Statistica Sinica **27** (2017), 389-413
doi:http://dx.doi.org/10.5705/ss.2013.277

VARIABLE SELECTION AND MODEL AVERAGING FOR LONGITUDINAL DATA INCORPORATING GEE APPROACH

Hui Yang, Peng Lin, Guohua Zou and Hua Liang

Amgen Inc., Shandong University of Technology, Capital Normal University and George Washington University

Abstract: The Akaike Criterion, which is based on maximum likelihood estimation and cannot be applied directly to the situations when likelihood functions are not available, has been modified for variable selection in longitudinal data with generalized estimating equations via a working independence model. This paper proposes another modification to AIC, the difference between the quasi-likelihood functions of a candidate model and of a narrow model plus a penalty term. Such a difference avoids calculating complex integration from quasi-likelihood, but inherits theoretical asymptotic properties from AIC. We also propose a focused information criterion for variable selection on the basis of the quasi-score function. Further, this paper develops a frequentist model average estimator for longitudinal data with generalized estimating equations. Simulation studies provide evidence of the superiority of the proposed procedures. The procedures are further applied to a data example.

Key words and phrases: FIC, local misspecification, marginal likelihood, model averaging, QIC, quasi-likelihood, working independence.

1. Introduction

Longitudinal data, in the form of repeated measurements on the same unit over time or space, arise in a broad range of fields including medical and public health research. An example is an AIDS clinical study A5055, that aimed to predict the long-term antiviral treatment responses of HIV-1 infected patients by considering pharmacokinetics (PK), drug adherence, and susceptibility. In this study, each patient was visited multiple times over 24 weeks after entry; the observations' correlations within each patient were expected and had to be taken into account during the analysis.

Mixed-effects models, introduced by Laird and Ware (1982), have been widely used for analyzing longitudinal data. As a likelihood-based approach, it relies on the assumption that data are drawn from some distributions of known form, which may be unknown in reality. Even if the distributions are specified, it

can still be challenging to derive the full likelihood, especially for non-Gaussian data. Instead of specifying the full joint distribution of responses, Liang and Zeger (1986) developed generalized estimating equations (GEE) approach that provides consistent estimates by only specifying the first two marginal moments and a working correlation matrix; when the specified correlation is the true correlation, the estimates are most efficient.

Since no assumption is made about the distributions of the longitudinal data in GEE-based methods, such traditional likelihood-based model selection criteria, as Akaike's (1974) Information Criterion (AIC), Schwarz's (1978) Bayes Information Criterion (BIC), and Mallows' (1973) C_p, cannot be applied directly for model selection incorporating the GEE approach. As a remedy, Pan (2001a) developed a quasi-likelihood-based AIC-type information criterion, known as QIC, by replacing the likelihood component in AIC with the quasi-likelihood under a working *independence* model. Such a replacement, along with the independence setup, makes implementation simpler with a acceptable loss. As a consequence, QIC can be easily computed using some well-developed statistical packages such as S-plus/R and SAS. The negligence of significant part in its derivation and its reliance on working independence make QIC lack theoretical asymptotic properties. Cantoni, Flemming and Ronchetti (2005) proposed a generalized C_p criterion by using weighted quadratic predictive risk as a measure of model's adequacy for prediction. This requires bootstrap sampling or Monte Carlo simulation, which can be computationally expensive. Another extended cross-validation approach based on expected predictive bias was suggested by Pan (2001b). Fu (2003) proposed penalized generalized estimating equations for variable selection, and Wang and Qu (2009) proposed a BIC-type model selection criterion based on a quadratic inference function. They both require an extra searching algorithm for the tuning parameter.

This article proposes a quasi-likelihood-based AIC-type variable selection criterion for longitudinal data incorporating the GEE approach. We choose a narrow model as a benchmark and consider the quasi-likelihood difference between a candidate model and the narrow model; this can avoid the complicated calculation of the full quasi-likelihood and make the implementation feasible and simpier. The idea was inspired by the local misspecification framework setting in Hjort and Claeskens (2003). Under certain regularity conditions, the proposed criterion is shown to have similar asymptotic properties as AIC.

The criteria mentioned are data-oriented and select the model with the best overall fit, regardless of the different estimation interests. However, as Hansen (2005) pointed out, "models should be evaluated based on their purpose", different parameters of interest may result in different models. From this perspective, Claeskens and Hjort (2003) proposed a focused information criterion (FIC) that

leads to the model with the smallest estimated mean square error of a focus parameter's estimation, and developed the corresponding large sample properties. Following Claeskens and Hjort (2003), we propose the quasi-likelihood based focused information criterion (QFIC) for longitudinal data, incorporating the GEE approach.

One concern about a model selection procedure is over-optimistic confidence intervals. Inference based on a single final model ignores the uncertainty introduced by the model selection process and underestimates the variability. The corresponding confidence intervals are either too narrow or their shift from the correct location, as shown in Danilov and Magnus (2004b), Shen, Huang and Ye (2004). Instead of relying on one model, model averaging procedure combines the estimates from different models in the form of a certain weighting mean. By avoiding the model selection process, the corresponding inference reduces the risk of ending up with the bad model and improves coverage probability. This strategy has been studied by such as Draper (1995), Buckland, Burnham and Augustin (1997), Burnham and Anderson (1998), Danilov and Magnus (2004a), and Leeb and Pötscher (2006). Hjort and Claeskens (2003) also considered a frequentist model averaging (FMA) procedure using the weights obtained based on certain model selection criteria, and derived nice asymptotic properties of it. Another purpose of our article is to develop the quasi-likelihood based frequentist model averaging procedure (QFMA) for longitudinal data incorporating the GEE approach which will inherit some good asymptotic properties due to the similarity of quasi-likelihood and likelihood.

