2kr + \frac{1}{n}\{\alpha_1(\hat{\boldsymbol{\beta}}) + \alpha_2(\hat{\boldsymbol{\beta}}) - \alpha_3(\hat{\boldsymbol{\beta}})\}. \tag{3.2.15}$$

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:

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

where *Risk* is the risk function given by (3.2.2).

3.3 Numerical Studies

In this section, we conduct numerical studies to show that the CAIC in (3.2.15) works better than the crude AIC in (3.2.3). 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

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

where $\mathcal{L}(\boldsymbol{\beta})$ is the loss function given by (3.2.1) and $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ is the MLE of $\boldsymbol{\beta}$ under the best model.

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, 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:

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

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 3.1 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.

Table 3.1: Selection probability of the model and the prediction error of the best model

			Selection Probability								
Case	m	Criterion	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	$\mathrm{PE_B}$
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.

Chapter 4

Consistent Selection of Working Correlation Structure in Generalized Estimating Equations Analysis Based on Stein's Loss Function

Chapter 4 is organized as follows: In Section 4.1, we introduce the GEE; In Section 4.2, we propose a criterion for selecting the true correlation structure; In Section 4.3, we derive its asymptotic behavior; In Section 4.4, we demonstrate the performance of our criterion in finite samples by presenting a numerical studies.

4.1 Generalized Estimating Equations

In this section, we introduce the GEE approach. For individuals $i = 1, \ldots, n$, we have an m-dimensional response vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$ and an $m \times p$ explanatory variable matrix $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im})'$. We allow the components of $\boldsymbol{y}_i$ to be correlated, but we assume that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent. Furthermore, we do not predetermine the distribution of each $\boldsymbol{y}_i$. In the GEE approach, we assume the marginal density function of y_{ij} to be the GLM, i.e.,

$$f(y_{ij};\theta_{ij},\phi) = \exp\left\{\frac{\theta_{ij}y_{ij} - a(\theta_{ij})}{\phi} + b(y_{ij},\phi)\right\},$$

where $a(\cdot)$ and $b(\cdot)$ are known functions, the unknown parameter θ_{ij} is referred to as the natural location parameter, and ϕ is referred to as the unknown scale parameter. Suppose that $\theta_{ij} \in \Theta^0$, where Θ^0 is the interior of the natural parameter space Θ. In order to use some of the properties of the MLE, we assume regularity assumptions; for details, see Fahrmeir & Kaufmann (1985) and Chapter 2. The linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ is related to $\mu_{ij} = \mathrm{E}[y_{ij}]$ by a link function $h(\cdot)$, i.e., $h(\mu_{ij}) = \eta_{ij}$, where $\boldsymbol{\beta}$ is a p-dimensional unknown regression coefficient. From the properties of the GLM, $\mu_{ij} = \partial a(\theta_{ij})/\partial\theta_{ij}$ and $\mathrm{Var}[y_{ij}] = \phi\partial^2 a(\theta_{ij})/\partial\theta_{ij}^2$. By using a working correlation matrix $\boldsymbol{R}$, the covariance matrix of the ith observation $\boldsymbol{y}_i$ is assumed to be

$$\boldsymbol{V}_i = \phi\boldsymbol{A}_i^{1/2}\boldsymbol{R}\boldsymbol{A}_i^{1/2}, \quad i = 1,\ldots,n, \tag{4.1.1}$$

where $\boldsymbol{A}_i = \mathrm{diag}\{\partial^2 a(\theta_{i1})/\partial\theta_{i1}^2,\ldots,\partial^2 a(\theta_{im})/\partial\theta_{im}^2\}$. Examples of a working correlation structure are

$$\begin{array}{ll} \text{Independent (Indep.)} & : \boldsymbol{R} = \boldsymbol{I}_m, \\ \text{Exchangeable (Ex.)} & : (\boldsymbol{R})_{jk} = \alpha, \\ \mathrm{AR}-1 & : (\boldsymbol{R})_{jk} = \alpha^{|j-k|}, \\ \text{Unstructured (Unst.)} & : (\boldsymbol{R})_{jk} = \alpha_{jk}, \end{array} \tag{4.1.2}$$

where $(\boldsymbol{R})_{jk}$ denotes the (j,k)th element of $\boldsymbol{R}$, and α and α_{jk} are correlation parameters. Note that $\boldsymbol{R}$ is symmetric and its diagonal elements are all ones, since it is a correlation matrix. Using this notation, the GEE is defined as follows.

Definition 4.1.1. *The GEE for $\boldsymbol{\beta}$ with a working correlation matrix $\boldsymbol{R}$ is defined as follows:*

$$\sum_{i=1}^{n}\boldsymbol{D}'_i\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_p, \tag{4.1.3}$$

where $\boldsymbol{D}_i = \boldsymbol{A}_i\boldsymbol{\Delta}_i\boldsymbol{X}_i$, $\boldsymbol{\Delta}_i = \mathrm{diag}\{\partial u(\eta_{i1})/\partial\eta_{i1},\ldots,\partial u(\eta_{im})/\partial\eta_{i1}\}$, $u(\eta_{ij}) = \theta_{ij}$, and $\boldsymbol{V}_i$ was defined in (4.1.1).

Denote $\hat{\boldsymbol{\beta}}(\boldsymbol{R})$ as a GEE estimator with $\boldsymbol{R}$, which is given by solving (4.1.3) with respect to $\boldsymbol{\beta}$. In actual use, unless $\boldsymbol{R}$ is a constant matrix, we need to estimate $\boldsymbol{R}$. Let $\boldsymbol{\alpha}$ be a correlation parameter constructing $\boldsymbol{R}$, i.e., $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\alpha})$. There are several methods for estimating $\boldsymbol{\alpha}$; see Wang & Carey (2003). In Section 4.4, we estimate $\boldsymbol{\alpha}$ by using a moment-based method.

4.2 Selection of Working Correlation Structure

In order to select the true correlation structure, let $\mathcal{M}$ be a set of working correlation structures. For instance, the elements of $\mathcal{M}$ are some particular working correlation structures introduced in (4.1.2). Examples with (4.1.2) are illustrated in Section 4.4. We assume $\mathcal{M}$ to involve at least one correct correlation structure. Let $\boldsymbol{R}_*$ be the true correlation matrix. For theoretical purposes, we divide $\mathcal{M}$ into an over-fitted set $\mathcal{M}^+$ and an under-fitted set $\mathcal{M}^-$, i.e.,

$$\mathcal{M}^+ = \{\boldsymbol{R} \in \mathcal{M} | {}^\exists \boldsymbol{\alpha} \in \mathcal{K} \text{ s.t. } \boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_*\},$$

where $\mathcal{K}$ is the parameter space, which is the compact set and $\mathcal{M}^- = \mathcal{M} \setminus \mathcal{M}^+$. For all $\boldsymbol{R} \in \mathcal{M}^+$, we assume that $\boldsymbol{\alpha} \in \mathcal{K}^0$ where $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_*$ and $\mathcal{K}^0$ is the interior of $\mathcal{K}$. Let the true correlation structure be $\boldsymbol{R}_0$, which has the fewest number of parameters among $\mathcal{M}^+$.

Let $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{A}}_i$, and $\hat{\phi}$ be estimators of $\boldsymbol{\mu}_i$, $\boldsymbol{A}_i$, and ϕ, respectively. For selecting $\boldsymbol{R}_0$ from $\mathcal{M}$, we define the following discrepancy function that is based on Stein's loss function:

$$SL_n(\boldsymbol{R}) = n \log\det(\boldsymbol{R}) + n\mathrm{tr}(\hat{\boldsymbol{R}}_U \boldsymbol{R}^{-1}), \tag{4.2.1}$$

where

$$\begin{aligned} (\hat{\boldsymbol{R}}_U)_{jk} &= \begin{cases} \hat{\phi}^{-1} \sum_{i=1}^n \hat{\varepsilon}_{ij}\hat{\varepsilon}_{ik}/n, & j \neq k, \\ 1, & j = k, \end{cases} \\ \hat{\boldsymbol{\varepsilon}}_i &= (\hat{\varepsilon}_{i1}, \ldots, \hat{\varepsilon}_{im})' = \hat{\boldsymbol{A}}_i^{-1/2}(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i). \end{aligned} \tag{4.2.2}$$

The adequacy of using $SL_n(\cdot)$ for $\boldsymbol{R}_0$ has been determined. It is known that for any correlation matrix $\boldsymbol{R}$,

$$\log\det(\boldsymbol{R}) + \mathrm{tr}(\boldsymbol{R}^{-1}\boldsymbol{R}_*) \geq \log\det(\boldsymbol{R}_*) + m$$

holds with equality if and only if $\boldsymbol{R} = \boldsymbol{R}_*$. Stein's loss function is almost the same as $-2 \times$ the Gaussian log-likelihood. Crowder (1985); Wang & Carey (2003) considered using the Gaussian log-likelihood for estimating the unknown parameter.

Recall that one of our aims is to derive a GIC-type criterion. The GIC is defined as $-2 \times$ the maximum log-likelihood + the number of parameters $\times$ the tuning parameter. By using (4.2.1) instead of the likelihood for $\boldsymbol{y}_i$, a GIC-type criterion is constructed.

Definition 4.2.1. *For a working correlation structure $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\alpha}) \in \mathcal{M}$, the GIC-type criterion is*

$$GIC_{\gamma_n}(\boldsymbol{R}) = SL_n(\hat{\boldsymbol{R}}) + q\gamma_n, \tag{4.2.3}$$

where $\hat{\boldsymbol{R}} = \boldsymbol{R}(\hat{\boldsymbol{\alpha}})$, $\hat{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}$, q is the number of elements in $\boldsymbol{\alpha}$, and γ_n is a tuning parameter.

