HIROSHIMA MATH. J.
45 (2015), 91–107

Consistent selection of working correlation structure in GEE analysis based on Stein's loss function

Shinpei IMORI

(Received April 16, 2014)
(Revised June 11, 2014)

ABSTRACT. In this paper, we attempt to select a working correlation structure for generalized estimating equations. We propose a selection criterion based on the Stein's loss function. Our criterion consistently selects the true correlation structure when the unknown parameters are sqrt(n)-consistent, where n is the sample size. We demonstrate the performance of the proposed methodology by a numerical study.

1. Introduction

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 and Zeger [13] for estimating regression coefficients in such correlated data; it is an expansion of the likelihood equation in the generalized linear model (GLM) that was proposed by Nelder and Wedderburn [14]. 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 (see, [13]). However, some studies have noted that a misspecification of the working correlation structure may induce undesirable results. For instance, Crowder [4] 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 [6] showed that a GEE estimator is less efficient when the independent structure is assumed to the working correlation matrix. Thus, it is important to adequately determine the working correlation structure,

2010 *Mathematics Subject Classification*. Primary 62F07; Secondary 62J12.

Key words and phrases. covariance structure selection, generalized estimating equations, Stein's loss function, 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 the 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 the working correlation structure. The Akaike information criterion (AIC) derived in [1] and the Bayesian information criterion (BIC) derived in [19] are often used to select the best model, due to their theoretical validity (see [15] or [18] for example). For example, the BIC and its generalization, the GIC derived in [15], can be used to select the true model since their selection probabilities of the true model goes to 1, which is called the consistency. Information criteria are typically based on the maximum log-likelihood and some penalty terms. Since the GEE approach does not assume a joint distribution of responses, Pan [16] 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) was derived in [10] 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 [17] attempted to select the working correlation structure that minimizes the mean squared prediction error estimated by a resampling method. Hin, et al. [9] proposed a criterion based on a measurement between the true correlation and the candidate correlation structure. Chen and Lazar [2] 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.

The primary aim of the present paper is to 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 derived in [12], which is one of the famous loss function for matrices. Moreover, we can show the consistency property of our criterion.

The present article is organized as follows: in Section 2, we introduce the GEE; in Section 3, we propose a criterion for selecting the true correlation structure; in Section 4, we derive its asymptotic behavior; in Section 5, we demonstrate the performance of our criterion with finite samples by presenting a numerical study; in Section 6, we present a discussion and our conclusions.

2. 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.,

$$f(y_{ij}; \theta_{ij}, \phi) = \exp[\phi^{-1}\{\theta_{ij} y_{ij} - a(\theta_{ij})\} + b(y_{ij}, \phi)],$$

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 [5]. A 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

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

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 the working correlation structure are

$$\begin{aligned} &\text{Independent (Indep.)} : \boldsymbol{R} = \boldsymbol{I}_m, \\ &\text{Exchangeable (Ex.)} \quad : (\boldsymbol{R})_{jk} = \alpha, \\ &\text{AR} - 1 \qquad\qquad\quad : (\boldsymbol{R})_{jk} = \alpha^{|j-k|}, \\ &\text{Unstructured (Unst.)} : (\boldsymbol{R})_{jk} = \alpha_{jk}, \end{aligned} \tag{2}$$

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

$$\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0}_p, \tag{3}$$

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 (1).

Denote $\hat{\boldsymbol{\beta}}(\boldsymbol{R})$ as the GEE estimator with $\boldsymbol{R}$, which is given by solving (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 [21]. In Section 5, we estimate $\boldsymbol{\alpha}$ by using a moment-based method.

