HIROSHIMA MATH. J.
48 (2018), 307–334

Model selection criterion based on the prediction mean squared error in generalized estimating equations

Yu INATSU and Shinpei IMORI

(Received January 7, 2017)
(Revised June 5, 2018)

ABSTRACT. The present paper considers a model selection criterion in regression models using generalized estimating equation (GEE). Using the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called PMSEG that reflects the correlation between responses. Numerical studies reveal that the PMSEG has better performance than previous other criteria for model selection.

1. Introduction

In real data analysis, correlated data are often discussed in health sciences, medical sciences, economics and many other fields. Longitudinal data, defined from observations of subjects measured repeatedly over time, often arise in these fields as an important example. In general, observations from each subject in longitudinal data are correlated. Liang and Zeger [11] introduced an extension of the generalized linear model (Nelder and Wedderburn, [13]) to the analysis of longitudinal data, known as the generalized estimating equation (GEE) method. Defining features of the GEE method are that we can use a working (but not necessarily correct) correlation matrix as the correlation matrix, and a full specification of the joint distribution is not required. Hence, the GEE method is widely used in many fields for analyzing longitudinal data.

In addition, the model selection problem in the GEE methodology is also important. The goodness of fit of the model is commonly measured by some risk function, and the model selection is performed by obtaining a certain estimator of the risk function. For example, the risk function based on the expected Kullback-Leibler (KL) information (Kullback and Leibler, [10]) is often used, and the most famous estimator of the risk function is Akaike's information criterion (AIC) proposed by Akaike [1, 2]. The AIC is obtained

2010 *Mathematics Subject Classification*. Primary 62H12; Secondary 62F07.
Key words and phrases. Generalized estimating equations, Longitudinal data, Prediction mean squared error, Model selection.

by using the likelihood, it can be simply defined as AIC $= -2 \times$ (the maximum log likelihood) $+ 2 \times$ (the number of parameters). Furthermore, Nishii [14], Rao [16] proposed the generalized information criterion (GIC) as a general extension of the AIC, which is widely applied for selecting the best model, and considered about various properties in many literatures.

However, in the GEE method, the maximum likelihood based model selection criteria such as the AIC or GIC, are not applicable directly because the GEE method is not likelihood based. Some model selection criteria like the AIC or GIC in the GEE method have been already proposed. For example, Pan [15] proposed the QIC (quasi-likelihood under the independence model criterion) and Imori [9] proposed the modified QIC. These criteria are estimators of the risk function based on the quasi likelihood (Wedderburn, [17]). Cantoni *et al.* [4] proposed the GC_p (generalized version of Mallows's C_p) as a general extension of the Mallows's C_p (Mallows, [12]). Hin and Wang [8], Gosho *et al.* [6] proposed the correlation information criterion (CIC) to select the correlation structure. Unfortunately, above model selection criteria are not derived by taking account into the correlation between responses. For example, the risk function of the QIC is based on the independent quasi likelihood. In this respect, these criteria may not be reflective of the significant feature in longitudinal data.

The principal aim of the present paper is to obtain a new model selection criterion that reflects the correlation between responses. In this study, we have focused on deciding the best subset of variables. By using the risk function based on the prediction mean squared error (PMSE) normalized by the covariance matrix, we propose a new model selection criterion called the PMSEG (the prediction mean squared error in the GEE).

The remainder of the present paper is organized as follows: In Section 2, we consider a stochastic expansion of a GEE estimator. In Section 3, we propose the PMSEG. In Section 4, we verify that the proposed PMSEG has good properties by conducting numerical experiments. In Section 5, we conclude our discussion. Technical details are provided in the Appendix.

2. Stochastic expansion of the GEE estimator

Let y_{ij} be a scalar response variable, and let $\boldsymbol{x}_{*,ij}$ be a l-dimensional nonstochastic vector consists of all possible explanatory variables for the ith subject at the jth occasion, where $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Assume that response variables from different subjects are independent and response variables from the same subject are correlated. For $i = 1, \ldots, n$, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^\top$, $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \ldots, \boldsymbol{x}_{*,im_i})^\top$, and let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{im_i})^\top$ be a $m_i \times p$ submatrix of the matrix $\boldsymbol{X}_{*,i}$. Liang and Zeger [11] used the GLM to model the

marginal density of y_{ij},

$$f(y_{ij}; \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \phi) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij})\}/\phi + b(y_{ij}, \phi)], \tag{1}$$

where, $a(\cdot)$ and $b(\cdot)$ are known functions, θ_{ij} is an unknown location parameter and ϕ is a scale parameter. In the GLM framework, the location parameter depends on x_{ij} as $\theta_{ij} = u(\eta_{ij}) = \theta_{ij}(\boldsymbol{\beta})$, where $u(\cdot)$ is a known function, $\eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is a p-dimensional unknown parameter. In the present paper, we assume that ϕ is known, and we also assume Θ to be the *natural parameter space* (see, Xie and Yang, [18]) of the exponential family of distributions presented in (1), and the interior of Θ is denoted as Θ^0. Then Θ is convex, and in the Θ^0, all derivatives of $a(\cdot)$ and all moments of y_{ij} exist. Under such model conditions, the first two moments of y_{ij} are given by

$$\mu_{ij}(\boldsymbol{\beta}) = \mathrm{E}(y_{ij}) = \dot{a}(\theta_{ij}), \qquad \sigma_{ij}^2(\boldsymbol{\beta}) = \mathrm{Cov}(y_{ij}) = \ddot{a}(\theta_{ij})\phi \equiv v\{\mu_{ij}(\boldsymbol{\beta})\} \ \text{(say)}.$$

In this situation, the expectation of the response y_{ij} is modeled by a link function $g(t) = (\dot{a} \circ u)^{-1}(t)$ and the linear predictor η_{ij}, i.e., $g\{\mu_{ij}(\boldsymbol{\beta})\} = \eta_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. When $u(s) = s$, we say that $g(t) = (\dot{a})^{-1}(t)$ is the natural link function. For example, the logistic regression model is defined with the natural link function. We call the model with $\boldsymbol{x}_{*,ij}$ or $\boldsymbol{x}_{ij}$ as the full or candidate model, respectively. We assume that the true probability density function of y_{ij} can be written as (1), i.e., the true model is one of the candidate models.

Denote $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{im_i}(\boldsymbol{\beta}))^\top$, $\boldsymbol{A}_i(\boldsymbol{\beta}) = \mathrm{diag}\{\sigma_{i1}^2(\boldsymbol{\beta}), \ldots, \sigma_{im_i}^2(\boldsymbol{\beta})\}$ and $\boldsymbol{\Delta}_i(\boldsymbol{\beta}) = \mathrm{diag}(\partial\theta_{i1}/\partial\eta_{i1}, \ldots, \partial\theta_{im}/\partial\eta_{im})$, where $\mathrm{diag}(a_1, \ldots, a_s)$ denotes the $s \times s$ diagonal matrix whose the (i,i)th element is a_i. We write $\boldsymbol{D}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i(\boldsymbol{\beta})\boldsymbol{\Delta}_i(\boldsymbol{\beta})\boldsymbol{X}_i$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{i,0}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})$, where $\boldsymbol{R}_{i,0}$ is the true correlation matrix, and $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})$. Here, $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the working correlation matrix that one can choose freely, which may possibly have a nuisance parameter $\boldsymbol{\alpha}$. Depending on the situation, we can choose some useful working correlation matrices. For example,

$$\begin{aligned}
\text{Independence:} \quad & (\boldsymbol{R})_{jk} = 0, \ \text{if} \ \ j \neq k, \\
\text{Exchangeable:} \quad & (\boldsymbol{R})_{jk} = \alpha, \ \text{if} \ \ j \neq k, \\
\text{AR-1:} \quad & (\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha^{j-k}, \ \text{if} \ \ j > k, \\
\text{1-dependent:} \quad & (\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha, \ \text{if} \ \ j = k+1 \\
& \text{and} \ \ (\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = 0, \ \text{if} \ \ j > k+1, \\
\text{Unstructured:} \quad & (\boldsymbol{R})_{jk} = (\boldsymbol{R})_{kj} = \alpha_{jk}, \ \text{if} \ \ j > k.
\end{aligned}$$

If $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is equal to $\boldsymbol{R}_{i,0}$, then $\boldsymbol{V}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}) = \boldsymbol{\Sigma}_i(\boldsymbol{\beta}_0) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0)\boldsymbol{R}_{i,0}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_0) = \mathrm{Cov}[\boldsymbol{y}_i]$, where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$.

Liang and Zeger [11] proposed the GEE as follows:

$$\boldsymbol{s}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^{\top}(\boldsymbol{\beta})\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = \boldsymbol{0}_p, \tag{2}$$

where $\boldsymbol{0}_p$ is a p-dimensional vector of zeros. An estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is defined as a solution of the equation (2), and the estimator is called the GEE estimator. In the present paper, we assume the following important assumptions:

Assumption 1: $m_1 = \cdots = m_n = m$.

Assumption 2: $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{R}(\boldsymbol{\alpha})$ and $\boldsymbol{R}_{i,0} = \boldsymbol{R}_0$.

In other words, we assume that all subjects have the common occasion size m (the number of occasions) and the common true correlation matrix $\boldsymbol{R}_0$. As a result, we consider a common working correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$, not $\boldsymbol{R}_i(\boldsymbol{\alpha})$. These strong assumptions are necessary for deriving a new model selection criterion. Hereafter, we replace m_i, $\boldsymbol{R}_i(\boldsymbol{\alpha})$ and $\boldsymbol{R}_{i,0}$ with m, $\boldsymbol{R}(\boldsymbol{\alpha})$ and $\boldsymbol{R}_0$, respectively. Moreover, to simplify our discussion, we also assume that the nuisance parameter $\boldsymbol{\alpha}$ is known. Hence, we write $\boldsymbol{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{V}_i(\boldsymbol{\beta})$.