Note that in the definitions of (4.2.1) and (4.2.3), we have not specified the working correlation structure for estimating the GEE estimator $\hat{\boldsymbol{\beta}}$ or the way in which to estimate $\hat{\phi}$ and $\hat{\boldsymbol{\alpha}}$.

By minimizing the GIC, the best working correlation structure is obtained.

Definition 4.2.2. *The best correlation structure selected by the GIC proposed in (4.2.3) is*

$$\boldsymbol{R}_{best} = \underset{\boldsymbol{R} \in \mathcal{M}}{\operatorname{argmin}} \{GIC_{\gamma_n}(\boldsymbol{R})\}.$$

Note that $\boldsymbol{R}_{best}$ depends on the data as well as the way in which ϕ and $\boldsymbol{\alpha}$ are estimated.

4.3 Properties of Criteria

In this section, we show the consistency of the GIC proposed in (4.2.3). Suppose that the mean structure has been correctly specified. The proof can then be obtained in a way similar to that in Nishii (1984). The following assumptions are sufficient conditions for the consistency of the GIC:

(C1) For all $\boldsymbol{R} \in \mathcal{M}$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(n^{-1/2})$ and $\hat{\phi} - \phi = O_p(n^{-1/2})$.

(C2) $u(\eta_{ij})$ is continuously differentiable.

(C3) For all $\boldsymbol{R} \in \mathcal{M}^+$, $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_p(n^{-1/2})$ and $\boldsymbol{R}(\cdot)$ is differentiable function at $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ satisfies $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_*$.

Note that if we consider $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{I}_m)$ and

$$\hat{\phi} = \frac{1}{nm - p} \sum_{i=1}^{n} \hat{\boldsymbol{\varepsilon}}_i' \hat{\boldsymbol{\varepsilon}}_i, \tag{4.3.1}$$

where $\hat{\boldsymbol{\varepsilon}}_i$ is defined in (4.2.2), then it follows from Liang & Zeger (1986) that the condition (C1) is established under the condition (C2). Under the conditions (C1)-(C3), an evaluation of the selection probability for an over-fitted correlation structure is obtained.

Theorem 4.3.1. *Under the conditions (C1)-(C3), for all $\boldsymbol{R} \in \mathcal{M}^+ \setminus \{\boldsymbol{R}_0\}$, when $\gamma_n \to \infty$,*

$$\lim_{n \to \infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) = 0.$$

Proof of Theorem 4.3.1. Denote q and q_* as the number of elements of correlation parameter for $\boldsymbol{R}$ and $\boldsymbol{R}_0$, respectively. From Definition 4.2.2, the selection probability of $\boldsymbol{R}$ is

$$\begin{aligned} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) &\leq Pr\{GIC_{\gamma_n}(\boldsymbol{R}_0) > GIC_{\gamma_n}(\boldsymbol{R})\} \\ &= Pr\{SL_n(\hat{\boldsymbol{R}}_0) - SL_n(\hat{\boldsymbol{R}}) > (q - q_*)\gamma_n\}. \end{aligned} \tag{4.3.2}$$

We evaluate $SL_n(\hat{\boldsymbol{R}}_0)$ and $SL_n(\hat{\boldsymbol{R}})$. Under the conditions (C1)-(C3), for all $\boldsymbol{R} \in \mathcal{M}^+$, it is established from the Taylor theorem that

$$\begin{aligned} n^{1/2}|(\hat{\boldsymbol{R}})_{jk} - (\boldsymbol{R}_*)_{jk}| &= n^{1/2}|(\boldsymbol{R}(\hat{\boldsymbol{\alpha}}))_{jk} - (\boldsymbol{R}_*)_{jk}| \\ &\leq n^{1/2}|\partial(\boldsymbol{R}(\tilde{\boldsymbol{\alpha}}))_{jk}/\partial\boldsymbol{\alpha}||\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}|, \end{aligned}$$

where $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_*$ and $\tilde{\boldsymbol{\alpha}}$ is a q-dimensional vector between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}$. Hence, it follows from $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_p(n^{-1/2})$ that

$$\hat{\boldsymbol{R}} - \boldsymbol{R}_* = O_p(n^{-1/2}). \tag{4.3.3}$$

On the contrary, let

$$\boldsymbol{R}_U^* = \frac{1}{n\phi} \sum_{i=1}^{n} \boldsymbol{A}_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\boldsymbol{A}_i^{-1/2}.$$

Since all elements of $\boldsymbol{R}_U^*$ are a differentiable function of $\boldsymbol{\beta}$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}| = O_p(n^{-1/2})$, it follows from a Taylor theorem that

$$\hat{\boldsymbol{R}}_U - \boldsymbol{R}_U^* = O_p(n^{-1/2}).$$

Note that in Θ^0, $a(\theta)$ is a C^∞-class function and all of the orders of the moments of y_{ij} exist and are bounded for all n under the regularity assumptions Fahrmeir & Kaufmann (1985). Additionally, these assure that the maximum eigenvalue of $\boldsymbol{A}_i^{-1}$ is upper bounded. Therefore, since the variance of $\sqrt{n}|(\boldsymbol{R}_U^*)_{jk} - (\boldsymbol{R}_*)_{jk}|$ is also upper bounded. Hence, by applying the Chebyshev inequality, for all $\delta > 0$, there exists a positive constant C such that

$$Pr\{\sqrt{n}|(\boldsymbol{R}_U^*)_{jk} - (\boldsymbol{R}_*)_{jk}| \geq \delta\} \leq C,$$

where $1 \leq j, k \leq m$. From this result

$$\boldsymbol{R}_U^* = \boldsymbol{R}_* + O_p(n^{-1/2}).$$

Hence,

$$\hat{\boldsymbol{R}}_U = \boldsymbol{R}_* + O_p(n^{-1/2}). \tag{4.3.4}$$

From (4.3.3) and (4.3.4), $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{R}}_U^{-1/2} \hat{\boldsymbol{R}} \hat{\boldsymbol{R}}_U^{-1/2} - \boldsymbol{I}_m = O_p(n^{-1/2})$. Hence, for all $\ell = 1, \ldots, m$, $\lambda_\ell(\hat{\boldsymbol{\Omega}}) = O_p(n^{-1/2})$, where $\lambda_\ell(\boldsymbol{A})$ is the ℓth smallest eigenvalue of $\boldsymbol{A}$ for any matrix $\boldsymbol{A}$. Hence, by applying a Taylor expansion, for all $\ell = 1, \ldots, m$,

$$\begin{aligned}
\log\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\} &= \lambda_\ell(\hat{\boldsymbol{\Omega}}) - \frac{1}{2}\lambda_\ell(\hat{\boldsymbol{\Omega}})^2 + O_p(n^{-3/2}), \\
\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\}^{-1} &= 1 - \lambda_\ell(\hat{\boldsymbol{\Omega}}) + \lambda_\ell(\hat{\boldsymbol{\Omega}})^2 + O_p(n^{-3/2}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\log\det(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}}) &= \sum_{\ell=1}^m \log\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\} \\
&= \mathrm{tr}(\hat{\boldsymbol{\Omega}}) - \frac{1}{2}\mathrm{tr}(\hat{\boldsymbol{\Omega}}^2) + O_p(n^{-3/2}), \\
\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} &= \sum_{\ell=1}^m \{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\}^{-1} \\
&= m - \mathrm{tr}(\hat{\boldsymbol{\Omega}}) + \mathrm{tr}(\hat{\boldsymbol{\Omega}}^2) + O_p(n^{-3/2}).
\end{aligned}$$

By substituting above results into (4.2.1),

$$\begin{aligned}
SL_n(\hat{\boldsymbol{R}}) &= n\log\det(\hat{\boldsymbol{R}}) + n\mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}^{-1}) \\
&= n\log\det(\hat{\boldsymbol{R}}_U) + n\log\det(\hat{\boldsymbol{R}}\hat{\boldsymbol{R}}_U^{-1}) + n\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} \\
&= n\log\det(\hat{\boldsymbol{R}}_U) + n\log\det(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}}) + n\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} \\
&= n\log\det(\hat{\boldsymbol{R}}_U) + nm + n\mathrm{tr}(\hat{\boldsymbol{\Omega}}^2)/2 + O_p(n^{-1/2}) \\
&= n\log\det(\hat{\boldsymbol{R}}_U) + nm + O_p(1). \qquad (4.3.5)
\end{aligned}$$

It follows from (4.3.5) that $SL_n(\hat{\boldsymbol{R}}_0) - SL_n(\hat{\boldsymbol{R}}) = O_p(1)$. Note that the definition of $\boldsymbol{R}_0$ implies that $q - q_* > 0$ holds. By substituting these results into (4.3.2), since $\gamma_n \to \infty$ as $n \to \infty$, then

$$\lim_{n\to\infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) = 0.$$

□

A similar result can be shown for the under-fitted structure.