3. 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 (2). Examples with (2) are illustrated in Section 5. 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 the over-fitted set $\mathcal{M}^+$ and the under-fitted set $\mathcal{M}^-$, i.e.,

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

where $\mathcal{K}$ is the parameter space, which is a compact set and $\mathcal{M}^- = \mathcal{M} \backslash \mathcal{M}^+$. For all $\boldsymbol{R} \in \mathcal{M}^+$, we assume that there exists $\boldsymbol{\alpha} \in \mathcal{K}^0$ such that $\boldsymbol{R}(\boldsymbol{\alpha}) = \boldsymbol{R}_*$, where $\mathcal{K}^0$ 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:

$$SL_n(\boldsymbol{R}) = n \log \det(\boldsymbol{R}) + n \,\mathrm{tr}(\hat{\boldsymbol{R}}_U \boldsymbol{R}^{-1}), \tag{4}$$

where

$$(\hat{\boldsymbol{R}}_U)_{jk} = \begin{cases} \hat{\phi}^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_{ij} \hat{\varepsilon}_{ik} / n, & j \neq k, \\ 1, & j = k, \end{cases}$$
$$\hat{\boldsymbol{\varepsilon}}_i = (\hat{\varepsilon}_{i1}, \ldots, \hat{\varepsilon}_{im})' = \hat{\boldsymbol{A}}_i^{-1/2} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i). \tag{5}$$

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 [3] and Wang and Carey [21] 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) instead of the likelihood for $\boldsymbol{y}_i$, we consider a GIC-type criterion as follows.

DEFINITION 2. For a working correlation structure $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\alpha}) \in \mathscr{M}$, the GIC-type criterion is

$$GIC_{\gamma_n}(\boldsymbol{R}) = SL_n(\hat{\boldsymbol{R}}) + q\gamma_n, \tag{6}$$

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) and (6), we have neither specified the working correlation structure for estimating the GEE estimator $\hat{\boldsymbol{\beta}}$ nor the way how to estimate $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\alpha}}$.

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

DEFINITION 3. The best correlation structure selected by the GIC proposed in (6) is

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

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

4. Properties of criteria

In this section, we show the consistency of the GIC proposed in (6). Suppose that the mean structure has been correctly specified. The proof can then be obtained in a way similar to that in [15]. The following assumptions are sufficient conditions for the consistency of the GIC:

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

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

(C3) For all $\boldsymbol{R} \in \mathscr{M}^+$, $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = O_p(n^{-1/2})$ 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

$$\hat{\phi} = \sum_{i=1}^{n} \hat{\boldsymbol{\varepsilon}}_i' \hat{\boldsymbol{\varepsilon}}_i / (nm - p), \tag{7}$$

where $\hat{\boldsymbol{\varepsilon}}_i$ is defined in (5), then it follows from [13] 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 1. *Under the conditions* (C1)–(C3), *for all* $\boldsymbol{R} \in \mathscr{M}^+ \backslash \{\boldsymbol{R}_0\}$, *when* $\gamma_n \to \infty$,

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

PROOF OF THEOREM 1. Denote q and q_* as the number of elements of correlation parameter for $\boldsymbol{R}$ and $\boldsymbol{R}_0$, respectively. From Definition 3, the selection probability of $\boldsymbol{R}$ is

$$\begin{aligned} Pr(\boldsymbol{R}_{best} = \boldsymbol{R}) &\leq Pr\{GIC_{\gamma_n}(\boldsymbol{R}_0) > GIC_{\gamma_n}(\boldsymbol{R})\} \\ &= Pr\{SL_n(\hat{\boldsymbol{R}}_0) - SL_n(\hat{\boldsymbol{R}}) > (q - q_*)\gamma_n\}. \end{aligned} \tag{8}$$

We evaluate $SL_n(\hat{\boldsymbol{R}}_0)$ and $SL_n(\hat{\boldsymbol{R}})$. Under the conditions (C1)–(C3), for all $\boldsymbol{R} \in \mathscr{M}^+$, 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

$$\hat{\boldsymbol{R}} - \boldsymbol{R}_* = O_p(n^{-1/2}). \tag{9}$$

On the contrary, let

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

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 [5]. 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,

$$\hat{\boldsymbol{R}}_U = \boldsymbol{R}_* + O_p(n^{-1/2}). \tag{10}$$

From (9) and (10), $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{R}}_U^{-1/2}\hat{\boldsymbol{R}}\hat{\boldsymbol{R}}_U^{-1/2} - \boldsymbol{I}_m = O_p(n^{-1/2})$. 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 $\ell = 1, \ldots, m$,

$$\log\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\} = \lambda_\ell(\hat{\boldsymbol{\Omega}}) - \lambda_\ell(\hat{\boldsymbol{\Omega}})^2/2 + O_p(n^{-3/2}),$$

$$\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\}^{-1} = 1 - \lambda_\ell(\hat{\boldsymbol{\Omega}}) + \lambda_\ell(\hat{\boldsymbol{\Omega}})^2 + O_p(n^{-3/2}).$$