In order to propose a new model selection criterion in Section 3, a stochastic expansion of $\hat{\boldsymbol{\beta}}$ is needed. In this section, we obtain the stochastic expansion of $\hat{\boldsymbol{\beta}}$ up to the order n^{-1}. For simplicity, we omit $(\boldsymbol{\beta})$ from functions of $\boldsymbol{\beta}$. For example, $\mu_{ij}(\boldsymbol{\beta}) = \mu_{ij}$, $\boldsymbol{A}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_i$, $\boldsymbol{s}_n(\boldsymbol{\beta}) = \boldsymbol{s}_n$ and so on. Furthermore, in order to distinguish a function of $\boldsymbol{\beta}$ evaluated at the true parameter $\boldsymbol{\beta}_0$, we write such as $\mu_{ij}(\boldsymbol{\beta}_0) = \mu_{ij,0}$, $\boldsymbol{A}_i(\boldsymbol{\beta}_0) = \boldsymbol{A}_{i,0}$, $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}_0) = \boldsymbol{\Sigma}_{i,0}$, $\boldsymbol{s}_n(\boldsymbol{\beta}_0) = \boldsymbol{s}_{n,0}$ and so on. Similarly, in the functions of the GEE estimator $\hat{\boldsymbol{\beta}}$, we write such as $\mu_{ij}(\hat{\boldsymbol{\beta}}) = \hat{\mu}_{ij}$, $\boldsymbol{A}_i(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{A}}_i$, $\boldsymbol{\Sigma}_i(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\Sigma}}_i$, $\boldsymbol{s}_n(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{s}}_n$ and so on. Furthermore, in order to ensure asymptotic properties of the GEE estimator, we consider the following regularity assumptions (see, e.g., Xie and Yang, [18]):

- **C1.:** $\boldsymbol{\beta}_0$ is in an admissible set $\mathscr{B}$, where $\mathscr{B}$ is an open set in $\mathbb{R}^p$ for the parameter $\boldsymbol{\beta}$.
- **C2.:** $\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} \in g(\mathscr{M})$ for all $\boldsymbol{\beta} \in \mathscr{B}$, where $\mathscr{M}$ is the image of $\dot{\boldsymbol{a}}(\Theta^0)$.
- **C3.:** $u(\eta_{ij})$ is four times continuously differentiable and $\dot{u}(\eta_{ij}) > 0$ in $g(\mathscr{M})^0$.
- **C4.:** $\boldsymbol{H}_{n,0}$ and $\boldsymbol{M}_{n,0}$ are positive definite when n is large, where

$$\boldsymbol{H}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \equiv \boldsymbol{H}_n, \qquad \boldsymbol{H}_{n,0} = \boldsymbol{H}_n(\boldsymbol{\beta}_0),$$

$$\boldsymbol{M}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{D}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \equiv \boldsymbol{M}_n, \qquad \boldsymbol{M}_{n,0} = \boldsymbol{M}_n(\boldsymbol{\beta}_0).$$

Conditions C1 and C2 are necessary to have the GLM for all $\boldsymbol{\beta}$. Conditions C3 and C4 are necessary to calculate the bias. In addition, in order to ensure the strong consistency, asymptotic normality and uniqueness of the GEE estimator, we consider the following additional assumptions, which can be derived slightly modifying the results reported by Xie and Yang [18]:

C5.: $\liminf_{n\to\infty} \lambda_{\min}(\boldsymbol{H}_{n,0}/n) > 0$, where $\lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of symmetric matrix $\boldsymbol{A}$.

C6.: Sequence $\{\boldsymbol{x}_{ij}\}$ lies in $\mathscr{X}$ with $u(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}) \in \Theta^0$, $\boldsymbol{\beta} \in \mathscr{B}$, where $\mathscr{X}$ is a compact set for regressors $\boldsymbol{x}_{ij}$.

C7.: In a neighborhood of $\boldsymbol{\beta}_0$, say N, there exist a constant $c_0 > 0$ and some $\delta > 0$, independent of n, such that, when $n \to \infty$, for any p-dimensional vector $\boldsymbol{\lambda}$, $\|\boldsymbol{\lambda}\| = 1$,

$$\boldsymbol{\lambda}^\top \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}^\top}\boldsymbol{\lambda} \geq c_0 \lambda_{\max}^{(1/2)+\delta}(\boldsymbol{M}_{n,0}), \qquad \text{a.s. for } \boldsymbol{\beta} \in N,$$

where $\lambda_{\max}(\boldsymbol{A})$ is the largest eigenvalue of symmetric matrix $\boldsymbol{A}$.

C8.: The equation (2) has a unique solution when n is large.

Note that conditions C1–C7 are sufficient conditions of Theorem 7 and Corollary 1 in Xie and Yang [18]. Hence, according to Theorem 7 and Corollary 1, $\hat{\boldsymbol{\beta}}$ has the strong consistency and asymptotic normality. Moreover, from the condition C8, $\hat{\boldsymbol{\beta}}$ is a unique solution of the GEE (2). Furthermore, from C5, $\boldsymbol{H}_{n,0} = O(n)$.

Based on the above conditions, to perform the stochastic expansion of $\hat{\boldsymbol{\beta}}$, we focus on the fact that $\hat{\boldsymbol{s}}_n = \boldsymbol{0}_p$. By applying Taylor's expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ to this equation, we obtain the stochastic expansion of $\hat{\boldsymbol{\beta}}$.

The GEE estimator $\hat{\boldsymbol{\beta}}$ can be expanded as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{b}_{1,0} + \boldsymbol{b}_{2,0} + O_p(n^{-3/2}), \tag{3}$$

where

$$\boldsymbol{b}_{1,0} = \boldsymbol{H}_{n,0}^{-1}\boldsymbol{s}_{n,0}, \qquad \boldsymbol{b}_{2,0} = \boldsymbol{H}_{n,0}^{-1}(\boldsymbol{b}_{1,0}^\top \otimes \boldsymbol{I}_p)\boldsymbol{G}_{3,0}\boldsymbol{b}_{1,0}/2 - \boldsymbol{G}_{1,0}\boldsymbol{b}_{1,0} - \boldsymbol{G}_{2,0}\boldsymbol{b}_{1,0},$$

and $\boldsymbol{G}_{1,0}$, $\boldsymbol{G}_{2,0}$ and $\boldsymbol{G}_{3,0}$ are given by

$$\boldsymbol{G}_{1,0} = -\boldsymbol{H}_{n,0}^{-1}\sum_{i=1}^{n}\boldsymbol{D}_{i,0}^\top\left(\frac{\partial}{\partial\boldsymbol{\beta}^\top}\otimes\boldsymbol{V}_i^{-1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\{\boldsymbol{I}_p\otimes(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\},$$

$$\boldsymbol{G}_{2,0} = -\boldsymbol{H}_{n,0}^{-1}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\beta}^\top}\otimes\boldsymbol{D}_i^\top\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)[\boldsymbol{I}_p\otimes\{\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\}],$$

$$\boldsymbol{G}_{3,0} = \mathrm{E}\{\boldsymbol{L}_1(\boldsymbol{\beta}_0)\}, \qquad \boldsymbol{L}_1(\boldsymbol{\beta}_0) = \left(\frac{\partial}{\partial\boldsymbol{\beta}}\otimes\frac{\partial\boldsymbol{s}_n}{\partial\boldsymbol{\beta}^\top}\right)\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

The derivation of (3) is given in Appendix. Note that for a matrix $\boldsymbol{W} = (w_{ij})$, the derivative of $\boldsymbol{W}$ by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and β_k are respectively defined as follows:

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top} \otimes \boldsymbol{W} = \left(\frac{\partial \boldsymbol{W}}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{W}}{\partial \beta_p} \right), \qquad \frac{\partial}{\partial \boldsymbol{\beta}} \otimes \boldsymbol{W} = \left(\frac{\partial \boldsymbol{W}^\top}{\partial \beta_1}, \ldots, \frac{\partial \boldsymbol{W}^\top}{\partial \beta_p} \right)^\top,$$

$$\frac{\partial \boldsymbol{W}}{\partial \beta_k} = \left(\frac{\partial w_{ij}}{\partial \beta_k} \right).$$

Also note that $\boldsymbol{b}_{1,0}$ and $\boldsymbol{b}_{2,0}$ are $O_p(n^{-1/2})$ and $O_p(n^{-1})$, respectively.

3. Main result

In this section, we propose a new model selection criterion. The goodness of fit of the model is measured by the risk function based on the PMSE normalized by the covariance matrix as follows:

$$\mathrm{Risk_P} = \mathrm{PMSE} - mn = \mathrm{E}_y \left[\mathrm{E}_z \left\{ \sum_{i=1}^{n} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)^\top \boldsymbol{\Sigma}_{i,0}^{-1} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i) \right\} \right] - mn,$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^\top$ is an m-dimensional random vector that is independent of $\boldsymbol{y}_i$ and has the same distribution as $\boldsymbol{y}_i$. It is easy to show that if $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\beta}_0$, then $\mathrm{Risk_P}$ has the minimum value of zero, i.e., the PMSE has the minimum value of mn. For this reason, we would like to select the model, which has the minimum PMSE. However, since the PMSE is typically unknown, we must estimate it.

We define $\tilde{\boldsymbol{R}}(\boldsymbol{\beta})$, $\mathscr{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\tilde{\mathscr{L}}(\boldsymbol{\beta})$ as follows:

$$\tilde{\boldsymbol{R}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_i^{-1/2} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{A}_i^{-1/2},$$

$$\mathscr{L}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \sum_{i=1}^{n} \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)\}^\top \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2) \tilde{\boldsymbol{R}}^{-1}(\boldsymbol{\beta}_2) \boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2) \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)\},$$

$$\tilde{\mathscr{L}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_{i,0}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i).$$

Based on the above, we estimate PMSE by $\mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$, where $\hat{\boldsymbol{\beta}}_f$ is a GEE estimator that is obtained under the "full" model. Specifically, $\hat{\boldsymbol{\beta}}_f$ is defined as the solution to the following equation:

$$\boldsymbol{s}_{f,n}(\boldsymbol{\beta}_*) = \sum_{i=1}^{n} \boldsymbol{D}_i^\top(\boldsymbol{\beta}_*) \boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_*, \boldsymbol{\alpha}_\dagger) \{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_*)\} = \boldsymbol{0}_l,$$

where $\boldsymbol{D}_i(\boldsymbol{\beta}_*) = \boldsymbol{A}_i(\boldsymbol{\beta}_*)\boldsymbol{\Delta}_i(\boldsymbol{\beta}_*)\boldsymbol{X}_{*,i}$, $\boldsymbol{V}_i(\boldsymbol{\beta}_*, \boldsymbol{\alpha}_\dagger) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_*)\overline{\boldsymbol{R}}_i(\boldsymbol{\alpha}_\dagger)\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta}_*)$ and $\overline{\boldsymbol{R}}_i(\boldsymbol{\alpha}_\dagger)$ is a positive definite working correlation matrix that one can choose freely. We assume that the nuisance parameter $\boldsymbol{\alpha}_\dagger$ is known. Note that $\boldsymbol{\beta}_*$ is an l-dimensional unknown parameter under the full model. Also note that $\overline{\boldsymbol{R}}_i(\boldsymbol{\alpha}_\dagger)$ is the same for all candidate models. The reason for using $\hat{\boldsymbol{\beta}}_f$ is discussed later. For simplicity, we write $\mathscr{L}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_2) = \mathscr{L}(\boldsymbol{\beta}_2)$ and $\tilde{\mathscr{L}}(\boldsymbol{\beta}_0) = \tilde{\mathscr{L}}$.