Theorem 4.3.2. *Under the conditions (C1)-(C3), for all $\boldsymbol{R} \in \mathcal{M}^-$, when $\gamma_n/n \to 0$,*

$$\lim_{n\to\infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) = 0.$$

Proof of Theorem 4.3.2. As in (4.3.2), the selection probability of $\boldsymbol{R} \in \mathcal{M}^-$ is evaluated by

$$Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) \le Pr\left\{\frac{1}{n}SL_n(\hat{\boldsymbol{R}}_0) - \frac{1}{n}SL_n(\hat{\boldsymbol{R}}) > \frac{(q-q_*)\gamma_n}{n}\right\},$$

where q and q_* are the number of elements in $\boldsymbol{R} \in \mathcal{M}^-$ and $\boldsymbol{R}_0$, respectively. $SL_n(\hat{\boldsymbol{R}}_0)/n - SL_n(\hat{\boldsymbol{R}})/n$ can be separated by using

$$\rho(\boldsymbol{A}) = -\log\det(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{A}) - m$$

as follows:

$$\begin{aligned}
&\frac{1}{n}SL_n(\hat{\boldsymbol{R}}_0) - \frac{1}{n}SL_n(\hat{\boldsymbol{R}}) \\
&= -\log\det(\hat{\boldsymbol{R}}_0^{-1}) + \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) + \log\det(\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}^{-1}) \\
&= -\{\log\det(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) - \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) + m\} \\
&\quad + \{\log\det(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + m\} \\
&\quad + \{\log\det(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) - \mathrm{tr}(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) + m\} \\
&\quad - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \\
&= \rho(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) - \rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \rho(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) \\
&\quad - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\}.
\end{aligned} \tag{4.3.6}$$

It follows from $\hat{\boldsymbol{R}}_U \to \boldsymbol{R}_*$ and $\hat{\boldsymbol{R}}_0 \to \boldsymbol{R}_*$ in probability under the conditions (C1)-(C3) that

$$\rho(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) = o_p(1), \quad \rho(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) = o_p(1). \tag{4.3.7}$$

Let $c = \inf_{\boldsymbol{\alpha}\in\mathcal{K}} \rho(\boldsymbol{R}_*\boldsymbol{R}(\boldsymbol{\alpha})^{-1})$. If $c = 0$, from the compactness of $\mathcal{K}$, there exists a sequence $\{\boldsymbol{\alpha}_\ell | \ell = 1, 2, \ldots\}$ such that $\boldsymbol{\alpha}_\ell \to \boldsymbol{\alpha}_* \in \mathcal{K}$ which satisfies $\rho(\boldsymbol{R}(\boldsymbol{\alpha}_\ell)^{-1}\boldsymbol{R}_*) \to 0$. Since $\rho(\boldsymbol{A})$ is a continuous function on $\mathcal{A}_L = \{\boldsymbol{A} | \rho(\boldsymbol{A}) \le L\}$ for all $L > 0$, $\boldsymbol{R}(\boldsymbol{\alpha}_*) = \boldsymbol{R}_*$ holds which contradicts that $\boldsymbol{R} \in \mathcal{M}_-$. Hence, $c > 0$ is established.

Here, for all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{A}_L$, and $t \in [0, 1]$

$$\begin{aligned}
&t\rho(\boldsymbol{A}) + (1-t)\rho(\boldsymbol{B}) - \rho(t\boldsymbol{A} + (1-t)\boldsymbol{B}) \\
&= \log\det\{t\boldsymbol{A}\boldsymbol{B}^{-1} + (1-t)\boldsymbol{I}_m\} - \log\det(t\boldsymbol{A}\boldsymbol{B}^{-1}) \\
&= \sum_{\ell=1}^{m}[\log\{t\lambda_\ell(\boldsymbol{A}\boldsymbol{B}^{-1}) + (1-t)\} - \log\{t\lambda_\ell(\boldsymbol{A}\boldsymbol{B}^{-1})\}] \ge 0.
\end{aligned}$$

The last inequality is established from the fact that the logarithm is concave. Hence, $\rho(t\boldsymbol{A} + (1-t)\boldsymbol{B}) \le t\rho(\boldsymbol{A}) + (1-t)\rho(\boldsymbol{B}) \le L$ holds, and then $t\boldsymbol{A} + (1-t)\boldsymbol{B} \in \mathcal{A}_L$. Therefore, $\mathcal{A}_L$ is a convex set.

Let $\boldsymbol{A}_{[t]} = \boldsymbol{I}_m + t(\hat{\boldsymbol{R}}^{-1}\boldsymbol{R}_* - \boldsymbol{I}_m)$. Then, $\rho(\boldsymbol{A}_{[0]}) = \rho(\boldsymbol{I}_m) = 0$ and $\rho(\boldsymbol{A}_{[1]}) = \rho(\hat{\boldsymbol{R}}^{-1}\boldsymbol{R}_*) \ge c$. Since for all $L > 0$, $\mathcal{A}_L$ is the convex set and $\rho(\cdot)$ is continuous on $\mathcal{A}_L$, there exists $t \in [0, 1]$ such that $\rho(\boldsymbol{A}_{[t]}) = c$. It follows from the convexness of

$$g(t) = \rho(\boldsymbol{A}_{[t]}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{A}_{[t]} - \boldsymbol{I}_m)\}$$

that

$$\begin{aligned}
&\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \\
&= \frac{g(1)-g(0)}{1-0} \geq \frac{g(t)-g(0)}{t-0} \geq g(t) \\
&= c + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{A}_{[t]} - \boldsymbol{I}_m)\} \\
&\geq c - \sqrt{\mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)^2\}\mathrm{tr}\{(\boldsymbol{A}_{[t]} - \boldsymbol{I}_m)^2\}} \\
&\geq c - \sqrt{\mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)^2\}b} \qquad (4.3.8)
\end{aligned}$$

where

$$b = \max\{\mathrm{tr}\{(\boldsymbol{A} - \boldsymbol{I}_m)^2\} | \rho(\boldsymbol{A}) = c\} > 0.$$

Denote E as the event that $\{\mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)^2\} < c^2/(4b)\}$ and E^c as the complement of E. Under the event E, from (4.3.8), it is established that

$$\begin{aligned}
&\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \\
&\geq c - c/2 = c/2.
\end{aligned} \qquad (4.3.9)$$

On the other hand, we can see that

$$\begin{aligned}
&Pr\{-\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} < -c/2\} \\
&= Pr\{\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} > c/2\} \\
&= 1 - Pr\{\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \leq c/2\} \\
&= 1 - Pr(\{\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \leq c/2\} \cap E) \\
&\quad - Pr(\{\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \leq c/2\} \cap E^c).
\end{aligned}$$

Thereby, it follows from (4.3.9) that

$$\begin{aligned}
&Pr\{-\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} < -c/2\} \\
&\geq 1 - Pr(E^c) \to 1,
\end{aligned} \qquad (4.3.10)$$

where the last convergence is established from $\hat{\boldsymbol{R}}_U \to \boldsymbol{R}_*$ in probability.

From (4.3.6), (4.3.7), (4.3.10), and $(q-q_*)\gamma_n/n \to 0$, for all $\boldsymbol{R} \in \mathcal{M}^-$,

$$\lim_{n\to\infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) = 0.$$

□

From these theorems, a sufficient condition for the consistency of our criterion is obtained.

Theorem 4.3.3. *Suppose $\gamma_n \to \infty$ and $\gamma_n/n \to 0$. Under the conditions (C1)-(C3),*

$$\lim_{n\to\infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}_0) = 1$$

holds.

Proof of Theorem 4.3.3. The probability of the true correlation structure selection is divided into two parts, as follows:

$$\begin{aligned} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}_0) &= 1 - Pr(\boldsymbol{R}_{best} \neq \boldsymbol{R}_0) \\ &\geq 1 - \sum_{\boldsymbol{R}\in\mathcal{M}\setminus\{\boldsymbol{R}_0\}} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) \\ &\geq 1 - \sum_{\boldsymbol{R}\in\mathcal{M}^+\setminus\{\boldsymbol{R}_0\}} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) - \sum_{\boldsymbol{R}\in\mathcal{M}^-} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}). \end{aligned}$$

From Theorem 4.3.1 and Theorem 4.3.2, it follows that

$$\lim_{n\to\infty} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}_0) = 1.$$

□

4.4 Numerical Studies

In this section, we present numerical studies to illustrate the performance of our criterion in a finite sample situation. We prepared $\gamma_n = 2$, $2\log\log n$ and $\log n$, respectively, as the AIC-type, the Hannan & Quinn's IC(HQIC)-type proposed in Hannan & Quinn (1979), and the BIC-type tuning parameters for the GIC proposed in (4.2.3). For convenience, the GICs with $\gamma_n = 2$, $2\log\log n$ and $\log n$ are called the AIC, the HQIC and the BIC, respectively. We compared some properties of the AIC, the HQIC and the BIC with those of the QIC and the CIC. The QIC and the CIC for the working correlation structure $\boldsymbol{R}$ are defined as follows:

$$\begin{aligned} QIC(\boldsymbol{R}) &= \hat{\phi}^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m} L(\hat{\mu}_{ij}; y_{ij}) + 2\mathrm{tr}(\hat{\boldsymbol{V}}_s\hat{\boldsymbol{\Sigma}}_I), \\ CIC(\boldsymbol{R}) &= \mathrm{tr}(\hat{\boldsymbol{V}}_s\hat{\boldsymbol{\Sigma}}_I), \end{aligned}$$

where $\hat{\mu}_{ij}$ is the estimator of μ_{ij}, $L(\mu_{ij}; y_{ij}) = y_{ij} \log \mu_{ij} + (1 - y_{ij}) \log(1 - \mu_{ij})$,

$$\boldsymbol{V}_s = \boldsymbol{\Sigma}_R^{-1} \left\{ \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right\} \boldsymbol{\Sigma}_R^{-1},$$

$$\boldsymbol{\Sigma}_R = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \quad \boldsymbol{\Sigma}_I = \phi^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{A}_i^{-1} \boldsymbol{D}_i,$$

where $\hat{\boldsymbol{V}}_s$ and $\hat{\boldsymbol{\Sigma}}_I$ are estimators of $\boldsymbol{V}_s$ and $\boldsymbol{\Sigma}_I$ obtained by substituting the GEE estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{R})$ and $\hat{\boldsymbol{\alpha}}$ into $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively, and $\boldsymbol{V}_i$ is defined in (4.1.1). Note that the CIC is the same as half of the second term in the QIC. Throughout this section, we assumed $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{I}_m)$ and that $\hat{\phi}$ is as given in (4.3.1), for calculating the GIC.