Hence,

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

$$\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} = \sum_{\ell=1}^{m}\{1 + \lambda_\ell(\hat{\boldsymbol{\Omega}})\}^{-1} = m - \mathrm{tr}(\hat{\boldsymbol{\Omega}}) + \mathrm{tr}(\hat{\boldsymbol{\Omega}}^2) + O_p(n^{-3/2}).$$

By substituting above results into (4),

$$\begin{aligned} SL_n(\hat{\boldsymbol{R}}) &= n\log\det(\hat{\boldsymbol{R}}) + n\,\mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}^{-1}) \\ &= n\log\det(\hat{\boldsymbol{R}}_U) + n\log\det(\hat{\boldsymbol{R}}\hat{\boldsymbol{R}}_U^{-1}) + n\,\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} \\ &= n\log\det(\hat{\boldsymbol{R}}_U) + n\log\det(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}}) + n\,\mathrm{tr}\{(\boldsymbol{I}_m + \hat{\boldsymbol{\Omega}})^{-1}\} \\ &= n\log\det(\hat{\boldsymbol{R}}_U) + nm + n\,\mathrm{tr}(\hat{\boldsymbol{\Omega}}^2)/2 + O_p(n^{-1/2}) \\ &= n\log\det(\hat{\boldsymbol{R}}_U) + nm + O_p(1). \end{aligned} \tag{11}$$

It follows from (11) 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 (8), since $\gamma_n \to \infty$ as $n \to \infty$, then

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

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

THEOREM 2. *Under the conditions* (C1)–(C3), *for all* $\boldsymbol{R} \in \mathscr{M}^-$, *when* $\gamma_n/n \to 0$,

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

PROOF OF THEOREM 2. As in (8), the selection probability of $\boldsymbol{R} \in \mathscr{M}^-$ is evaluated as

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

where q and q_* are the number of elements in $\boldsymbol{R} \in \mathscr{M}^-$ 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:

$$\begin{aligned}
&SL_n(\hat{\boldsymbol{R}}_0)/n - SL_n(\hat{\boldsymbol{R}})/n \\
&\quad = -\log\det(\hat{\boldsymbol{R}}_0^{-1}) + \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) + \log\det(\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}^{-1}) \\
&\quad = -\log\det(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) + \mathrm{tr}(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) - m + \log\det(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) \\
&\qquad - \mathrm{tr}(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + m + \log\det(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) - \mathrm{tr}(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) + m \\
&\qquad - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \\
&\quad = \rho(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) - \rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \rho(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) \\
&\qquad - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\}. \qquad (12)
\end{aligned}$$

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

$$\rho(\hat{\boldsymbol{R}}_U\hat{\boldsymbol{R}}_0^{-1}) = o_p(1), \qquad \rho(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1}) = o_p(1). \qquad (13)$$

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

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

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

The last inequality is established from the fact that the logarithm is concave. Hence, $\rho(t\boldsymbol{A} + (1-t)\boldsymbol{B}) \leq t\rho(\boldsymbol{A}) + (1-t)\rho(\boldsymbol{B}) \leq L$ holds, and then $t\boldsymbol{A} + (1-t)\boldsymbol{B} \in \mathscr{A}_L$. Therefore, $\mathscr{A}_L$ 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 $\rho(\boldsymbol{A}_{[1]}) = \rho(\hat{\boldsymbol{R}}^{-1}\boldsymbol{R}_*) \geq c$. Since for all $L > 0$, $\mathscr{A}_L$ is the convex set and $\rho(\cdot)$ is continuous on $\mathscr{A}_L$, 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