Since $\mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$ is not an asymptotically unbiased estimator of PMSE, we evaluate the asymptotic bias in order to propose the new model selection criterion. The bias when we estimate the PMSE by $\mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$ is given as

$$\begin{aligned}\text{Bias} = \text{PMSE} - \text{E}_y\{\mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)\} &= [\text{Risk}_\text{P} - \text{E}_y\{\tilde{\mathscr{L}}(\hat{\boldsymbol{\beta}})\}] + [\text{E}_y\{\tilde{\mathscr{L}}(\hat{\boldsymbol{\beta}}) - \tilde{\mathscr{L}}\}] \\ &\quad + [\text{E}_y\{\tilde{\mathscr{L}} - \mathscr{L}(\hat{\boldsymbol{\beta}}_f)\}] + [\text{E}_y\{\mathscr{L}(\hat{\boldsymbol{\beta}}_f) - \mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)\}] \\ &= \text{Bias1} + \text{Bias2} + \text{Bias3} + \text{Bias4}. \end{aligned} \tag{4}$$

Here, we can see that Bias3 in (4) satisfies

$$\begin{aligned}\text{Bias3} &= \text{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top\{\boldsymbol{\Sigma}_{i,0}^{-1} - \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\right] \\ &= mn - \text{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})\right\}.\end{aligned}$$

Therefore, we can ignore the calculation of Bias3 because it is not dependent on candidate models.

On the other hand, Bias1 + Bias2 + Bias4 in (4) is given by

$$\text{Bias1} + \text{Bias2} + \text{Bias4} = 2p + O(n^{-1}). \tag{5}$$

The derivation of (5) is given in Appendix.

Consequently, by substituting (5) into (4), we obtain the asymptotic expansion of Bias up to order 1 as

$$\text{Bias} = 2p + \text{Bias3} + O(n^{-1}). \tag{6}$$

Recall that Bias3 is not dependent on candidate models. Hence, the PMSEG can then be defined by adding an estimated (Bias − Bias3) to $\mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f)$, i.e.,

$$\text{PMSEG} = \mathscr{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_f) + 2p. \tag{7}$$

(7) is our proposed model selection criterion called the PMSEG (the prediction mean squared error in the GEE). Recall that $\hat{\boldsymbol{\beta}}_f$ is estimated under the full model and it is not dependent on candidate models. Since the covariance

matrix in the PMSEG is estimated by $\hat{\boldsymbol{\beta}}_f$, the PMSEG can be simply defined. If the covariance matrix is estimated by $\hat{\boldsymbol{\beta}}$, Bias3 in (7) is different for each candidate model. Unfortunately, it is difficult and too expensive to calculate Bias3. This is one of the advantages of using $\hat{\boldsymbol{\beta}}_f$ for estimating the covariance matrix. For actual use, we recommend to use the independence working correlation matrix in order to obtain $\hat{\boldsymbol{\beta}}_f$ since we can get $\hat{\boldsymbol{\beta}}_f$ easily by omitting the calculations of the working correlation matrix. Fitzmaurice [5] mentioned that the GEE estimator under the working independence assumption is often inefficient. Nevertheless, from some simulation results, we confirmed that the estimation of the covariance matrix using the inefficient estimator does not dramatically influence the performance of the PMSEG (see, Subsection 4.2).

4. Numerical studies

In this section, we confirm a usefulness of the PMSEG through comparisons with the QIC and modified QIC (called mQIC in this paper), which are representative model selection criteria. The QIC and mQIC are estimators of a risk function based on the quasi likelihood under the independence assumption. The risk function that is estimated by the QIC (or mQIC), which is called $\mathrm{Risk_Q}$ in this paper, and the quasi likelihood $Q(\cdot)$ are defined as follows:

$$\mathrm{Risk_Q} = \mathrm{E}_y\left[\mathrm{E}_z\left\{-2\sum_{i=1}^{n}\sum_{j=1}^{m} Q(\hat{\boldsymbol{\beta}}; z_{ij})\right\}\right], \qquad Q(\hat{\boldsymbol{\beta}}; z_{ij}) = \int_{z_{ij}}^{g^{-1}(\boldsymbol{x}_{ij}^{\top}\hat{\boldsymbol{\beta}})} \frac{z_{ij}-t}{v(t)}\,dt,$$

where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^{\top}$ is an m-dimensional random vector that is independent of $\boldsymbol{y}_i$ and has the same distribution as $\boldsymbol{y}_i$. Note that since the $\mathrm{Risk_Q}$ is considered under the independence assumption, all of the $\mathrm{Risk_Q}$, QIC and mQIC are not reflective of the correlation between responses. Also note that the PMSEG and QIC (or mQIC) are estimators of the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$, respectively. In general, the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$ are different, i.e., the PMSEG and QIC (or mQIC) are criteria from different viewpoints.

4.1. Selection probability and prediction error. At the beginning, we examine the numerical studies for the frequencies of the selected candidate models and the prediction error of the best models selected by the PMSEG, QIC and mQIC. We prepare the fifteen candidate models with $n = 100$ and $m = 3$. First, we construct a 3×5 explanatory variable matrix $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \boldsymbol{x}_{*,i2}, \boldsymbol{x}_{*,i3})^{\top}$, $i = 1, \ldots, 100$. The first column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_3$, where $\mathbf{1}_3$ is a 3-dimensional vector of ones, and the second column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_3 \times \varsigma_i$, where $\varsigma_1, \ldots, \varsigma_{100}$ are independent and identically distributed (i.i.d.) random variables from binomial distribution $\mathrm{B}(1, 0.5)$. The third column of $\boldsymbol{X}_{*,i}$ is $(0, 1, 2)^{\top}$,

and the remaining six elements of $\boldsymbol{X}_{*,i}$ are defined by realizations of independent variables with uniform distribution on the interval $[-1, 1]$.

In this simulation, we prepare two situations, as follows:

$$\text{Case 1:}\quad \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.85^{|j-j^*|},\ \boldsymbol{\beta}_0 = (0.25, -0.25, -0.25, 0, 0)^\top,$$

$$\text{Case 2:}\quad \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.35 + \delta_{jj^*}0.65,\ \boldsymbol{\beta}_0 = (0.25, -0.25, -0.25, 0, 0)^\top,$$

where δ_{jj^*} is the Kronecker delta, i.e., $\delta_{jj^*} = 1$ if $j = j^*$ and $\delta_{jj^*} = 0$ if $j \neq j^*$. The explanatory variables matrix for the ith subject in the $(5a + b)$th model consists of the first b column of $\boldsymbol{X}_{*,i}$, $a = 0, 1, 2$, $b = 1, \ldots, 5$. For the working correlation matrix, we prepare three different matrices, exchangeable working correlation matrix $(a = 0)$, AR-1 working correlation matrix $(a = 1)$ and independence working correlation matrix $(a = 2)$. Thus, in Case 1, the true model is the eighth model, and in Case 2, the true model is the third model. We simulate 10,000 realizations of $\boldsymbol{y} = (y_{11}, \ldots, y_{13}, \ldots, y_{100,1}, \ldots, y_{100,3})^\top$ in the logistic regression model, i.e., $y_{ij} \sim \mathrm{B}(1, p_{ij})$, where $p_{ij} = \mathrm{logit}^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_0)$, $i = 1, \ldots, 100$, $j = 1, 2, 3$. Note that we use the independence working correlation matrix for obtaining $\hat{\boldsymbol{\beta}}_f$ in this simulation. The average values of estimates of the GEE estimator for each model and its variances are given in Tables 1 and 2. Furthermore, Tables 3 and 4 list the following characteristics.

Table 1. Average values of estimates of the GEE estimator and its variances for each model in Case 1

	1		2		**3**		4		5	
	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$
Exchangeable	−0.119	0.035	−0.002	0.070	0.254	0.078	0.255	0.078	0.255	0.078
			−0.248	0.149	−0.251	0.154	−0.251	0.154	−0.251	0.155
					−0.256	0.006	−0.256	0.006	−0.256	0.006
							0.000	0.013	0.000	0.013
									0.000	0.011
AR-1	−0.118	0.034	−0.002	0.069	0.254	0.077	0.254	0.077	0.255	0.077
			−0.247	0.147	−0.251	0.152	−0.251	0.152	−0.251	0.153
					−0.256	0.006	−0.256	0.006	−0.256	0.006
							0.000	0.011	0.000	0.009
									0.000	0.011
Independence	−0.119	0.035	−0.002	0.070	0.254	0.077	0.255	0.078	0.256	0.078
			−0.248	0.149	−0.251	0.153	−0.251	0.154	−0.252	0.155
					−0.256	0.006	−0.257	0.006	−0.258	0.006
							−0.002	0.042	−0.003	0.044
									0.003	0.039

Table 2. Average values of estimates of the GEE estimator and its variances for each model in Case 2

	1		2		**3**		4		5	
	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$	$\hat{\beta}$	$\mathrm{Var}(\hat{\beta})$
Exchangeable	−0.119	0.022	0.000	0.044	0.255	0.059	0.256	0.059	0.256	0.060
			−0.249	0.096	−0.252	0.099	−0.253	0.099	−0.254	0.100
					−0.255	0.014	−0.256	0.014	−0.257	0.014
							0.001	0.036	0.001	0.037
									0.002	0.035
AR-1	−0.119	0.022	−0.001	0.045	0.254	0.060	0.255	0.060	0.256	0.061
			−0.249	0.097	−0.252	0.100	−0.253	0.101	−0.254	0.102
					−0.255	0.014	−0.256	0.014	−0.257	0.014
							0.001	0.038	0.000	0.039
									0.002	0.037
Independence	−0.119	0.022	0.000	0.044	0.255	0.059	0.256	0.059	0.257	0.060
			−0.249	0.096	−0.252	0.099	−0.253	0.099	−0.254	0.100
					−0.255	0.014	−0.256	0.014	−0.257	0.014
							0.000	0.044	0.000	0.045
									0.003	0.042

(1): Prediction error of the best model in the $\mathrm{Risk_P}/\mathrm{Risk_Q}$ ($\mathrm{PEB_P}/\mathrm{PEB_Q}$): the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$ of the model selected by the criterion as the best model, which are respectively estimated as

$$\mathrm{PEB_P} = \frac{1}{10000}\sum_{\upsilon=1}^{10000} \mathrm{E}_z\left[\sum_{i=1}^{n}\{\boldsymbol{z}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon})\}^\top \boldsymbol{\Sigma}_{i,0}^{-1}\{\boldsymbol{z}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon})\}\right] - mn,$$

$$\mathrm{PEB_Q} = \frac{1}{10000}\sum_{\upsilon=1}^{10000} \mathrm{E}_z\left\{-2\sum_{i=1}^{n}\sum_{j=1}^{m} Q(\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}; z_{ij})\right\}.$$

(2): Selection probability: the frequency of the best model chosen by minimizing each criterion. In particular, the $\mathrm{SP_C}/\mathrm{SP_M}$ is the frequency that the working correlation matrix/mean structure of the selected model is correctly specified.