We prepared four candidate models, each with 50, 100, 200, 500 and 1,000 samples. For each sample, we had a four-dimensional response vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i4})'$ and a 4×2 explanatory matrix $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i4})'$. Let $\boldsymbol{x}_{ij} = (1, x_{ij})'$, and assume that the x_{ij} were independent and identically distributed as the uniform distribution $U(-1, 1)$. We assumed that y_{ij} was distributed as $B(1, p_{ij})$ according to a logistic regression model, i.e., $p_{ij} = 1/\{1 + \exp(-\boldsymbol{x}_{ij}'\boldsymbol{\beta})\}$ and $\boldsymbol{\beta} = (1, -1)'$. A set of candidate correlation structures $\mathcal{M}$ was considered in the following case, introduced in (4.1.2):

$$\mathcal{M} = \{\text{“Indep.”}, \text{“Ex.”}, \text{“AR} - 1\text{”}, \text{“Unst.”}\}.$$

In all simulations, we assumed that the true correlation structure of $\boldsymbol{y}_i$ was an element of $\mathcal{M}$, as defined below:

$$\begin{aligned} \text{Indep.} &: \boldsymbol{R}_0 = \boldsymbol{I}_4, \\ \text{Ex.} &: \boldsymbol{R}_0 = \boldsymbol{I}_4/2 + \boldsymbol{1}_4\boldsymbol{1}_4'/2, \text{ where } \boldsymbol{1}_4 = (1,1,1,1)', \\ \text{AR-1} &: (\boldsymbol{R}_0)_{jk} = 2^{-|j-k|}, \\ \text{Unst.} &: \boldsymbol{R}_0 = \boldsymbol{H}_d^{-1/2} \boldsymbol{H} \boldsymbol{H}_d^{-1/2}, \text{ where } \boldsymbol{H} = (h_{ij})_{1 \le i,j \le 4} = \boldsymbol{W}'\boldsymbol{W} + \\ &\quad \boldsymbol{I}_4,\ (\boldsymbol{W})_{jk} \overset{i.i.d.}{\sim} U(-1,1) \text{ and } \boldsymbol{H}_d = \text{diag}\{h_{11}, \ldots, h_{44}\}. \end{aligned}$$

The correlation parameter $\boldsymbol{\alpha}$ was estimated for each candidate correlation structure by using the following moment-based method:

$$\text{Ex.} : \hat{\alpha} = \frac{2}{nm(m-1)} \sum_{i=1}^{n} \sum_{j>k} \hat{\varepsilon}_{ij} \hat{\varepsilon}_{ik},$$

$$\begin{aligned} \text{AR-1}: \ \hat{\alpha} &= \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m-1} \hat{\varepsilon}_{ij} \hat{\varepsilon}_{i,j+1}, \\ \text{Unst.}: \ \hat{\alpha}_{jk} &= \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_{ij} \hat{\varepsilon}_{ik}. \end{aligned}$$

Note that the conditions (C1)-(C3) held in this simulation setting. The BIC satisfied the assumptions of Theorem 4.3.3 but the AIC satisfies only the assumption of Theorem 4.3.2. For the situations described above, we simulated 1,000 repetitions.

In the numerical studies, we considered three measurements to evaluate the criteria: the selection probability of the true structure, the predictive mean squared error (PMSE), and the average value of the variance of $\hat{\boldsymbol{\beta}}$ (VAR) with the best correlation structure selected by each criterion. The definitions of the PMSE and VAR are

$$\begin{aligned} \text{PMSE}: \ & \frac{1}{1000} \sum_{\ell=1}^{1000} \sum_{i=1}^{n} \{\hat{\boldsymbol{\mu}}_{i,best}^{(\ell)} - \boldsymbol{\mu}_i\}' \boldsymbol{V}_i^{-1} \{\hat{\boldsymbol{\mu}}_{i,best}^{(\ell)} - \boldsymbol{\mu}_i\}, \\ \text{VAR}: \ & \frac{1}{1000} \sum_{\ell=1}^{1000} \left| \hat{\boldsymbol{\beta}}_{best}^{(\ell)} - \frac{1}{1000} \sum_{\ell=1}^{1000} \hat{\boldsymbol{\beta}}_{best}^{(\ell)} \right|^2, \end{aligned}$$

where $\hat{\boldsymbol{\mu}}_{i,best}^{(\ell)}$ and $\hat{\boldsymbol{\beta}}_{best}^{(\ell)}$ are the estimators of $\boldsymbol{\mu}_i = \mathrm{E}[\boldsymbol{y}_i]$, and $\boldsymbol{\beta}$ is from the best correlation structure in the ℓth iteration.

Tables 4.1-4.4 listed the results of the selection probability and the ratios of the PMSE and VAR to the values of the BIC. From Tables 4.1-4.4, we could look the consistency of the BIC, and we saw that on many occasions, the QIC and CIC did not select the true correlation structure frequently. In all cases except “Unstructured” with $n = 50$ and $n = 100$, the BIC performed better than other criteria. When the sample size was small, the improvements from the BIC were especially good. In the “Unstructured” case, the AIC, the HQIC and the CIC performed better than the BIC. This result implied that the penalty term of the BIC was too big to select the “Unstructured” correlation structure when the sample size was not large in comparison with the true correlation parameter. The QIC and the CIC might have a tendency to select the over-fitted structure rather than the true structure. Based on these results, we recommend using the BIC to select the true correlation structure when the sample size is large. However, if the sample size is

Table 4.1: Selection probability, predictive mean square error, and variance of $\hat{\boldsymbol{\beta}}$ when the true correlation structure is “Independent”

n	IC	Indep.	Ex.	AR-1	Unst.	PMSE	VAR
50	AIC	**704**	138	109	49	1.00	1.00
	HQIC	**824**	91	66	19	1.00	1.00
	BIC	**924**	44	30	2	1.00	1.00
	CIC	**39**	50	52	859	1.05	1.06
	QIC	**121**	103	124	652	1.01	1.01
100	AIC	**714**	120	123	43	1.00	1.00
	HQIC	**858**	66	72	4	1.00	1.00
	BIC	**941**	31	28	0	1.00	1.00
	CIC	**57**	64	56	823	1.03	1.04
	QIC	**124**	125	113	638	1.01	1.01
200	AIC	**719**	123	112	46	1.00	1.00
	HQIC	**870**	62	65	3	1.00	1.00
	BIC	**956**	21	23	0	1.00	1.00
	CIC	**54**	49	56	841	1.01	1.01
	QIC	**116**	106	130	648	1.00	1.01
500	AIC	**716**	119	124	41	1.00	1.01
	HQIC	**907**	43	49	1	1.00	1.00
	BIC	**976**	9	15	0	1.00	1.00
	CIC	**49**	55	59	837	1.01	1.01
	QIC	**116**	109	125	650	1.01	1.01
1000	AIC	**727**	103	119	51	1.00	1.00
	HQIC	**914**	42	44	0	1.00	1.00
	BIC	**983**	9	8	0	1.00	1.00
	CIC	**52**	58	42	848	1.00	1.00
	QIC	**115**	121	112	652	1.00	1.00

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

Table 4.2: Selection probability, predictive mean square error, and variance of $\hat{\boldsymbol{\beta}}$ when the true correlation structure is “Exchangeable”

n	IC	Indep.	Ex.	AR-1	Unst.	PMSE	VAR
50	AIC	0	**680**	39	281	1.02	1.01
	HQIC	0	**826**	46	128	1.01	1.00
	BIC	0	**908**	54	38	1.00	1.00
	CIC	0	**103**	25	872	1.02	1.00
	QIC	134	**253**	78	535	1.12	1.26
100	AIC	0	**727**	1	272	1.01	1.01
	HQIC	0	**899**	8	93	1.01	1.01
	BIC	0	**974**	9	17	1.00	1.00
	CIC	0	**121**	8	871	1.02	1.02
	QIC	112	**291**	69	528	1.14	1.21
200	AIC	0	**703**	0	297	1.00	1.00
	HQIC	0	**930**	0	70	1.00	1.00
	BIC	0	**992**	0	8	1.00	1.00
	CIC	0	**117**	0	883	1.00	1.00
	QIC	120	**307**	52	521	1.13	1.20
500	AIC	0	**726**	0	274	1.00	1.00
	HQIC	0	**959**	0	41	1.00	1.00
	BIC	0	**1000**	0	0	1.00	1.00
	CIC	0	**107**	0	893	1.00	1.00
	QIC	107	**339**	46	508	1.12	1.18
1000	AIC	0	**731**	0	269	1.00	1.00
	HQIC	0	**968**	0	32	1.00	1.00
	BIC	0	**1000**	0	0	1.00	1.00
	CIC	0	**136**	0	864	1.00	1.00
	QIC	131	**336**	56	477	1.17	1.25

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

Table 4.3: Selection probability, predictive mean square error, and variance of $\hat{\boldsymbol{\beta}}$ when the true correlation structure is "AR-1"

n	IC	Indep.	Ex.	AR-1	Unst.	PMSE	VAR
50	AIC	0	32	**756**	212	1.02	1.02
	HQIC	0	38	**883**	79	1.01	1.01
	BIC	0	38	**940**	22	1.00	1.00
	CIC	0	13	**118**	869	1.04	1.05
	QIC	105	126	**264**	505	1.15	1.24
100	AIC	0	3	**802**	195	1.01	1.01
	HQIC	0	7	**938**	55	1.00	1.00
	BIC	0	9	**985**	6	1.00	1.00
	CIC	0	4	**129**	867	1.01	1.01
	QIC	87	125	**289**	499	1.14	1.22
200	AIC	0	0	**797**	203	1.01	1.01
	HQIC	0	0	**951**	49	1.00	1.00
	BIC	0	0	**997**	3	1.00	1.00
	CIC	0	0	**123**	877	1.01	1.01
	QIC	78	140	**307**	475	1.12	1.19
500	AIC	0	0	**803**	197	1.00	1.00
	HQIC	0	0	**977**	23	1.00	1.00
	BIC	0	0	**1000**	0	1.00	1.00
	CIC	0	0	**120**	880	1.00	1.00
	QIC	81	152	**296**	471	1.11	1.17
1000	AIC	0	0	**810**	190	1.00	1.00
	HQIC	0	0	**985**	15	1.00	1.00
	BIC	0	0	**1000**	0	1.00	1.00
	CIC	0	0	**147**	853	1.00	1.00
	QIC	83	141	**326**	450	1.11	1.18