$$\begin{aligned} & \rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \\ & \quad = \{g(1) - g(0)\} \geq \{g(t) - g(0)\}/t \geq g(t) \\ & \quad = c + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{A}_{[t]} - \boldsymbol{I}_m)\} \\ & \quad \geq c - \sqrt{\mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)^2\}\ \mathrm{tr}\{(\boldsymbol{A}_{[t]} - \boldsymbol{I}_m)^2\}} \\ & \quad \geq c - \sqrt{\mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)^2\}b} \end{aligned} \tag{14}$$

where

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

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 (14), it is established that

$$\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) + \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} \geq c - c/2 = c/2. \tag{15}$$

On the other hand, we can see that

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

Thereby, it follows from (15) that

$$Pr\{-\rho(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1}) - \mathrm{tr}\{(\hat{\boldsymbol{R}}_U\boldsymbol{R}_*^{-1} - \boldsymbol{I}_m)(\boldsymbol{R}_*\hat{\boldsymbol{R}}^{-1} - \boldsymbol{I}_m)\} < -c/2\}$$
$$\geq 1 - Pr(E^c) \to 1, \tag{16}$$

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

From (12), (13), (16), and $(q - q_*)\gamma_n/n \to 0$, for all $\boldsymbol{R} \in \mathscr{M}^-$,

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

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

THEOREM 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 3. The probability of the true correlation structure selection is divided into two parts, as follows:

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

From Theorem 1 and Theorem 2, it follows that

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

5. Numerical study

In this section, we present a numerical study 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, Hannan and Quinn's IC(HQIC)-type proposed in [7], and BIC-type tuning parameters for the GIC proposed in (6). 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:

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

$$CIC(\boldsymbol{R}) = \mathrm{tr}(\hat{\boldsymbol{V}}_s \hat{\boldsymbol{\Sigma}}_I),$$

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

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

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

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 (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 (7), 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 $\mathscr{M}$ was considered in the following case, introduced in (2):

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

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

$$\text{Indep.} : \boldsymbol{R}_0 = \boldsymbol{I}_4,$$

$$\text{Ex.} : \boldsymbol{R}_0 = \boldsymbol{I}_4/2 + \boldsymbol{1}_4\boldsymbol{1}_4'/2, \qquad \text{where } \boldsymbol{1}_4 = (1, 1, 1, 1)',$$

$$\text{AR-1} : (\boldsymbol{R}_0)_{jk} = 2^{-|j-k|},$$

$$\text{Unst.} : \boldsymbol{R}_0 = \boldsymbol{H}_d^{-1/2} \boldsymbol{H} \boldsymbol{H}_d^{-1/2}, \qquad \text{where } \boldsymbol{H} = (h_{ij})_{1 \le i,j \le 4} = \boldsymbol{W}'\boldsymbol{W} + \boldsymbol{I}_4,$$

$$(\boldsymbol{W})_{jk} \overset{i.i.d.}{\sim} U(-1, 1) \qquad \text{and} \qquad \boldsymbol{H}_d = \mathrm{diag}(h_{11}, \ldots, h_{44}).$$

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

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

$$\text{AR-1} : \hat{\alpha} = \sum_{i=1}^{n} \sum_{j=1}^{m-1} \hat{\varepsilon}_{ij}\hat{\varepsilon}_{i,j+1}/\{n(m-1)\},$$

$$\text{Unst.} : \hat{\alpha}_{jk} = \sum_{i=1}^{n} \hat{\varepsilon}_{ij}\hat{\varepsilon}_{ik}/n.$$

Note that the conditions (C1)–(C3) held in this simulation setting. The BIC satisfied the assumptions of Theorem 3 but the AIC satisfies only the assumption of Theorem 2. For the situations described above, we simulated 1,000 repetitions.