Here $\boldsymbol{z}_i$ is a future observation, and $\hat{\boldsymbol{\beta}}_{\mathrm{B}_\upsilon}$ is the value of $\hat{\boldsymbol{\beta}}$ of the selected model at the υth iteration. In particular, both the $\mathrm{PEB_P}$ and $\mathrm{PEB_Q}$ are important properties because these are equivalent to the $\mathrm{Risk_P}$ and $\mathrm{Risk_Q}$ of the best model selected by the criterion, respectively. Moreover, the values in parenthesis indicate standard errors in Tables 3 and 4. We would like to note that the PMSEG selects the model which minimizes the $\mathrm{Risk_P}$, and the QIC

Table 3. Selection probability and prediction error in Case 1

Criterion	b	1	2	**3**	4	5	SP_C	SP_M	PEB_P	PEB_Q
PMSEG	Exchangeable	0.95	0.08	11.02	2.62	1.10	70.49 (0.46)	**95.61** (0.20)	**4.37** (0.04)	**416.66** (0.06)
	AR-1	2.61	0.48	**50.71**	9.77	6.92				
	Independence	0.22	0.05	12.72	0.67	0.08				
QIC	Exchangeable	4.36	0.01	2.60	3.46	3.60	**72.79** (0.45)	55.50 (0.50)	10.54 (0.07)	418.09 (0.07)
	AR-1	38.36	0.26	**22.20**	6.11	5.86				
	Independence	1.51	0.00	3.09	5.17	3.41				
mQIC	Exchangeable	4.58	0.02	2.30	3.52	4.06	72.02 (0.45)	55.25 (0.50)	10.70 (0.07)	418.05 (0.06)
	AR-1	38.62	0.23	**21.34**	5.96	5.87				
	Independence	1.30	0.00	3.01	5.08	4.11				

Table 4. Selection probability and prediction error in Case 2

Criterion	b	1	2	**3**	4	5	SP_C	SP_M	PEB_P	PEB_Q
PMSEG	**Exchangeable**	13.97	0.89	**26.28**	8.06	6.76	55.96 (0.50)	**72.79** (0.45)	**4.88** (0.03)	**416.24** (0.05)
	AR-1	7.67	1.77	9.10	1.97	1.20				
	Independence	2.06	0.85	17.00	1.73	0.69				
QIC	**Exchangeable**	29.70	1.14	**20.06**	7.46	5.64	**64.00** (0.48)	57.93 (0.49)	5.53 (0.04)	416.61 (0.05)
	AR-1	2.55	0.19	1.70	1.17	1.14				
	Independence	7.53	0.96	16.64	2.56	1.56				
mQIC	**Exchangeable**	21.63	3.27	**19.76**	6.27	4.77	55.70 (0.50)	58.21 (0.49)	5.54 (0.04)	416.53 (0.05)
	AR-1	9.62	1.25	6.72	2.65	2.16				
	Independence	5.13	0.89	12.27	2.15	1.46				

(or mQIC) selects the model which minimizes the $Risk_Q$. Thus, the model selected by the PMSEG does not necessarily minimize the $Risk_Q$ and the model selected by the QIC (or mQIC) does not necessarily minimize the $Risk_P$. In other words, in order to evaluate the goodness of the criterion, the PEB_P and PEB_Q are favorable indicators for the PMSEG and both QIC and mQIC, respectively.

From Tables 3 and 4, we can see that the value of the PEB_P from the model selected by the PMSEG is smaller than that from the model selected by the QIC (or mQIC). This result is justified since the PEB_P is the favorable indicator for the PMSEG. However, surprisingly, although the PEB_Q is the favorable indicator for the QIC (or mQIC), the value of the PEB_Q from the model selected by the PMSEG is also smaller. This result means that the PMSEG is better than the QIC and mQIC whether evaluating the goodness of the criterion by the PEB_P or PEB_Q. Moreover, both the frequency of selecting

the true model and SP_{M} of the PMSEG are larger than those of the QIC and mQIC in two cases. On the other hand, the SP_{C} of the QIC is larger in two cases. Furthermore, by comparing Tables 3 and 4, we can see that the difference between the PMSEG and both QIC and mQIC is more salient when the correlation is large. We simulated several other models and obtained similar results.

4.2. Estimation of the covariance matrix using an inefficient GEE estimator. Recall that we must estimate the covariance matrix to calculate the PMSEG. As we mentioned before, in order to calculate the covariance matrix in the PMSEG, we suggest using the $\hat{\boldsymbol{\beta}}_f$ obtained by solving the GEE with the independence working correlation matrix and all explanatory variables. In this subsection, we use working independence, exchangeable and AR-1 matrices to obtain the $\hat{\boldsymbol{\beta}}_f$, and using these GEE estimators we calculate the covariance matrix in the PMSEG. In this setting, we confirm the influence on the prediction mean by using different working correlation matrices through numerical experiments.

We prepare the fifteen candidate models with $n = 100$ and $m = 5$. First, we construct a 5×5 matrix $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \boldsymbol{x}_{*,i2}, \boldsymbol{x}_{*,i3}, \boldsymbol{x}_{*,i4}, \boldsymbol{x}_{*,i5})^\top$, $i = 1, \ldots, 100$. The first column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_5$, and the second column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_5 \times \varsigma_i$, where $\varsigma_1, \ldots, \varsigma_{100}$ are i.i.d. random variables from $\mathrm{B}(1, 0.5)$. The third column of $\boldsymbol{X}_{*,i}$ is $(0, 1, 2, 3, 4)^\top$. The fourth and the fifth columns of $\boldsymbol{X}_{*,i}$ are defined by realizations of independent variables with uniform distribution on the interval $[-0.5, 0.5]$ and $[-1.5, -0.5]$, respectively.

In this simulation, we consider the following situation:

$$\mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.3^{|j-j^*|}, \qquad \boldsymbol{\beta}_0 = (-0.45, -0.35, 0.3, 0, 0)^\top.$$

The explanatory variables matrix for the ith subject in the $(5a + b)$th model consists of the first b column of $\boldsymbol{X}_{*,i}$, $a = 0, 1, 2$, $b = 1, \ldots, 5$. For the working correlation matrix, we prepare three different matrices, AR-1 working correlation matrix $(a = 0)$, exchangeable working correlation matrix $(a = 1)$, and independence working correlation matrix $(a = 2)$. We simulate 10,000 realizations of $\boldsymbol{y} = (y_{11}, \ldots, y_{15}, \ldots, y_{100,1}, \ldots, y_{100,5})^\top$ in the logistic regression model, i.e., $y_{ij} \sim \mathrm{B}(1, p_{ij})$, where $p_{ij} = \mathrm{logit}^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_0)$, $i = 1, \ldots, 100$, $j = 1, \ldots, 5$. Under this setting, average values of estimates, variances and variance ratios (VR) of the GEE estimator for using each working correlation matrix are given in Table 5. Here, the VR is defined by the ratio of the variance of the GEE estimator to that for using the AR-1 working correlation matrix. For example, the value of the variance of the GEE estimator using the AR-1 and independence structures in β_4 are 0.093 and 0.108, respectively. Hence, the

Table 5. Average of estimate (Est), variance (Var), variance ratio (VR) of the GEE estimator of the full model, and $\mathrm{PEB_P}$ for using each working correlation matrix

	AR-1			Exchangeable			Independence		
	Est	Var	VR	Est	Var	VR	Est	Var	VR
β_1	−0.456	0.137	1.000	−0.457	0.145	1.058	−0.458	0.152	1.109
β_2	−0.354	0.056	1.000	−0.353	0.057	1.018	−0.353	0.057	1.018
β_3	0.304	0.005	1.000	0.304	0.005	1.000	0.304	0.005	1.000
β_4	−0.007	0.093	1.000	−0.006	0.101	1.086	−0.008	0.108	1.161
β_5	0.001	0.086	1.000	−0.001	0.093	1.081	−0.001	0.100	1.163
$\mathrm{PEB_P}$	**4.147** (0.034)			4.148 (0.034)			4.150 (0.034)		

VR is calculated as $\mathrm{VR} = 0.108/0.093 \approx 1.161$. Moreover, the $\mathrm{PEB_P}$ for using each working correlation matrix is also given in Table 5 and the values in parenthesis indicate standard errors.

From Table 5, we can see that the GEE estimator using the AR-1 (Independence) structure is the most efficient (inefficient). However, the values of the $\mathrm{PEB_P}$ using working AR-1 and independence matrices are 4.147 and 4.150, respectively. Therefore, we can see that the estimation of the covariance matrix using the inefficient estimator does not dramatically influence the $\mathrm{PEB_P}$. We simulated several other settings and obtained similar results.

4.3. Selection probability when the number of occasions is not small. In this subsection, we calculate the selection probabilities using the PMSEG, QIC and mQIC when the number of occasions m is not small. We prepare the ten candidate models with n individuals and m occasions. First, we construct a $m \times 5$ explanatory variable matrix $\boldsymbol{X}_{*,i} = (\boldsymbol{x}_{*,i1}, \ldots, \boldsymbol{x}_{*,im})^\top$, $i = 1, \ldots, n$. The first column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_m$, and the second column of $\boldsymbol{X}_{*,i}$ is $\mathbf{1}_m \times \varsigma_i$, where $\varsigma_1, \ldots, \varsigma_n$ are i.i.d. random variables from $\mathrm{B}(1, 0.5)$. The third column of $\boldsymbol{X}_{*,i}$ is $(0, \ldots, m-1)^\top$, and the remaining $2 \times m$ elements of $\boldsymbol{X}_{*,i}$ are defined by realizations of independent variables with uniform distribution on the interval $[-1, 1]$.