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

Table 4.4: Selection probability, predictive mean square error, and variance of $\hat{\boldsymbol{\beta}}$ when the true correlation structure is "Unstructured"

n	IC	Indep.	Ex.	AR-1	Unst.	PMSE	VAR
50	AIC	41	1	51	**907**	0.93	0.91
	HQIC	113	4	98	**785**	0.95	0.94
	BIC	326	3	179	**492**	1.00	1.00
	CIC	0	1	2	**997**	0.93	0.92
	QIC	31	21	167	**781**	0.99	1.00
100	AIC	0	0	4	**996**	0.99	0.99
	HQIC	10	0	13	**977**	0.99	0.99
	BIC	52	1	64	**883**	1.00	1.00
	CIC	0	0	0	**1000**	0.99	0.99
	QIC	24	16	163	**797**	1.06	1.04
200	AIC	0	0	0	**1000**	1.00	1.00
	HQIC	0	0	0	**1000**	1.00	1.00
	BIC	0	0	1	**999**	1.00	1.00
	CIC	0	0	0	**1000**	1.00	1.00
	QIC	13	8	174	**805**	1.08	1.06
500	AIC	0	0	0	**1000**	1.00	1.00
	HQIC	0	0	0	**1000**	1.00	1.00
	BIC	0	0	0	**1000**	1.00	1.00
	CIC	0	0	0	**1000**	1.00	1.00
	QIC	7	4	157	**832**	1.07	1.06
1000	AIC	0	0	0	**1000**	1.00	1.00
	HQIC	0	0	0	**1000**	1.00	1.00
	BIC	0	0	0	**1000**	1.00	1.00
	CIC	0	0	0	**1000**	1.00	1.00
	QIC	9	5	164	**822**	1.08	1.06

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

not large, we recommend using the AIC or the HQIC for a conservative selection.

Chapter 5

Conclusions and Discussions

In this paper, we proposed three model selection criteria in models related to the GLMs, from two different perspectives, i.e., bias-correction and consistency of the model selection. In the former of the present paper, we derived a simple formula for calculating the CAIC in the GLM and 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 $O(n^{-1})$. Furthermore, both models are widely used in the real data analysis. The GLM can express a number of statistical models by changing the distribution and the link function and can be easily fitted to the real data using the function "glm" in the "R" software, which implements some distributions and link functions. On the contrary, by using the function "vglm" in the "R" software, the estimation results in the multinomial logistic regression model can be obtained without difficulty. Hence, the researcher can easily obtain the CAIC using formula manipulation software, even if a researcher wants to use the CAIC in a model for which an example CAIC has not yet been derived. Thereby, the proposed CAIC is confirmed to be useful in real data analysis.

For using the CAIC, we deal primarily with variable selection. However, in the simulation of the real data analysis in Section 2.3, we also considered the selection of the link function in the GLM. If we choose the link function by minimizing the original AIC, the optimal link function is determined only by maximizing the log-likelihood function. On the other hand, if we use the CAIC to select the link function, the optimal link function is not determined only by maximizing the log-likelihood function. Thus, using the CAIC will allow us to select an appropriate link function in the GLM. These results will be able to be expanded to a multivariate data of the GLM based on the derivation of the CAIC in the multinomial logistic regression model

The numerical studies revealed that the CAIC was better than the original AIC. In all situations of the simulation studies, 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 when the sample size is small so that the bias of the AIC is corrected. Needless to say, the AIC and the CAIC are asymptotically equivalent. Hence, the difference between two criteria becomes small when the sample size is increased. Nevertheless, in Section 2.3 and 3.3, we can see that 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 the best model.

In the latter of the present paper, for selecting the true correlation matrix in the GEE frameworks, we proposed a GIC-type criterion based on Stein′s loss function, which is the discrepancy between the true correlation structure and a working correlation structure, and we derived sufficient conditions to establish its consistency. The GEE is an expansion of the GLM into the dependent data.

Since the consistency of our criterion is shown from the property of Stein′s loss function and the $n^{1/2}$-consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, we will be able to expand this class of criteria and its consistency to general semi-parametric frameworks. Moreover, it may be possible to show that the criterion based on other loss functions (such as the quadratic loss function) has the consistency property.

Through the simulation results, we confirmed that the proposed criterion with $\gamma_n = \log n$ often selects the true correlation structure in large sample situations. Furthermore, this selection method improves the PMSE and the variance of $\hat{\boldsymbol{\beta}}$, which are of primary interest in the GEE frameworks. However, when the true correlation structure is “Unstructured” and n is not sufficiently large, the BIC-type criterion did not work well in the simulation. This may arise from that the number of the correlation parameter for “Unstructured” is too many with respect to the sample size.

In order to solve this problem, we consider two approaches. One is to consider this situation as a high-dimensional setting, i.e., we allow the number of correlation parameters to be as large as the sample size. This indicates that the dimension of the responses m is assumed to be large as same as the sample size n. Another approach is to construct a risk function based on Stein′s loss function and to derive a bias-corrected criterion, as was done in Sugiura (1978); Hurvich & Tsai (1989); Imori, et al. (2014), and Sections 2 and 3. We expect that these approaches will yield more adequate criteria or assumptions for selecting the true correlation structure.

Furthermore, the primary aim of using the GEE is to estimate the regression coefficients. Then, to consider a subset selection of explanatory variables is important, and we had already proposed several adequate criteria to select the best subset of explanatory variables (see, e.g., Imori, 2013; Inatsu & Imori, 2013). However, the properties of these criteria such as efficiency and optimality (Shibata, 1980; Shao, 1997) have been not shown yet. Moreover, by generalizing penalty term of these criteria, we will be able to derive a criterion which has a consistency of the true subset, which is like as the result shown in Section 4.3. We will consider above problems in the future.

Acknowledgements

I would like to express the deepest gratitude to Professor Hirofumi Wakaki of Hiroshima University for his valuable comments, encouragement and introducing me many fields of mathematical statistics.

I would also like to express my sincere gratitude to Dr. Hirokazu Yanagihara of Hiroshima University for his helpful advice about my research work.

I wish to thank Dr. Kenichi Satoh of Hiroshima University and Dr. Ken-ichi Kamo of Sapporo Medical University for their constructive discussion. Also, I wish to thank Professor Yasunori Fujikoshi of Hiroshima University and Professor Megu Ohtaki of Hiroshima University for their encouragement. Many thanks to my colleagues at Hiroshima University for their cooperation.

Further, I am grateful to Professor Dietrich von Rosen of Swedish University of Agricultural Sciences who helped and supported me while I studied in Sweden.

This thesis is dedicated to my wife, Hisako.

Chapter A

Appendix

A.1 Derivation of the Third Term of (2.2.4)

In order to calculate the moments of $\boldsymbol{b}_1$ and $\boldsymbol{Z}_2$, we rewrite the third term of (2.2.4) using $\boldsymbol{b}_1$, $\boldsymbol{Z}_2$, and $\boldsymbol{M}_3$ as

$$\begin{aligned}&\frac{2}{n\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i-a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_2+\frac{1}{2}c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)^2\right\}\right]\\&=\frac{1}{\sqrt{n}}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{M}_3(\boldsymbol{b}_1\otimes\boldsymbol{b}_1)]+\frac{3}{\sqrt{n}}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_2\boldsymbol{b}_1],\end{aligned}\tag{A.1.1}$$

where $a_{ij}=\partial^j a(\theta_i)/\partial\theta_i^j$, $c_{ij}=\partial^j\theta_i/\partial\eta_i^j$. Let $\varphi_{\boldsymbol{b}_1}(\boldsymbol{t})$ be the characteristic function of the distribution of $\boldsymbol{b}_1$, defined as

$$\begin{aligned}\varphi_{\boldsymbol{b}_1}(\boldsymbol{t})&=\mathrm{E}[\exp(i\boldsymbol{t}'\boldsymbol{b}_1)]=\mathrm{E}\left[\exp\left\{\sum_{j=1}^{n}i(y_j-a_{j1})s_j\right\}\right],\\s_j&=-\frac{1}{\sqrt{n}\phi}c_{j1}\boldsymbol{t}'\boldsymbol{M}_2^{-1}\boldsymbol{x}_j,\end{aligned}$$

where $\boldsymbol{t}=(t_1,\ldots,t_p)'$. Note that $\mathrm{E}[\exp\{i(y-\mu)s\}]$ is the characteristic function of $y-\mu$, which is expressed as

$$\mathrm{E}[\exp\{i(y-\mu)s\}]=\exp\left\{\frac{b(\theta+is\phi)-b(\theta)}{\phi}-i\mu s\right\}.$$

Therefore, we have

$$\varphi_{\boldsymbol{b}_1}(\boldsymbol{t}) = \exp\left[\sum_{j=1}^{n}\left\{\frac{b(\theta_j + is_j\phi) - b(\theta_j)}{\phi} - ia_{j1}s_j\right\}\right].$$

Based on the property of the random variable with mean zero, the third moment is equivalent to the third cumulant. Since $|s_j| = O(n^{-1/2})$, $\log\varphi_{\boldsymbol{b}_1}(\boldsymbol{t})$ can be expanded as

$$\begin{aligned}\log\varphi_{\boldsymbol{b}_1}(\boldsymbol{t}) &= \sum_{j=1}^{n}\left\{\frac{b(\theta_j + is_j\phi) - b(\theta_j)}{\phi} - ia_{j1}s_j\right\}\\ &= \frac{1}{\phi}\sum_{j=1}^{n}\left\{\frac{1}{2}a_{j2}(is_j\phi)^2 + \frac{1}{6}a_{j3}(is_j\phi)^3 + \frac{1}{24}a_{j4}(is_j\phi)^4\right\} + O(n^{-3/2}).\end{aligned}$$