In this numerical study, 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

$$\text{PMSE} : \frac{1}{1000}\sum_{\ell=1}^{1000}\sum_{i=1}^{n}\{\hat{\boldsymbol{\mu}}_{i,best}^{(\ell)} - \boldsymbol{\mu}_i\}' \boldsymbol{V}_i^{-1}\{\hat{\boldsymbol{\mu}}_{i,best}^{(\ell)} - \boldsymbol{\mu}_i\},$$

$$\text{VAR} : \frac{1}{1000}\sum_{\ell=1}^{1000}\left|\hat{\boldsymbol{\beta}}_{best}^{(\ell)} - \frac{1}{1000}\sum_{\ell=1}^{1000}\hat{\boldsymbol{\beta}}_{best}^{(\ell)}\right|^2,$$

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 $\boldsymbol{\beta}$, respectively, with using the best correlation structure in the ℓth iteration.

Tables 1–4 listed the results of the selection probability and the ratios of the PMSE and VAR to the values of the BIC. From Tables 1–4, 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 the 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

Table 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

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

6. Discussion

In this paper, we proposed a GIC-type criterion based on Stein's loss function (the discrepancy between the true correlation structure and a working correlation structure) in order to select the true correlation structure, and we derived sufficient conditions for its consistency. 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 semiparametric frameworks. Moreover, it may be

Table 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

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 approach. 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

Table 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

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. 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 [8, 11, 20]. We expect that these approaches will yield more adequacy criteria or assumptions for selecting the true correlation structure.

Acknowledgement

The author is grateful to the referee for the helpful suggestion and the valuable advice. In addition, I would like to thank Professor Hirofumi Wakaki of Hiroshima University for his various comments.

Table 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

References

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

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

[3] M. Crowder, Gaussian estimation for correlated binomial data. J. R. Statist. Soc. B., **47** (1985), 229–237.

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

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

[6] G. M. Fitzmaurice, A caveat concerning independence estimating equations with multivariate binary data. Biometrics, **51** (1995), 309–317.
[7] E. J. Hannan, & B. G. Quinn, The determination of the order of an autoregression. J. R. Statist. Soc. B., **41** (1979), 190–195.
[8] C. M. Hurvich, & C. L. Tsai, Regression and time series model selection in small samples. Biometrika, **76** (1989), 297–307.
[9] L.-Y. Hin, V. J. Carey, & Y.-G. Wang, Criteria for working-correlation-structure selection in GEE: Assessment via simulation. Amer. Statistician, **61** (2007), 360–364.
[10] L.-Y. Hin, & Y.-G. Wang, Working-correlation-structure identification in generalized estimating equations. Statist. Med., **28** (2009), 642–658.
[11] S. Imori, H. Yanagihara, & H. Wakaki, Simple Formula for Calculating Bias-Corrected AIC in Generalized Linear Models. Scand. J. Statist., **41** (2014), 535–555.
[12] W. James, & C. Stein, Estimation with quadratic loss. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, **1** (1961), 361–379.
[13] K.-Y. Liang, & S. L. Zeger, Longitudinal data analysis using generalized linear models. Biometrika., **73** (1986), 13–22.
[14] J. A. Nelder, & W. M. Wedderburn, Generalized linear models. J. R. Statist. Soc. A., **135** (1972), 370–384.
[15] R. Nishii, Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist., **12** (1984), 758–765.
[16] W. Pan, Akaike's information criterion in generalized estimating equations. Biometrics., **57** (2001), 120–125.
[17] W. Pan, & J. E. Connett, Selecting the working correlation structure in generalized estimating equations with application to the lung health study. Statist. Sinica, **12** (2002), 475–490.
[18] J. Shao, An asymptotic theory for linear model selection. Statist. Sinica, **7** (1997), 221–264.
[19] G. Schwarz, Estimating the dimension of a model. Ann. Statist., **6** (1978), 461–464.
[20] N. Sugiura, Further analysis of the data by Akaike's information criterion and the finite corrections. Commun. Statist.-Theory Meth., **7** (1978), 1, 13–26.
[21] Y.-G. Wang, & V. Carey, Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. Biometrika, **90** (2003), 29–41.

Shinpei Imori
Graduate School of Engineering Science
Osaka University
1-3 *Machikaneyama-cho*
Toyonaka, Osaka 560-8531, *Japan*
E-mail: imori.stat@gmail.com