In this simulation, we prepared two situations, as follows:

$$\text{Case 1:}\quad \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.2^{|j-j^*|},\ \boldsymbol{\beta}_0 = (-1.2, -0.2, 0.2, 0, 0)^\top,\ m = 2 + \frac{n}{25},$$

$$\text{Case 2:}\quad \mathrm{Corr}[y_{ij}, y_{ij^*}] = 0.2^{|j-j^*|},\ \boldsymbol{\beta}_0 = (-1.2, -0.2, 0.2, 0, 0)^\top,$$
$$m = \max\{6, 0.1n\}.$$

The explanatory variables matrix for the ith subject in the $(5a+b)$th model consists of the first b column of $\boldsymbol{X}_{*,i}$, $a=0,1$, $b=1,\ldots,5$. For the working correlation matrix, we prepare two different matrices, independence working correlation matrix ($a=0$) and AR-1 working correlation matrix ($a=1$). We simulated 1,000 realizations of $\boldsymbol{y}=(y_{11},\ldots,y_{1m},\ldots,y_{n,1},\ldots,y_{n,m})^\top$ in the logistic regression model, i.e., $y_{ij}\sim \mathrm{B}(1,p_{ij})$, where $p_{ij}=\mathrm{logit}^{-1}(\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}_0)$, $i=1,\ldots,n$, $j=1,\ldots,m$. Note that we used the independence working correlation matrix for obtaining $\hat{\boldsymbol{\beta}}_f$ in this simulation.

From Tables 6 and 7, we can see that the selection probability of an underspecified model, which does not include the true model, converges to zero for all criteria. In addition, its convergence speed becomes faster when the number of occasions becomes large. On the other hand, we can also see that the selection probability of the true model does not tend to one for all criteria. Incidentally, in multivariate normal linear regression models, it is known that the AIC has consistency under the high-dimensional framework (see, Yanagihara et al. [19]). However, although the PMSEG, QIC and mQIC

Table 6. Selection probability (%) by using each criterion for ten candidate models in Case 1

			1		2		**3**		4		5	
N	m	Criterion	Ind	AR-1	Ind	AR-1	Ind	**AR-1**	Ind	AR-1	Ind	AR-1
50	4	QIC	20.0	**29.7**	4.5	6.4	8.8	17.7	2.4	5.1	1.8	3.6
		mQIC	16.6	**33.3**	3.5	7.4	7.8	18.7	1.9	5.4	1.4	4.0
		PMSEG	**21.1**	17.9	4.2	6.9	12.9	16.8	4.5	7.2	3.2	5.3
100	6	QIC	0.6	1.6	0.1	0.3	22.6	**54.7**	2.7	9.8	1.4	6.2
		mQIC	0.5	1.7	0.1	0.3	13.9	**63.4**	1.7	10.9	1.0	6.5
		PMSEG	0.8	0.9	0.1	0.1	28.9	**47.5**	4.5	8.2	2.2	6.8
150	8	QIC	0.0	0.0	0.0	0.0	18.1	**54.9**	3.6	12.6	2.9	7.9
		mQIC	0.0	0.0	0.0	0.0	12.9	**60.1**	3.1	13.2	2.3	8.4
		PMSEG	0.0	0.0	0.0	0.0	26.1	**44.7**	5.0	11.9	3.0	9.3
200	10	QIC	0.0	0.0	0.0	0.0	19.5	**57.3**	3.0	10.7	1.0	8.5
		mQIC	0.0	0.0	0.0	0.0	12.3	**64.7**	2.6	11.1	1.1	8.2
		PMSEG	0.0	0.0	0.0	0.0	30.8	**45.2**	4.6	8.9	2.7	7.8
250	12	QIC	0.0	0.0	0.0	0.0	20.5	**57.0**	2.9	11.1	1.3	7.2
		mQIC	0.0	0.0	0.0	0.0	14.2	**63.2**	2.1	11.9	1.2	7.4
		PMSEG	0.0	0.0	0.0	0.0	30.4	**45.7**	3.8	9.5	2.9	7.7
300	14	QIC	0.0	0.0	0.0	0.0	24.0	**53.5**	2.8	11.7	1.3	6.7
		mQIC	0.0	0.0	0.0	0.0	14.3	**63.2**	2.3	12.3	1.3	6.6
		PMSEG	0.0	0.0	0.0	0.0	30.9	**45.4**	4.6	9.8	2.6	6.7

Table 7. Selection probability (%) by using each criterion for ten candidate models in Case 2

			1		2		**3**		4		5	
N	m	Criterion	Ind	AR-1	Ind	AR-1	Ind	**AR-1**	Ind	AR-1	Ind	AR-1
50	6	QIC	7.8	10.5	0.8	1.5	21.5	**37.6**	4.8	9.0	1.7	4.8
		mQIC	6.3	12.2	0.9	1.4	16.2	**42.8**	4.0	9.7	0.8	5.7
		PMSEG	5.5	5.7	1.8	1.2	24.7	**33.9**	6.5	9.4	3.2	8.1
100	10	QIC	0.0	0.0	0.0	0.0	23.6	**51.1**	5.0	10.1	1.7	8.5
		mQIC	0.0	0.0	0.0	0.0	15.2	**59.7**	3.5	11.4	1.4	8.8
		PMSEG	0.0	0.0	0.0	0.0	30.9	**42.0**	5.7	9.2	4.5	7.7
150	15	QIC	0.0	0.0	0.0	0.0	28.1	**48.9**	2.8	10.4	2.2	7.6
		mQIC	0.0	0.0	0.0	0.0	19.6	**57.5**	2.2	11.0	1.7	8.0
		PMSEG	0.0	0.0	0.0	0.0	33.5	**39.1**	6.1	8.4	4.7	8.2
200	20	QIC	0.0	0.0	0.0	0.0	29.7	**47.2**	3.8	11.3	1.6	6.4
		mQIC	0.0	0.0	0.0	0.0	21.7	**55.6**	2.9	11.7	1.3	6.8
		PMSEG	0.0	0.0	0.0	0.0	35.2	**38.6**	5.9	9.2	3.8	7.3
250	25	QIC	0.0	0.0	0.0	0.0	27.7	**50.3**	3.2	9.6	1.9	7.3
		mQIC	0.0	0.0	0.0	0.0	20.6	**57.1**	2.6	10.3	1.9	7.5
		PMSEG	0.0	0.0	0.0	0.0	35.1	**38.4**	6.9	7.1	4.8	7.7
300	30	QIC	0.0	0.0	0.0	0.0	26.8	**50.2**	3.3	10.1	1.8	7.8
		mQIC	0.0	0.0	0.0	0.0	21.9	**54.9**	2.9	10.6	1.7	8.0
		PMSEG	0.0	0.0	0.0	0.0	35.5	**39.4**	4.1	8.8	5.5	6.7

are similar to the AIC, from Tables 6 and 7 we can see that these criteria do not have consistency. Actually, the number of regression parameters for a candidate model considered by Yanagihara et al. [19] becomes large when the number of dimensions becomes large. As a result, the penalty term of the AIC becomes large, and the difference between penalty terms of two candidate models also becomes large. This is the key point to prove consistency of the AIC. By contrast, in our considered model, the penalty term of the PMSEG is $2 \times$ the number of explanatory variables even if the number of occasions m becomes large. Similarly, the order of penalty terms of the QIC and mQIC is $O(p)$. Hence, for all criteria, the difference between penalty terms of two candidate models is a constant even if the number of occasions becomes large. This is probably one of reasons why the consistency of the PMSEG, QIC and mQIC is not confirmed in our simulation study.

4.4. Real data analysis. Next, for the purpose of analyzing the GEE method, we consider the Mother's Stress and Child Morbidity (MSCM) data reported in Alexander and Markowitz [3], who studied the relationship between maternal

employment and pediatric health care utilization. The MSCM data contain the information of mothers and children in the study for 28 days, and there are 167 mothers and preschool children enrolled. In this analysis, we focus on the child illness for the first 9 days. The response variable is the child illness on the study day (1 = yes, 0 = no), and there were six predictor variables: Race (child race, 1 = non-white, 0 = white), Household (size of household, 1 = more than 3 people, 0 = other), Stress (today's mother's stress, 1 = yes, 0 = no), and additionally, St1, St2 and St3 are the mother's stress of one, two and three days before, respectively. This data have a few missing value, and we assume that the missingness mechanism is missing completely at random. Thus, we use complete 146 mothers and children data. We also assume that the response variable y_{ij} is distributed according to $\mathrm{B}(1, p_{ij})$, $i = 1, \ldots, 146$, $j = 1, \ldots, 9$. For the link function, we prepare the logistic link function. For the working correlation matrix, we prepare three matrices: the working AR-1, exchangeable and independence matrices. In this analysis, we select the working correlation matrix and variables.

Table 8 shows the selection probability of the model selected by minimizing the criterion and the estimated prediction error of the best model selected by the criterion, where the values in parenthesis indicate standard errors. We divide the MSCM data into calibration data and validation data. The numbers of subjects in the calibration data and validation data were 136 and 10, respectively. The best subset of variables and working correlation matrices are selected by each criterion derived from the calibration data. The selection probabilities are obtained from only the calibration data. The estimated prediction errors are obtained as follows. Let $\boldsymbol{d}_t = (d_{1t}, \ldots, d_{146t})^\top$ be a 146-dimensional binary vector that contains 136 zeros and 10 ones at the tth iteration, $t = 1, \ldots, 100$, i.e., $d_{it} = 0$ or 1 and $\sum_{i=1}^{146} d_{it} = 10$. In addition, we denote that $\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]}$ is a GEE estimator $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_t]}$ of the best model selected from

Table 8. Selection probability and estimated prediction error with the MSCM data

Variables	PMSEG			QIC			mQIC		
	Auto	Ex	Ind	Auto	Ex	Ind	Auto	Ex	Ind
Race, Household, Stress, St1	55	7	0	0	0	74	0	0	74
Race, Household, Stress, St2	18	12	0	0	0	26	0	0	26
Race, Household, Stress, St1, St2	5	3	0	0	0	0	0	0	0
$\widehat{\mathrm{PEB}}_{\mathrm{P}}$	**1.92** (0.10)			4.39 (1.55)			4.39 (1.55)		
$\widehat{\mathrm{PEB}}_{\mathrm{Q}}$	**1001.88** (6.26)			1003.82 (6.31)			1003.82 (6.31)		

the calibration data, where $\hat{\boldsymbol{\beta}}_{[-\boldsymbol{d}_t]}$ is the solution of the following equation:

$$\sum_{i=1}^{146}(1-d_{it})\boldsymbol{D}_i^{\top}\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)=\mathbf{0}_p.$$

Finally, the estimated $\mathrm{PEB_P}$ and $\mathrm{PEB_Q}$ are given as

$$\widehat{\mathrm{PEB}}_{\mathrm{P}}=136\times\left\{\frac{1}{100}\sum_{t=1}^{100}\frac{1}{10}\mathscr{L}(\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]},\hat{\boldsymbol{\beta}}_{f,t},t)-9\right\},$$

$$\widehat{\mathrm{PEB}}_{\mathrm{Q}}=136\times\frac{1}{100}\sum_{t=1}^{100}\frac{1}{10}\left\{-2\sum_{i=1}^{146}\sum_{j=1}^{9}d_{it}\times Q(\hat{\boldsymbol{\beta}}_{\mathrm{B},[-\boldsymbol{d}_t]};y_{ij})\right\},$$

where $\mathscr{L}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,t)$ is defined as follows:

$$\mathscr{L}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,t)=\sum_{i=1}^{146}d_{it}\times\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_1)\}^{\top}\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2)\tilde{\boldsymbol{R}}^{-1}(\boldsymbol{\beta}_2)\boldsymbol{A}_i^{-1/2}(\boldsymbol{\beta}_2)\{\boldsymbol{y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta}_1)\}.$$

Note that we used the independence working correlation matrix for obtaining $\hat{\boldsymbol{\beta}}_f$ from the calibration data in this study. Also note that we denote $\hat{\boldsymbol{\beta}}_{f,t}$ as the value of $\hat{\boldsymbol{\beta}}_f$ from the calibration data at the tth iteration. From Table 8, we can see that the model most selected by each criterion is different. However, both the $\widehat{\mathrm{PEB}}_{\mathrm{P}}$ and $\widehat{\mathrm{PEB}}_{\mathrm{Q}}$ of the PMSEG were smaller than those of the QIC and mQIC. Hence, using the PMSEG is better than using the QIC (or mQIC) for selecting models in this study.