Thus, the third cumulant of $\boldsymbol{b}_1 = (b_{11}, \dots, b_{1p})'$ is computed through the derivative of $\log\varphi_{\boldsymbol{b}_1}(\boldsymbol{t})$, i.e.,

$$\begin{aligned}\mathrm{E}[b_{1\alpha_1}b_{1\alpha_2}b_{1\alpha_3}] &= i^{-3}\left.\frac{\partial^3\log\varphi_{\boldsymbol{b}_1}(\boldsymbol{t})}{\partial t_{\alpha_1}\partial t_{\alpha_2}\partial t_{\alpha_3}}\right|_{\boldsymbol{t}=\boldsymbol{0}}\\ &= \phi^2\sum_{j=1}^{n}a_{j3}\frac{\partial s_j}{\partial t_{\alpha_1}}\frac{\partial s_j}{\partial t_{\alpha_2}}\frac{\partial s_j}{\partial t_{\alpha_3}} + O(n^{-3/2}).\end{aligned}$$

Note that

$$\frac{\partial s_j}{\partial t_{\alpha_l}} = -\frac{1}{\sqrt{n}\phi}c_{j1}\boldsymbol{e}_{\alpha_l}'\boldsymbol{M}_2^{-1}\boldsymbol{x}_j,$$

where $\boldsymbol{e}_j$ is the p-dimensional vector, the jth element of which is 1 and the other elements of which are 0. Thus, using the Equation (2.2.8), we obtain

$$\begin{aligned}&\frac{1}{\sqrt{n}}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{M}_3(\boldsymbol{b}_1\otimes\boldsymbol{b}_1)]\\ &= \frac{1}{n^3\phi}\sum_{i,j}^{n}a_{j3}c_{j1}^3(a_{i3}c_{i1}^3 + 3a_{i2}c_{i1}c_{i2})u_{ij}^3 + O(n^{-2}).\end{aligned} \tag{A.1.2}$$

Let $\varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2}(\boldsymbol{t},\boldsymbol{T}_1)$ denote the characteristic function of the joint distribution for $\boldsymbol{b}_1$ and $\boldsymbol{Z}_2$ as

$$\varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2}(\boldsymbol{t},\boldsymbol{T}_1) = \exp\left[\sum_{j=1}^{n}\left\{\frac{b(\theta_j + iv_j\phi) - b(\theta_j)}{\phi} - ia_{j1}v_j\right\}\right],$$

where $(\boldsymbol{T}_1)_{ij} = t_{ij}^{(1)}$, $i, j = 1, \ldots, p$ and

$$v_j = \frac{1}{\sqrt{n}\phi}(-c_{j1}\boldsymbol{t}'\boldsymbol{M}_2^{-1}\boldsymbol{x}_j + c_{j2}\boldsymbol{x}_j'\boldsymbol{T}_1\boldsymbol{x}_j).$$

In the same manner as in the calculation of $\log \phi_{\boldsymbol{b}_1}(\boldsymbol{t})$, we have

$$\begin{aligned}
&\log \varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2}(\boldsymbol{t}, \boldsymbol{T}_1)\\
&= \sum_{j=1}^{n} \left\{ \frac{b(\theta_j + iv_j\phi) - b(\theta_j)}{\phi} - ia_{j1}v_j \right\}\\
&= \frac{1}{\phi}\sum_{j=1}^{n} \left\{ \frac{1}{2}a_{j2}(iv_j\phi)^2 + \frac{1}{6}a_{j3}(iv_j\phi)^3 + \frac{1}{24}a_{j4}(iv_j\phi)^4 \right\} + O(n^{-3/2}).
\end{aligned}$$

Note that

$$\frac{\partial v_k}{\partial t_i} = -\frac{1}{\sqrt{n}\phi}c_{k1}\boldsymbol{e}_i'\boldsymbol{M}_2^{-1}\boldsymbol{x}_k, \quad \frac{\partial v_k}{\partial t_{ij}^{(1)}} = \frac{1}{\sqrt{n}\phi}c_{k2}(\boldsymbol{e}_i'\boldsymbol{x}_k)(\boldsymbol{e}_j'\boldsymbol{x}_k).$$

Hence, we obtain

$$\frac{1}{\sqrt{n}}\mathrm{E}[\boldsymbol{b}_1\boldsymbol{Z}_2\boldsymbol{b}_1] = \frac{1}{n^2\phi}\sum_{i=1}^{n} a_{i3}c_{i1}^2c_{i2}u_{ii}^2 + O(n^{-2}). \tag{A.1.3}$$

Substituting (A.1.2) and (A.1.3) into (A.1.1), the third term of (2.2.4) is given by (2.2.7).

A.2 Derivation of the Fourth Term of (2.2.4)

In order to use the asymptotic properties, we express the fourth term of (2.2.4) in terms of $\boldsymbol{b}_1$, $\boldsymbol{Z}_2$, $\boldsymbol{Z}_3$ $\boldsymbol{M}_2$ $\boldsymbol{M}_3$, and $\boldsymbol{M}_4$ as

$$\begin{aligned}
&\frac{2}{n\sqrt{n}\phi}\sum_{i=1}^{n}\mathrm{E}\left[(y_i - a_{i1})\left\{c_{i1}\boldsymbol{x}_i'\boldsymbol{b}_3 + c_{i2}(\boldsymbol{x}_i'\boldsymbol{b}_1)(\boldsymbol{x}_i'\boldsymbol{b}_2) + \frac{1}{6}c_{i3}(\boldsymbol{x}_i'\boldsymbol{b}_1)^3\right\}\right]\\
&= -\frac{1}{n}\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] + \frac{1}{3n}\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_4(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)]\\
&\quad - \frac{3}{n}\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1] - \frac{4}{n}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_2\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1]\\
&\quad + \frac{4}{3n}\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)].
\end{aligned} \tag{A.2.1}$$

Let $\varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3}(\cdot)$ denote the characteristic function of the joint distribution for $\boldsymbol{b}_1$, $\boldsymbol{Z}_2$, and $\boldsymbol{Z}_3$ defined by

$$\varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3}(\boldsymbol{t},\boldsymbol{T}_2,\boldsymbol{T}_3) = \exp\left[\sum_{j=1}^{n}\left\{\frac{b(\theta_j + ir_j\phi) - b(\theta_j)}{\phi} - ia_{j1}r_j\right\}\right],$$

where $(\boldsymbol{T}_2)_{ij} = t_{ij}^{(2)}$, $i,j = 1,\ldots,p$, $(\boldsymbol{T}_3)_{ijk} = t_{ijk}^{(3)}$, $i,j,k = 1,\ldots,p$ and

$$r_j = \frac{1}{\sqrt{n\phi}}\{-c_{j1}\boldsymbol{t}'\boldsymbol{M}_2^{-1}\boldsymbol{x}_j + c_{j2}\boldsymbol{x}_j'\boldsymbol{T}_2\boldsymbol{x}_j + c_{j3}\boldsymbol{x}_j'\boldsymbol{T}_3(\boldsymbol{x}_j \otimes \boldsymbol{x}_j)\}.$$

In order to simplify the calculations, we define the following notation:

$$\begin{aligned}
\tau_{kj} &= \frac{(i\phi a_{jk}r_j)^k}{\phi k!},\\
\kappa_{ij} &= \sum_{\alpha=1}^{n}\frac{\partial^2\tau_{2\alpha}}{\partial t_i\partial t_j}\\
&= -\frac{1}{n\phi}\sum_{m=1}^{n} a_{m2}c_{m1}^2(\boldsymbol{e}_i'\boldsymbol{M}_2^{-1}\boldsymbol{x}_m)(\boldsymbol{e}_j'\boldsymbol{M}_2^{-1}\boldsymbol{x}_m), && \text{(A.2.2)}\\
\kappa_{i,kl} &= \sum_{\alpha=1}^{n}\frac{\partial^2\tau_{2\alpha}}{\partial t_i\partial t_{jk}^{(2)}}\\
&= \frac{1}{n\phi}\sum_{m=1}^{n} a_{m2}c_{m1}c_{m2}(\boldsymbol{e}_i'\boldsymbol{M}_2^{-1}\boldsymbol{x}_m)(\boldsymbol{e}_k'\boldsymbol{x}_m)(\boldsymbol{e}_l'\boldsymbol{x}_m), && \text{(A.2.3)}\\
\kappa_{ik,jl} &= \sum_{\alpha=1}^{n}\frac{\partial^2\tau_{2\alpha}}{\partial t_{ik}^{(2)}\partial t_{jl}^{(2)}}\\
&= -\frac{1}{n\phi}\sum_{m=1}^{n} a_{m2}c_{m2}^2(\boldsymbol{e}_i'\boldsymbol{x}_m)(\boldsymbol{e}_k'\boldsymbol{x}_m)(\boldsymbol{e}_j'\boldsymbol{x}_m)(\boldsymbol{e}_l'\boldsymbol{x}_m), && \text{(A.2.4)}\\
\kappa_{i,ijk} &= \sum_{\alpha=1}^{n}\frac{\partial^2\tau_{2\alpha}}{\partial t_i\partial t_{ijk}^{(3)}}\\
&= \frac{1}{n\phi}\sum_{m=1}^{n} a_{m2}c_{m1}c_{m3}(\boldsymbol{e}_i'\boldsymbol{M}_2^{-1}\boldsymbol{x}_m)(\boldsymbol{e}_i'\boldsymbol{x}_m)(\boldsymbol{e}_j'\boldsymbol{x}_m)(\boldsymbol{e}_k'\boldsymbol{x}_m). && \text{(A.2.5)}
\end{aligned}$$

By using the derivations of $\varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3}$, the first term of (A.2.1) is given by

$$\begin{aligned}
&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] \\
&= \sum_{i,j,k,l}^{p} (\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3)_{ijkl}\mathrm{E}[b_{i1}b_{j1}b_{k1}b_{l1}] \\
&= \sum_{i,j,k,l}^{p} (\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3)_{ijkl} \left.\frac{\partial^4 \varphi_{\boldsymbol{b}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3}(\boldsymbol{t},\boldsymbol{T}_2,\boldsymbol{T}_3)}{\partial t_i \partial t_j \partial t_k \partial t_l}\right|_{\boldsymbol{t}=\boldsymbol{0}_p,\boldsymbol{T}_2=\boldsymbol{O},\boldsymbol{T}_3=\boldsymbol{O}}.
\end{aligned}$$

By applying a Taylor expansion, we obtain

$$\begin{aligned}
&\frac{\partial^4}{\partial t_i \partial t_j \partial t_k \partial t_l} \exp\left[\sum_{j=1}^{n}\left\{\frac{b(\theta_j + ir_j\phi) - b(\theta_j)}{\phi} - ia_{j1}r_j\right\}\right]\Bigg|_{\boldsymbol{t}=\boldsymbol{0}_p,\boldsymbol{T}_2=\boldsymbol{O},\boldsymbol{T}_3=\boldsymbol{O}} \\
&= \frac{\partial^4}{\partial t_i \partial t_j \partial t_k \partial t_l} \exp\left\{\sum_{j=1}^{n}(\tau_{2j} + \tau_{3j} + \tau_{4j}) + O(n^{-3/2})\right\}\Bigg|_{\boldsymbol{t}=\boldsymbol{0}_p,\boldsymbol{T}_2=\boldsymbol{O},\boldsymbol{T}_3=\boldsymbol{O}} \\
&= \left\{\kappa_{ij}\kappa_{kl} + \kappa_{ik}\kappa_{jl} + \kappa_{jk}\kappa_{il} + \sum_{\alpha=1}^{n}\frac{\partial^4 \tau_{4\alpha}}{\partial t_i \partial t_j \partial t_k \partial t_l} + O(n^{-2/3})\right\}\exp\{1 + O(n^{-3/2})\}.
\end{aligned}$$

Note that $|r_j| = O(n^{-1/2})$ and

$$\frac{\partial^4 \tau_{4\alpha}}{\partial t_i \partial t_j \partial t_k \partial t_l} = \sum_{\alpha=1}^{n}\phi^3 a_{\alpha 4}\frac{\partial r_\alpha}{\partial t_i}\frac{\partial r_\alpha}{\partial t_j}\frac{\partial r_\alpha}{\partial t_k}\frac{\partial r_\alpha}{\partial t_l} = O(n^{-1}).$$

Hence, the first term of (A.2.1) is expressed as

$$\begin{aligned}
&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] \\
&= \sum_{i,j,k,l}^{p} (\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3)_{ijkl}(\kappa_{ij}\kappa_{kl} + \kappa_{ik}\kappa_{jl} + \kappa_{jk}\kappa_{il}) + O(n^{-1}). \qquad \text{(A.2.6)}
\end{aligned}$$

By substituting (A.2.2) into (A.2.6), we obtain

$$\begin{aligned}
&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{M}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] \\
&= \frac{1}{n^2\phi^2}\sum_{i,j}^{n}(a_{i2}c_{i1}^3 + 3a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}^3 + 3a_{j2}c_{j1}c_{j2}) \qquad \text{(A.2.7)} \\
&\qquad \times (u_{ii}u_{ij}u_{jj} + 2u_{ij}^3) + O(n^{-1}).
\end{aligned}$$

The remaining terms of (A.2.1), as well as the first term of (A.2.1), will be calculated. The second term of (A.2.1) is similarly obtained from (A.2.7) as follows:

$$\begin{aligned}&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_4(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)]\\ &= -\frac{3}{n\phi}\sum_{i=1}^{n}(a_{i4}c_{i1}^4 + 6a_{i3}c_{i1}^2c_{i2} + 3a_{i2}c_{i2}^2 + 4a_{i2}c_{i1}c_{i3})u_{ii}^2 + O(n^{-1}).\end{aligned} \tag{A.2.8}$$

Next, we calculate the third term of (A.2.1). The third term of (A.2.1) is expressed as follows:

$$\begin{aligned}&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1]\\ &= \sum_{i,j,k,l}^{p}(\boldsymbol{M}_3'\boldsymbol{M}_2^{-1})_{ijk}\mathrm{E}[b_{1i}b_{1j}b_{1l}\boldsymbol{Z}_{2,kl}]\\ &= \sum_{i,j,k,l}^{p}(\boldsymbol{M}_3'\boldsymbol{M}_2^{-1})_{ijk}(\kappa_{ij}\kappa_{l,kl} + \kappa_{ik}\kappa_{j,kl} + \kappa_{jk}\kappa_{i,kl}) + O(n^{-1}).\end{aligned}$$

Expression (A.2.3) implies that

$$\begin{aligned}&\mathrm{E}[(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)'\boldsymbol{M}_3'\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1]\\ &= -\frac{1}{n^2\phi^2}\sum_{i,j}^{n}(a_{i3}c_{i1}^3 + 3a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}c_{j2})(u_{ii}u_{ij}u_{jj} + 2u_{ij}^3) + O(n^{-1}).\end{aligned} \tag{A.2.9}$$

The fourth term of (A.2.1) is as follows:

$$\begin{aligned}&\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_2\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1]\\ &= \sum_{i,j,k,l}^{p}(\boldsymbol{M}_2^{-1})_{jk}(\kappa_{il}\kappa_{ik,jl} + \kappa_{i,ij}\kappa_{l,kl} + \kappa_{i,kl}\kappa_{l,ij}) + O(n^{-1}).\end{aligned} \tag{A.2.10}$$

It follows from (A.2.4) and (A.2.10) that

$$\begin{aligned}&\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_2\boldsymbol{M}_2^{-1}\boldsymbol{Z}_2\boldsymbol{b}_1]\\ &= \frac{1}{n^2\phi^2}\sum_{i,j}^{n}(a_{i2}c_{i1}c_{i2})(a_{j2}c_{j1}c_{j2})(u_{ii}u_{ij}u_{jj} + u_{ij}^3)\\ &\quad - \frac{1}{n\phi}\sum_{i=1}^{n}a_{i2}c_{i2}^2u_{ii} + O(n^{-1}).\end{aligned} \tag{A.2.11}$$

Finally, we calculate the fifth term of (A.2.1). Note that

$$\begin{aligned}
&\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] \\
&= \sum_{i,j,k}^{p} \mathrm{E}[\boldsymbol{Z}_{3,ijk} b_{1i} b_{1j} b_{1k}] \\
&= \sum_{i,j,k}^{p} (\kappa_{ij}\kappa_{k,ijk} + \kappa_{ik}\kappa_{j,ijk} + \kappa_{jk}\kappa_{i,ijk}) + O(n^{-1}). \qquad \text{(A.2.12)}
\end{aligned}$$

Substituting (A.2.5) into (A.2.12) yields

$$\mathrm{E}[\boldsymbol{b}_1'\boldsymbol{Z}_3(\boldsymbol{b}_1 \otimes \boldsymbol{b}_1)] = \frac{3}{n}\sum_{i=1}^{n} a_{i2}c_{i1}c_{i3}u_{ii}^2 + O(n^{-1}). \qquad \text{(A.2.13)}$$

Consequently, from (A.2.7), (A.2.8), (A.2.9), (A.2.11), and (A.2.13), we obtain the fourth term of (2.2.4) as (2.2.9).

A.3 Explicit Forms of (3.1.5), (3.1.6) and (3.1.7)

In this section, 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

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

where $\boldsymbol{e}_j$ is the jth coordinate unit vector, which is used in equation (3.1.8). This result and equation (3.1.2) imply that

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

Substituting the above result into the definition of $\boldsymbol{G}_2(\boldsymbol{\beta})$ yields equation (3.1.5). 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 (3.1.6) and (3.1.7), 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:

$$\begin{aligned}
\boldsymbol{\Delta}_{3,i}(\boldsymbol{\beta}) &= \sum_{a,b}^{r} \frac{\partial}{\partial \boldsymbol{\beta}'}(p_{ia}\delta_{ab} - p_{ia}p_{ib}) \otimes \boldsymbol{e}_a\boldsymbol{e}_b', \\
\boldsymbol{\Delta}_{4,i}(\boldsymbol{\beta}) &= \sum_{a,b}^{r} \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(p_{ia}\delta_{ab} - p_{ia}p_{ib}) \otimes \boldsymbol{e}_a\boldsymbol{e}_b'.
\end{aligned} \tag{A.3.1}$$

Derivatives of p_{ia} are calculated as

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

By substituting the above derivatives into (A.3.1), we have

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

and

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

where $\boldsymbol{q}_{i,a} = \boldsymbol{e}_a - \boldsymbol{p}_i$. The above two equations indicate that explicit forms of $\boldsymbol{G}_3(\boldsymbol{\beta})$ and $\boldsymbol{G}_4(\boldsymbol{\beta})$ are given in (3.1.6) and (3.1.7), respectively.