Consequently, from Tables 3, 4 and 8, we recommend the use of the PMSEG rather than the QIC (or mQIC) for selecting models in the GEE method.

5. Conclusion and discussion

In the present paper, we proposed the PMSEG as a model selection criterion that reflects the correlation in the GEE method. The PMSEG is the simple criterion such as the AIC. Nowadays, the GEE method is one of the mainstream of longitudinal analysis methods and many statistical softwares (e.g., SAS, R, etc.) support the GEE method. For these reasons, it is important to propose a more useful criterion for analyzing longitudinal data using the GEE method.

In all situations of the simulation results of Section 4, we showed that the PMSEG has better performance than the QIC and mQIC for the variable selection, and the difference between the performances of the PMSEG and both QIC and mQIC is more salient when the correlation is large. Recall that the

PMSEG reflects the correlation between responses, however, both the QIC and mQIC are not reflective. This is probably one of the reasons why the PMSEG has better performance than the QIC and mQIC. In the study of the MSCM data, we also showed that the PMSEG is useful as same as the QIC and mQIC. Nevertheless, computational costs of the PMSEG are lower than those of the QIC and mQIC because the bias term of the PMSEG is $2 \times$ (the number of parameters). On the other hand, many previous studies including the QIC and mQIC require the calculation of the bias term for each candidate model. Therefore, the PMSEG is useful and user friendly.

As for the future work, we will consider the following three problems. First, we recall that, in deriving the PMSEG, we assume that the nuisance parameter $\boldsymbol{\alpha}$ is known. Actually, we often estimate $\boldsymbol{\alpha}$ because $\boldsymbol{\alpha}$ is unknown in many cases. In fact, we estimate $\boldsymbol{\alpha}$ in Section 4. However, Liang and Zeger [11] showed that an estimator of $\boldsymbol{\alpha}$ is consistent under the standard assumption, and we confirmed that the estimation of $\boldsymbol{\alpha}$ dose not dramatically influence the performance of the PMSEG from some simulation results. Theoretical study of the influence of estimating $\boldsymbol{\alpha}$ to the PMSEG is left for the future work. Second, in this paper, we assume that all individuals have the same number of repeated observations and a common correlation structure. Since this is a strong assumption, we should consider to relax it. Finally, in Subsection 4.3, we confirmed the performance of the selection probability of the PMSEG when the number of occasions is not small. However, since this is merely a numerical result, we should show the theoretical property of the PMSEG.

Appendix

Derivation of (3)

The GEE is expanded as follows:

$$\begin{aligned}
\mathbf{0}_p &= \boldsymbol{s}_{n,0} + \left.\frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \otimes \boldsymbol{I}_p\} \left.\left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}^\top}\right)\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\star} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&= \boldsymbol{s}_{n,0} - \boldsymbol{H}_{n,0}(\boldsymbol{I}_p + \boldsymbol{G}_{1,0} + \boldsymbol{G}_{2,0})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \otimes \boldsymbol{I}_p\} \boldsymbol{L}_1(\boldsymbol{\beta}^\star)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \qquad (8)
\end{aligned}$$

where $\boldsymbol{\beta}^\star$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ (i.e., $^\exists\varepsilon_1 \in (0,1)$ s.t. $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_0 + \varepsilon_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$), $\boldsymbol{I}_p$ is a p-dimensional identity matrix, $\boldsymbol{G}_{1,0}$, $\boldsymbol{G}_{2,0}$ and $\boldsymbol{L}_1(\boldsymbol{\beta}^\star)$ are defined by

$$\boldsymbol{G}_{1,0} = -\boldsymbol{H}_{n,0}^{-1} \sum_{i=1}^{n} \boldsymbol{D}_{i,0}^{\top} \left(\frac{\partial}{\partial \boldsymbol{\beta}^{\top}} \otimes \boldsymbol{V}_i^{-1} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) \{ \boldsymbol{I}_p \otimes (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0}) \},$$

$$\boldsymbol{G}_{2,0} = -\boldsymbol{H}_{n,0}^{-1} \sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}^{\top}} \otimes \boldsymbol{D}_i^{\top} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \right) [\boldsymbol{I}_p \otimes \{ \boldsymbol{V}_{i,0}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0}) \}],$$

$$\boldsymbol{L}_1(\boldsymbol{\beta}^{\star}) = \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}^{\top}} \right) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star}}.$$

Note that $\boldsymbol{L}_1(\boldsymbol{\beta}^{\star}) = O_p(n)$, $\boldsymbol{s}_{n,0} = O_p(n^{1/2})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, $\boldsymbol{G}_{1,0}$ and $\boldsymbol{G}_{2,0} = O_p(n^{-1/2})$. Thus, (8) yields

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \boldsymbol{H}_{n,0}^{-1} \boldsymbol{s}_{n,0} + O_p(n^{-1}) = \boldsymbol{b}_{1,0} + O_p(n^{-1}). \tag{9}$$

Similarly, the GEE can also be expanded as follows:

$$\begin{aligned} \boldsymbol{s}_{n,0} = {} & \boldsymbol{H}_{n,0} (\boldsymbol{I}_p + \boldsymbol{G}_{1,0} + \boldsymbol{G}_{2,0}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ & - \frac{1}{2} \{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\top} \otimes \boldsymbol{I}_p \} \{ \boldsymbol{G}_{3,0} + (\boldsymbol{L}_1(\boldsymbol{\beta}_0) - \boldsymbol{G}_{3,0}) \} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ & - \frac{1}{6} \{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\top} \otimes \boldsymbol{I}_p \} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}^{\top}} \otimes \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{s}_n}{\partial \boldsymbol{\beta}^{\top}} \right) \right\} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{\star\star}} \\ & \times \{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \}, \end{aligned} \tag{10}$$

where $\boldsymbol{\beta}^{\star\star}$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ (i.e., $^{\exists}\varepsilon_2 \in (0,1)$ s.t. $\boldsymbol{\beta}^{\star\star} = \boldsymbol{\beta}_0 + \varepsilon_2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$), and $\boldsymbol{G}_{3,0} = \mathrm{E}\{\boldsymbol{L}_1(\boldsymbol{\beta}_0)\}$.

Note that the order of the last term of equation (10) is $O_p(n^{-1/2})$. Also note that $\boldsymbol{G}_{3,0} = O(n)$ and $\boldsymbol{L}_1(\boldsymbol{\beta}_0) - \boldsymbol{G}_{3,0} = O_p(n^{1/2})$. Therefore, by using equation (9) and (10), we get the stochastic expansion (3), and $\boldsymbol{b}_{1,0} = O(n^{-1/2})$ and $\boldsymbol{b}_{2,0} = O(n^{-1})$.

Calculation of Bias1

Bias1 in (4) is expanded as

$$\begin{aligned} \text{Bias1} &= \mathrm{E}_y \left[\mathrm{E}_z \left\{ \sum_{i=1}^{n} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i)^{\top} \boldsymbol{\Sigma}_{i,0}^{-1} (\boldsymbol{z}_i - \hat{\boldsymbol{\mu}}_i) \right\} - \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)^{\top} \boldsymbol{\Sigma}_{i,0}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i) \right] \\ &= 2\mathrm{E}_y \left\{ \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^{\top} \boldsymbol{\Sigma}_{i,0}^{-1} (\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_{i,0}) \right\}. \end{aligned} \tag{11}$$

Applying Taylor's expansion around $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ to $\hat{\boldsymbol{\mu}}_i$ yields

$$\begin{aligned}\hat{\boldsymbol{\mu}}_i &= \boldsymbol{\mu}_{i,0} + \left.\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\quad + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \otimes \boldsymbol{I}_m\} \left.\left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\right)\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\quad + \frac{1}{6}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \otimes \boldsymbol{I}_m\} \left.\left\{\frac{\partial}{\partial \boldsymbol{\beta}^\top} \otimes \left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^\top}\right)\right\}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{***}} \{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \otimes (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} \\ &= \boldsymbol{\mu}_{i,0} + \boldsymbol{D}_{i,0}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \otimes \boldsymbol{I}_m\} \boldsymbol{D}_{i,0}^{(1)} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + O_p(n^{-3/2}), \end{aligned} \tag{12}$$

where $\boldsymbol{D}_{i,0}^{(1)}$ is given by

$$\boldsymbol{D}_{i,0}^{(1)} = \left.\left(\frac{\partial}{\partial \boldsymbol{\beta}} \otimes \boldsymbol{D}_i\right)\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

Here, $\boldsymbol{\beta}^{***}$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Substituting the stochastic expansion of $\hat{\boldsymbol{\beta}}$ in (3) into (12) yields the following:

$$\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_{i,0} = \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0} + \left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}^\top \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\} + O_p(n^{-3/2}). \tag{13}$$

By combining (11) and (13), we obtain

$$\begin{aligned}\frac{1}{2}\mathrm{Bias1} &= \mathrm{E}_y\left\{\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right\} \\ &\quad + \mathrm{E}_y\left[\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1}\left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0} + \frac{1}{2}(\boldsymbol{b}_{1,0}^\top \otimes \boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\}\right] \\ &\quad + \mathrm{E}_y\{O_p(n^{-1/2})\}. \end{aligned} \tag{14}$$

Note that $\mathrm{E}\{(\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top (\boldsymbol{y}_j - \boldsymbol{\mu}_{j,0})\} = 0$, $(i \neq j)$, the first term of (14) can be calculated as

$$\begin{aligned}&\mathrm{E}_y\left\{\sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right\} \\ &\quad = \mathrm{E}_y\left[\sum_{i=1}^n \sum_{j=1}^n (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0} \boldsymbol{H}_{n,0}^{-1} \boldsymbol{D}_{j,0}^\top \boldsymbol{V}_{j,0}^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_{j,0})\right] \end{aligned}$$

$$= \mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1}\boldsymbol{D}_{i,0}\boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}^\top \boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\right]$$

$$= \mathrm{tr}\left[\boldsymbol{H}_{n,0}^{-1}\sum_{i=1}^{n}\boldsymbol{D}_{i,0}^\top \boldsymbol{V}_{i,0}^{-1}\boldsymbol{D}_{i,0}\right] = \mathrm{tr}[\boldsymbol{H}_{n,0}^{-1}\boldsymbol{H}_{n,0}] = p. \tag{15}$$

Similarly, since $\mathrm{E}[(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\otimes(\boldsymbol{y}_j-\boldsymbol{\mu}_{j,0})^\top(\boldsymbol{y}_k-\boldsymbol{\mu}_{k,0})] = \boldsymbol{0}_m$, (not $i=j=k$), the second term of (14) can be expanded as

$$\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1}\left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0}+\frac{1}{2}(\boldsymbol{b}_{1,0}^\top\otimes\boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\}\right]$$

$$= \mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1}\left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2i,0}+\frac{1}{2}(\boldsymbol{b}_{1i,0}^\top\otimes\boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1i,0}\right\}\right],$$

where $\boldsymbol{b}_{1i,0} = \boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}^\top \boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})$, $\boldsymbol{b}_{2i,0} = \boldsymbol{H}_{n,0}^{-1}(\boldsymbol{b}_{1i,0}^\top\otimes\boldsymbol{I}_p)\boldsymbol{G}_{3,0}\boldsymbol{b}_{1i,0}/2 - \boldsymbol{G}_{1i,0}\boldsymbol{b}_{1i,0} - \boldsymbol{G}_{2i,0}\boldsymbol{b}_{1i,0}$,

$$\boldsymbol{G}_{1i,0} = -\boldsymbol{H}_{n,0}^{-1}\boldsymbol{D}_{i,0}^\top\left(\frac{\partial}{\partial\boldsymbol{\beta}^\top}\otimes\boldsymbol{V}_i^{-1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)\{\boldsymbol{I}_p\otimes(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\},$$

$$\boldsymbol{G}_{2i,0} = -\boldsymbol{H}_{n,0}^{-1}\left(\frac{\partial}{\partial\boldsymbol{\beta}^\top}\otimes\boldsymbol{D}_i^\top\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)[\boldsymbol{I}_p\otimes\{\boldsymbol{V}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\}].$$

Note that $\boldsymbol{D}_{i,0}\boldsymbol{b}_{2i,0} + (\boldsymbol{b}_{1i,0}^\top\otimes\boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1i,0}/2 = O_p(n^{-2})$, the second term of (14) can be obtained as

$$\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top \boldsymbol{\Sigma}_{i,0}^{-1}\left\{\boldsymbol{D}_{i,0}\boldsymbol{b}_{2,0}+\frac{1}{2}(\boldsymbol{b}_{1,0}^\top\otimes\boldsymbol{I}_m)\boldsymbol{D}_{i,0}^{(1)}\boldsymbol{b}_{1,0}\right\}\right] = O(n^{-1}). \tag{16}$$

Under certain conditions, the limit of the expectation is equal to the expectation of the limit. Furthermore, in many cases of practical interest, a moment of statistic can be expanded as power series in n^{-1} (see e.g., Hall, [7]). Therefore, by substituting (15) and (16) into (14), we obtain the asymptotic expansion of Bias1 up to order 1 as

$$\mathrm{Bias1} = 2p + O(n^{-1}). \tag{17}$$

Calculation of Bias2 + Bias4

We calculate Bias2 + Bias4. Bias2 in (4) can be calculated as

$$\begin{aligned}\mathrm{Bias2} &= \mathrm{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)^\top\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)-\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\right\}\\ &= \mathrm{E}_y\left\{2\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right\}\\ &\quad+\mathrm{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)^\top\boldsymbol{\Sigma}_{i,0}^{-1}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right\},\end{aligned}$$

and Bias4 in (4) can also be calculated as

$$\begin{aligned}\mathrm{Bias4} &= \mathrm{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})\right\}\\ &\quad-\mathrm{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)^\top\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)\right\}\\ &= -\mathrm{E}_y\left\{2\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right\}\\ &\quad-\mathrm{E}_y\left\{\sum_{i=1}^{n}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)^\top\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right\}.\end{aligned}$$

Thus, we obtain Bias2 + Bias4 as follows:

$$\begin{aligned}\mathrm{Bias2}+\mathrm{Bias4} &= \mathrm{E}_y\left[2\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^\top\{\boldsymbol{\Sigma}_{i,0}^{-1}-\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\right.\\ &\qquad\left.\times\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right] \qquad (18)\\ &\quad+\mathrm{E}_y\left[\sum_{i=1}^{n}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)^\top\{\boldsymbol{\Sigma}_{i,0}^{-1}-\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\right.\\ &\qquad\left.\times\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\}(\boldsymbol{\mu}_{i,0}-\hat{\boldsymbol{\mu}}_i)\right]. \qquad (19)\end{aligned}$$

In order to calculate (18) and (19), we perform the stochastic expansion of $\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$, $\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)$, $\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f)$ and $\hat{\boldsymbol{\beta}}_f$. Denote $\boldsymbol{D}_{*,i}=\boldsymbol{A}_i(\boldsymbol{\beta}_*)\boldsymbol{\Delta}_i(\boldsymbol{\beta}_*)\boldsymbol{X}_{*,i}$ and $\boldsymbol{D}_{*,i,0}=\boldsymbol{A}_{i,0}\boldsymbol{\Delta}_{i,0}\boldsymbol{X}_{*,i}$. Considering the same argument in section 2 and 3, $\hat{\boldsymbol{\beta}}_f$ and $\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f)$ can be expanded as

$$\hat{\boldsymbol{\beta}}_f - \boldsymbol{\beta}_{*,0} = \boldsymbol{b}_{f,0} + O_p(n^{-1}), \qquad \boldsymbol{b}_{f,0} = \boldsymbol{H}_{f,n,0}^{-1}\boldsymbol{s}_{f,n}(\boldsymbol{\beta}_{*,0}),$$
$$\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}_f) - \boldsymbol{\mu}_{i,0} = \boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0} + O_p(n^{-1}), \tag{20}$$

where $\boldsymbol{\beta}_{*,0}$ is the true value of $\boldsymbol{\beta}_*$, and $\boldsymbol{H}_{f,n,0}$ is defined as

$$\boldsymbol{H}_{f,n,0} = \sum_{i=1}^{n} \boldsymbol{D}_{*,i,0}^{\top}\boldsymbol{A}_{i,0}^{-1/2}\overline{\boldsymbol{R}}_i^{-1}(\boldsymbol{\alpha}_\dagger)\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{D}_{*,i,0}.$$

Let $\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)$ denote the m-dimensional vector and the jth element of which is the (j,j)th element of $\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$. Note that $\mathrm{diag}\{\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)\} = \boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)$. By applying Taylor's expansion around $\hat{\boldsymbol{\beta}}_f = \boldsymbol{\beta}_{*,0}$, $\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)$ is expanded as

$$\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f) = \boldsymbol{a}_{f,i}(\boldsymbol{\beta}_{*,0}) + \boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0} + O_p(n^{-1}), \qquad \boldsymbol{A}_{f,i,0}^{*} = \frac{\partial}{\partial\boldsymbol{\beta}_*^{\top}}\boldsymbol{a}_{f,i}(\boldsymbol{\beta}_*)\bigg|_{\boldsymbol{\beta}_*=\boldsymbol{\beta}_{*,0}}.$$

Hence, we obtain

$$\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f) = \mathrm{diag}\{\boldsymbol{a}_{f,i}(\hat{\boldsymbol{\beta}}_f)\} = \boldsymbol{A}_{i,0}^{-1/2} + \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0}) + O_p(n^{-1}). \tag{21}$$

Note that $\boldsymbol{b}_{f,0}, \boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0}, \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0}) = O_p(n^{-1/2})$. Moreover, substituting (20) and (21) into $\tilde{\boldsymbol{R}}(\hat{\boldsymbol{\beta}}_f)$ yields following:

$$\begin{aligned}
\tilde{\boldsymbol{R}}(\hat{\boldsymbol{\beta}}_f) = &-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}\{\boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top} + (\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{D}_{*,i,0}\boldsymbol{b}_{f,0})^{\top}\}\boldsymbol{A}_{i,0}^{-1/2} \\
&+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2} \\
&+\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2} \\
&+\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*}\boldsymbol{b}_{f,0}) + O_p(n^{-1}).
\end{aligned} \tag{22}$$

By the Lindeberg central limit theorem, the first term of (22) is $O_p(n^{-1})$. Thus, using this fact and (22), we obtain

$$\begin{aligned}
&\boldsymbol{R}_0^{-1/2}\tilde{\boldsymbol{R}}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1/2} \\
&\quad = \boldsymbol{I}_m - \boldsymbol{R}_0^{-1/2}\left\{\boldsymbol{R}_0 - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}\right.
\end{aligned}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\,\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1/2}+O_p(n^{-1}).$$

Therefore, by calculating the inverse matrix of $\boldsymbol{R}_0^{-1/2}\tilde{\boldsymbol{R}}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{R}_0^{-1/2}$, $\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)$ can be expanded as

$$\begin{aligned}\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)=\boldsymbol{R}_0^{-1}+\boldsymbol{R}_0^{-1}\Bigg\{&\boldsymbol{R}_0-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}\\&-\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}\\&-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\,\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1}\\&+O_p(n^{-1}).\end{aligned}\tag{23}$$

Note that the second term of (23) is $O_p(n^{-1/2})$.