A.4 Expectations of Derivatives of the Negative Log-Likelihood Function

In this section, we derive general formulas of the expectations of derivatives of the negative 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 negative 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

$$\begin{aligned}&\dot{L}_a = -\frac{\dot{f}_a}{f}, \quad \dot{L}_{ab} = \dot{L}_a\dot{L}_b - \frac{\dot{f}_{ab}}{f}, \quad \dot{L}_{abc} = -\dot{L}_a\dot{L}_b\dot{L}_c + \sum_{[3]}\dot{L}_a\dot{L}_{bc} - \frac{\dot{f}_{abc}}{f}, \\ &\dot{L}_{abcd} = \dot{L}_a\dot{L}_b\dot{L}_c\dot{L}_d - \sum_{[6]}\dot{L}_a\dot{L}_b\dot{L}_{cd} + \sum_{[3]}\dot{L}_{ab}\dot{L}_{cd} + \sum_{[4]}\dot{L}_a\dot{L}_{bcd} - \frac{\dot{f}_{abcd}}{f},\end{aligned} \tag{A.4.1}$$

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

$$\begin{aligned}
\mathrm{E}\left[\frac{\dot{f}_{a_1\cdots a_j}}{f}\right] &= \int \dot{f}_{a_1\cdots a_j} d\boldsymbol{u} \\
&= \int \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} f d\boldsymbol{u} \qquad \text{(A.4.2)} \\
&= \frac{\partial^j}{\partial\theta_{a_1}\cdots\partial\theta_{a_j}} \int f d\boldsymbol{u} = 0.
\end{aligned}$$

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.4.1) and (A.4.2) imply that

$$\mathrm{E}[\dot{L}_a] = 0, \quad \mathrm{E}[\dot{L}_{ab}] = \mathrm{E}[\dot{L}_a\dot{L}_b], \quad \mathrm{E}[\dot{L}_{abc}] = -\mathrm{E}[\dot{L}_a\dot{L}_b\dot{L}_c] + \sum_{[3]} \mathrm{E}[\dot{L}_a\dot{L}_{bc}],$$

$$\mathrm{E}[\dot{L}_{abcd}] = \mathrm{E}[\dot{L}_a\dot{L}_b\dot{L}_c\dot{L}_d] - \sum_{[6]} \mathrm{E}[\dot{L}_a\dot{L}_b\dot{L}_{cd}] + \sum_{[3]} \mathrm{E}[\dot{L}_{ab}\dot{L}_{cd}] + \sum_{[4]} \mathrm{E}[\dot{L}_a\dot{L}_{bcd}]. \qquad \text{(A.4.3)}$$

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

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

From the expectations in (A.4.3), we obtain $\mathrm{E}[\boldsymbol{H}(\boldsymbol{\theta})]$, $\mathrm{E}[\boldsymbol{C}(\boldsymbol{\theta})]$, and

$\mathrm{E}[\boldsymbol{Q}(\boldsymbol{\theta})]$ as

$$\begin{aligned}
\mathrm{E}[\boldsymbol{H}(\boldsymbol{\theta})] &= \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \\
\mathrm{E}[\boldsymbol{C}(\boldsymbol{\theta})] &= -\mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] + \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{H}(\boldsymbol{\theta})] \\
&\quad + \mathrm{E}[\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{g}(\boldsymbol{\theta})'] + \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'], \\
\mathrm{E}[\boldsymbol{Q}(\boldsymbol{\theta})] &= \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] \\
&\quad - (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)\mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{H}(\boldsymbol{\theta})](\boldsymbol{I}_{q^2} + \boldsymbol{K}_q) \\
&\quad - \mathrm{E}[\mathrm{vec}(\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})')\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'] - \mathrm{E}[\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})')'] \\
&\quad + (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)\mathrm{E}[\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{H}(\boldsymbol{\theta})] + \mathrm{E}[\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'] \\
&\quad + \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta}) \otimes \boldsymbol{C}(\boldsymbol{\theta})](\boldsymbol{I}_{q^2} + \boldsymbol{K}_q) + (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)\mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{C}(\boldsymbol{\theta})'].
\end{aligned} \tag{A.4.4}$$

Recall that $\mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})] = \boldsymbol{0}_q$ 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 $\boldsymbol{Q}(\boldsymbol{\theta})$ in (A.4.4) become simpler, as follows:

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{\theta}) &= \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \quad \boldsymbol{C}(\boldsymbol{\theta}) = -\mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'], \\
\boldsymbol{Q}(\boldsymbol{\theta}) &= \mathrm{E}[\boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})' \otimes \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})'] - (\boldsymbol{I}_{q^2} + \boldsymbol{K}_q)\{\boldsymbol{H}(\boldsymbol{\theta}) \otimes \boldsymbol{H}(\boldsymbol{\theta})\} \\
&\quad - \mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))\mathrm{vec}(\boldsymbol{H}(\boldsymbol{\theta}))'.
\end{aligned} \tag{A.4.5}$$

Bibliography

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory*, eds. Petrov, B. N. & Csáki, F., 267–281, Akadémiai Kiadó, Budapest.

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

Barnett, S. B. L. & Nurmagambetov, T. A. (2010). Cost of Asthma in the United States: 2002-2007. *J. Allergy Clin. Immun.*, **127**, 145–152.

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

Brown, B. W. (1980). Prediction Analysis for Binary Data. In *Biostatistics Casebook*. (eds. Miller Jr., G. R., Efron., B., Brown, B. W., & Moses, L. E.), 3–18, Wiley, New York.

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

Chen, J. & Lazar, N. A. (2012). Selection of working correlation structure in generalized estimating equations via empirical likelihood. *J. Comput. Graph. Statist.*, **21**, 18–41.

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.

Crowder, M. (1985). Gaussian Estimation for Correlated Binomial Data. *J. R. Statist. Soc. B.*, **47**, 229–237.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, **82**, 407–410.

Davies, S. L., Neath, A. A. & Cavanaugh, J. E. (2006). Estimation Optimality of Corrected AIC and Modified *Cp* in Linear Regression. *Internat. Statist. Rev.*, **74**, 2, 161–168.

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

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models. *Ann. Statist.*, **13**, 342–368.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, **51**, 309–317.

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

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J. R. Statist. Soc. B.*, **41**, 190–195.

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

Hin, L.-Y., Carey, V. J. & Wang, Y.-G. (2007). Criteria for working-correlation-structure selection in GEE: Assessment via simulation. *Amer. Statistician*, **61**, 360–364.

Hin, L.-Y. & Wang, Y.-G. (2009). Working-correlation-structure identification in generalized estimating equations. *Statist. Med.*, **28**, 642–658.

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

Hurvich, C. M. & Tsai, C. L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, **76**, 297–307.

Imori, S. (2014). Consistent Selection of Working Correlation Structure in GEE Analysis Based on Stein′s Loss Function. *Hiroshima Math. J.*, to appear.

Imori, S. (2013). On Properties of QIC in Generalized Estimating Equations. *TR 13-01, Statistical Research Group, Hiroshima University, Hiroshima.*

Imori, S., Yanagihara, H. & Wakaki, H. (2014). Simple Formula for Calculating Bias-Corrected AIC in Generalized Linear Models. *Scand. J. Stat.*, **41**, 535–555.

Inatsu, Y. & Imori, S. (2013). Model Selection Criterion Based on the Prediction Mean Squared Error in Generalized Estimating Equations. *TR 13-10, Statistical Research Group, Hiroshima University, Hiroshima.*

James, W. & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, **1**, 361–379.

Kamo, K., Yanagihara, H. & Satoh, K. (2013). Bias-Corrected AIC for Selecting Variables in Poisson Regression Models. *Comm. Statist. Theory Methods*, **42**, 1911–1921.

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

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

Kullback, S. & Leibler, R. (1951). On Information and Sufficiency. *Ann. Math. Statist.*, **22**, 79–86.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika.*, **73**, 13–22.

Matas, M., Picornell, A., Cifuentes, C., Payeras, A., Bassa, A., Homar, F., López-Labrador, F. X., Moya, A., Ramon, M. M. & Castro, J. A. (2010). Relating the Liver Damage with Hepatitis C Virus Polymorphism in Core Region and Human Variables in HIV-1-Coinfected Patients. *Infect. Genet. Evol.*, **10**, 1252–1261.

McCullagh, P. & Cox, D. R. (1986). Invariants and likelihood ratio statistics, *Ann. Statist.*, **14**, 1419–1430.

McCullagh, P. & Nelder, J. A. (1989). Generalized linear models, 2nd edition. Chapman and Hall, London.

Meyers, R. H., Montgomery, D. C. & Vining, G. G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences.* Wiley Interscience, Canada.

Nelder, J. A. & Wedderburn, W. M. (1972). Generalized Linear Models. *J. R. Statist. Soc. A.*, **135**, 370–384.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics.*, **57**, 120–125.

Pan, W. & Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Statist. Sinica*, **12**, 475–490.

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/.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221–264.

Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *Ann. Math. Statist.*, **8**, 147–164.

Sugiura, N. (1978). Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections. *Commun. Statist. -Theory Meth.*, **7**, 1, 13–26.

Sánchez-Carneo, N., Couñago, E., Rodrigues-Perez, D. & Freire, J. (2011). Exploiting Oceanographic Satellite Data to Study the Small Scale Coastal Dynamics in a NE Atlantic Open Embayment. *J. Marine Syst.*, **87**, 123–132.

Teste, F. P. & Lieffers, V. J. (2011). Snow Damage in Lodgepole Pine Stands Brought into Thinning and Fertilization Regimes. *Forest Ecol. Manag.*, **261**, 2094–2104.

Wang, Y.-G. & Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika*, **90**, 29–41.

Wong, C. S. & Li, W. K. (1998). A Note on the Corrected Akaike Information Criterion for Threshold Autoregressive Models. *J. Time Ser. Anal.*, **19**, 113–124.

Yanagihara, H., Kamo, K., Imori, S. & Satoh, K. (2012). Bias-Corrected AIC for Selecting Variables in Multinomial Logistic Regression Models. *Linear Algebra Appl.*, **436**, 4329–4341.

Yanagihara, H., Sekiguchi, R. & Fujikoshi, Y. (2003). Bias Correction of AIC in Logistic Regression Models. *J. Statist. Plann. Inference*, **115**, 349–360.