Next, we calculate (19). By using (21) and (23), we obtain

$$\begin{aligned}&\boldsymbol{\Sigma}_{i,0}^{-1}-\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\\&\quad=-\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}-\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\,\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\\&\qquad-\boldsymbol{A}_{i,0}^{-1/2}\boldsymbol{R}_0^{-1}\Bigg\{\boldsymbol{R}_0-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}\\&\qquad-\frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\boldsymbol{A}_{i,0}^{-1/2}\\&\qquad-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,0}^{-1/2}(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})(\boldsymbol{y}_i-\boldsymbol{\mu}_{i,0})^{\top}\,\mathrm{diag}(\boldsymbol{A}^{*}_{f,i,0}\boldsymbol{b}_{f,0})\Bigg\}\boldsymbol{R}_0^{-1}\boldsymbol{A}_{i,0}^{-1/2}\\&\qquad+O_p(n^{-1}).\end{aligned}\tag{24}$$

Note that $\boldsymbol{\Sigma}_{i,0}^{-1}-\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)\tilde{\boldsymbol{R}}^{-1}(\hat{\boldsymbol{\beta}}_f)\boldsymbol{A}_i^{-1/2}(\hat{\boldsymbol{\beta}}_f)=O_p(n^{-1/2})$ and $\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}=O_p(n^{-1/2})$. Therefore, by substituting (13) and (24) into (19), we obtain

$$(19)=O(n^{-1}).\tag{25}$$

Recall that, in general, a moment of statistic can be expanded as power series in n^{-1}. Hence, the order of (25) is shown by $O(n^{-1})$, not $O(n^{-1/2})$.

Finally, we calculate (18). By substituting (13) and (24) into (18), we obtain

$$
\begin{aligned}
(18) = \mathrm{E}_y \Bigg[& 2\sum_{i=1}^{n} \boldsymbol{\kappa}_{i,0}^{\top} \{ \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0}) \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \\
& + \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0}) \} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \Bigg] \\
& - \mathrm{E}_y \left(2\sum_{i=1}^{n} \boldsymbol{\kappa}_{i,0}^{\top} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^{\top} \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \right) \\
& - \mathrm{E}_y \left\{ \frac{2}{n} \sum_{i,j} \boldsymbol{\kappa}_{i,0}^{\top} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \mathrm{diag}(\boldsymbol{A}_{f,j,0}^{*} \boldsymbol{b}_{f,0}) \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^{\top} \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \right\} \\
& - \mathrm{E}_y \left\{ \frac{2}{n} \sum_{i,j} \boldsymbol{\kappa}_{i,0}^{\top} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^{\top} \mathrm{diag}(\boldsymbol{A}_{f,j,0}^{*} \boldsymbol{b}_{f,0}) \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \right\} \\
& + \mathrm{E}_y \left(2\sum_{i=1}^{n} \boldsymbol{\kappa}_{i,0}^{\top} \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \right) + O(n^{-1}). \qquad (26)
\end{aligned}
$$

Here, $\boldsymbol{\kappa}_{i,0} = (\boldsymbol{y}_i - \boldsymbol{\mu}_{i,0})$ and $\sum_{i,j} = \sum_{i=1}^{n} \sum_{j=1}^{n}$. Moreover, in order to simplify the calculation, we define the following notation:

$$
\sum_{i \neq j} = \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n}, \qquad \boldsymbol{b}_{f,i,0} = \boldsymbol{H}_{f,n,0}^{-1} \boldsymbol{D}_{*,i,0}^{\top} \boldsymbol{A}_{i,0}^{-1} \boldsymbol{\kappa}_{i,0}.
$$

Recall that $\mathrm{E}(\boldsymbol{\kappa}_{i,0} \otimes \boldsymbol{\kappa}_{j,0}^{\top} \boldsymbol{\kappa}_{k,0}) = \boldsymbol{0}_m$, (not $i = j = k$). Hence, the first term of (26) is as follows:

$$
\begin{aligned}
& \mathrm{E}_y \left[2\sum_{i=1}^{n} \boldsymbol{\kappa}_{i,0}^{\top} \{ \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0}) \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} + \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,0}) \} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0} \right] \\
& = \mathrm{E}_y \Bigg[2\sum_{i=1}^{n} \boldsymbol{\kappa}_{i,0}^{\top} \{ \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,i,0}) \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \\
& \quad + \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \mathrm{diag}(\boldsymbol{A}_{f,i,0}^{*} \boldsymbol{b}_{f,i,0}) \} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1i,0} \Bigg] = O(n^{-1}). \qquad (27)
\end{aligned}
$$

Similarly, since $\mathrm{E}_y(\boldsymbol{\kappa}_{i,0}^\top \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{\kappa}_{k,0}) = 0$ unless $i = k$, the second term of (26) can be calculated as

$$
\begin{aligned}
&-\mathrm{E}_y\left(2\sum_{i=1}^n \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0}\right) \\
&\quad = -\mathrm{E}_y\left(2\sum_{i=1}^n \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \frac{1}{n}\sum_{j=1, i\neq j}^n \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1i,0}\right) \\
&\qquad + O(n^{-1}) \\
&\quad = -\mathrm{E}_y\left(2\sum_{i=1}^n \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{\Sigma}_{i,0}^{-1} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1i,0}\right) + O(n^{-1}) = -2p + O(n^{-1}). \qquad (28)
\end{aligned}
$$

Note that $-2p$ in (28) is obtained from (15). Furthermore, $\mathrm{E}_y[\boldsymbol{\kappa}_{i,0}^\top (\boldsymbol{\kappa}_{j,0} \otimes \boldsymbol{\kappa}_{k,0}^\top) \cdot (\boldsymbol{\kappa}_{k,0} \otimes \boldsymbol{\kappa}_{l,0})] = 0$ expect in the following cases:

$$
i = j = l \ \text{ or } \ i = j \neq k = l \ \text{ or } \ i = l \neq k = j \ \text{ or } \ j = l \neq k = i.
$$

Thus, the third term of (26) is given by

$$
\begin{aligned}
&-\mathrm{E}_y\left\{\frac{2}{n}\sum_{i,j} \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \operatorname{diag}(\boldsymbol{A}_{f,j,0}^* \boldsymbol{b}_{f,0}) \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0}\right\} \\
&\quad = -\mathrm{E}_y\left\{\frac{2}{n}\sum_{i=1}^n \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \operatorname{diag}(\boldsymbol{A}_{f,i,0}^* \boldsymbol{b}_{f,0}) \boldsymbol{\kappa}_{i,0} \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0}\right\} \\
&\qquad - \mathrm{E}_y\left\{\frac{2}{n}\sum_{i\neq j} \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \operatorname{diag}(\boldsymbol{A}_{f,j,0}^* \boldsymbol{b}_{f,i,0}) \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1j,0}\right\} \\
&\qquad - \mathrm{E}_y\left\{\frac{2}{n}\sum_{i\neq j} \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \operatorname{diag}(\boldsymbol{A}_{f,j,0}^* \boldsymbol{b}_{f,j,0}) \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1i,0}\right\} \\
&\qquad + O(n^{-1}) \\
&\quad = O(n^{-1}). \qquad (29)
\end{aligned}
$$

In the same manner as in the calculation of the third term of (26), the fourth term of (26) is calculated as

$$
\begin{aligned}
&-\mathrm{E}_y\left\{\frac{2}{n}\sum_{i,j} \boldsymbol{\kappa}_{i,0}^\top \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{R}_0^{-1} \boldsymbol{A}_{j,0}^{-1/2} \boldsymbol{\kappa}_{j,0} \boldsymbol{\kappa}_{j,0}^\top \operatorname{diag}(\boldsymbol{A}_{f,j,0}^* \boldsymbol{b}_{f,0}) \boldsymbol{R}_0^{-1} \boldsymbol{A}_{i,0}^{-1/2} \boldsymbol{D}_{i,0} \boldsymbol{b}_{1,0}\right\} \\
&\quad = O(n^{-1}). \qquad (30)
\end{aligned}
$$

Moreover, the fifth term of (26) is obtained from (15), as follows:

$$\mathrm{E}_y\left(2\sum_{i=1}^{n}\boldsymbol{\kappa}_{i,0}^{\top}\boldsymbol{\Sigma}_{i,0}^{-1}\boldsymbol{D}_{i,0}\boldsymbol{b}_{1,0}\right)=2p. \tag{31}$$

Substituting (27), (28), (29), (30) and (31) into (26), (18) is given by

$$(18)=O(n^{-1}). \tag{32}$$

Consequently, from (25) and (32), we obtain $\mathrm{Bias2}+\mathrm{Bias4}=O(n^{-1})$.

Acknowledgement

The authors are grateful to the referees for their valuable comments and suggestions. We would also like to thank Professor Hirofumi Wakaki of Hiroshima University for his helpful comments and suggestions.

References

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

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

[3] C. S. Alexander and R. Markowitz, Maternal employment and use of pediatric clinic services, Medical Care, **24** (1986), 134–147.

[4] E. Cantoni, J. M. Flemming and E. Ronchetti, Variable selection for marginal longitudinal generalized linear models, Biometrics, **61** (2005), 507–514.

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

[6] M. Gosho, C. Hamada and I. Yoshimura, Modifications of QIC and CIC for selecting a working correlation structure in the generalized estimating equation method, Japanese Journal of Biometrics, **32** (2011), 1–12.

[7] P. Hall, The bootstrap and edgeworth expansion, Springer-Verlag, New York, 1992.

[8] L. Y. Hin and Y. G. Wang, Working-correlation-structure identification in generalized estimating equations, Statistics in Medicine, **28** (2009), 642–658.

[9] S. Imori, Model selection criterion based on the multivariate quasi-likelihood for generalized estimating equations, Scand. J. Stat., **42** (2015), 1214–1224.

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

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

[12] C. L. Mallows, Some comments on C_p, Technometrics, **15** (1973), 661–675.

[13] J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, J. R. Statist. Soc. ser. A, **135** (1972), 370–384.

[14] R. Nishii, Asymptotic properties of criteria for selecting of variables in multiple regression, Ann. Statist., **12** (1984), 758–765.
[15] W. Pan, Akaike's information criterion in generalized estimating equations, Biometrics, **57** (2001), 120–125.
[16] C. R. Rao and Y. Wu, A strongly consistent procedure for model selection in a regression problem, Biometrika, **76** (1989), 369–374.
[17] R. W. M. Wedderburn, Quasi-likelihood functions, generalized linear models, and Gauss-Newton method, Biometrika, **61** (1974), 439–447.
[18] M. Xie and Y. Yang, Asymptotics for generalized estimating equations with large cluster sizes, Ann. Statist., **31** (2003), 310–347.
[19] H. Yanagihara, H. Wakaki and Y. Fujikoshi, A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large, Electron. J. Stat., **9** (2015), 869–897.

Yu Inatsu
RIKEN Center for Advanced Intelligence Project
1-4-1 *Nihonbashi, Chuo-ku, Tokyo,* 103-0027, *Japan*
E-mail: yu.inatsu@riken.jp

Shinpei Imori
Graduate School of Science
Hiroshima University
1-3-1 *Kagamiyama, Higashi-Hiroshima, Hiroshima,* 739-8526, *Japan*
E-mail: imori@hiroshima-u.ac.jp