This paper is organized as follows. In Section 2, we give a brief review on generalized estimation equations and QIC. Section 3 proposes a new variable selection criterion, ΔAIC, and provides the corresponding theoretical insights. Section 4 introduces the QFIC procedure. Section 5 considers the QFMA procedure and constructs the modified confidence intervals based on QFMA estimation. Simulation studies and the A5055 data analysis are reported in Sections 6 and 7, respectively. In the final section, we conclude with some remarks. The proofs of results are contained in the online Supplemental Material.

2. Generalized Estimating Equations and QIC

2.1. Generalized estimating equations

Consider a longitudinal study with n subjects and m_i visits for the ith subject: y_{ij} and $\mathbf{x}_{ij}$ are the response and a set of the covariates (fixed) for the ith subject at the jth visit; the mean of y_{ij}, denoted by μ_{ij}, can be connected to $\mathbf{x}_{ij}$ through a link function $g(\cdot)$:

$$\mathrm{E}(y_{ij}) = \mu_{ij} \quad \text{and} \quad g(\mu_{ij}) = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ is a vector of unknown parameters; the variance of y_{ij} can be expressed as a known function $\nu(\cdot)$ of μ_{ij}, with a nuisance parameter ϕ, $\mathrm{Var}\,(y_{ij}) = \phi\nu(\mu_{ij})$. Starting from these basic assumptions, Wedderburn (1974) defined the log quasi-likelihood function $K(\mu_{ij}, \phi, y_{ij})$ through the relationship

$$\frac{\partial K(\mu_{ij}, \phi; y_{ij})}{\partial \mu_{ij}} = \frac{y_{ij} - \mu_{ij}}{\phi\nu(\mu_{ij})}.$$

Let $\mathbf{y}_i = (y_{i1} \cdots, y_{im_i})^\top$ and $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})^\top$. In the context of longitudinal data incorporating the GEE approach, the log quasi-likelihood function can be defined similarly as

$$\frac{\partial Q(\boldsymbol{\beta}, \mathbf{R}_i(\boldsymbol{\alpha}), \phi; \mathbf{y}_i)}{\partial \boldsymbol{\beta}} = \mathbf{D}_i^\top \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{y}_i)$, $\mathbf{D}_i = \mathbf{D}_i(\boldsymbol{\beta}) = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^\top$, $\mathbf{V}_i = \phi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$, $\mathbf{R}_i(\boldsymbol{\alpha})$ is an $m_i \times m_i$ working correlation matrix, and $\mathbf{A}_i$ is an $m_i \times m_i$ diagonal matrix with the jth diagonal element $\nu(\mu_{ij})$. If $\mathcal{D} = \{(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_n, \mathbf{x}_n)\}$, the estimation of $\boldsymbol{\beta}$ can be reached by solving the corresponding quasi-score equations, known as generalized estimating equations (GEE),

$$\mathbf{U}(\boldsymbol{\beta}, \mathbf{R}(\boldsymbol{\alpha}), \phi; \mathcal{D}) = \sum_{i=1}^{n} \mathbf{D}_i^\top \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

The main advantage of the GEE estimator $\widehat{\boldsymbol{\beta}}_{\text{gee}}$ is consistency under the mild regularity conditions, regardless of the misspecified working correlation matrix. It has also been shown that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{gee}} - \boldsymbol{\beta})$ is asymptotically normal with mean zero and variance-covariance matrix $\mathbf{V}_{\text{gee}}$, where

$$\mathbf{V}_{\text{gee}} = \lim_{n\to\infty} n\Bigg(\sum_{i=1}^{n} \mathbf{D}_i^\top \mathbf{V}_i^{-1}\mathbf{D}_i\Bigg)^{-1}\Bigg\{\sum_{i=1}^{n} \mathbf{D}_i^\top \mathbf{V}_i^{-1}\,\mathrm{Cov}\,(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\Bigg\} \times\Bigg(\sum_{i=1}^{n} \mathbf{D}_i^\top \mathbf{V}_i^{-1}\mathbf{D}_i\Bigg)^{-1}.$$

By replacing $\mathrm{Cov}\,(\mathbf{y}_i)$ with $\{\mathbf{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}})\}\{\mathbf{y}_i - \boldsymbol{\mu}_i(\widehat{\boldsymbol{\beta}})\}^\top$, and substituting $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and ϕ by their $\sqrt{n}$-consistent estimators, $\mathbf{V}_{\text{gee}}$ can be estimated consistently, where the estimator is known as the sandwich estimator or the robust variance-covariance estimator (White (1980)).

Liang and Zeger (1986) suggested several commonly used working correlation matrices: the independent working correlation matrix (IN) with $\mathbf{R}_i = \mathbf{I}_{m_i}$, the exchangeable working correlation matrix (EX) with $[\mathbf{R}_i]_{jk} = \alpha$ $(j \neq k)$, the first-order autoregressive working correlation matrix (AR) with $[\mathbf{R}_i]_{jk} = \alpha^{|j-k|}$ $(j \neq k)$, and the unstructured working correlation matrix (UN) with $[\mathbf{R}_i]_{jk} = \alpha_{jk}$

($j \neq k$). Although the GEE approach provides robust estimates regardless of the choice of $\mathbf{R}_i$, choosing one that is close to the true correlation can increase efficiency.

2.2. Akaike information criterion in generalized estimating equations

For longitudinal data incorporating the GEE approach, no assumption is made about the distributions of the responses, so the likelihood-based Akaike Information Criterion (AIC), cannot be applied directly. Fortunately, Wedderburn has shown some similar properties in using the quasi-likelihood function. Accordingly, one might replace the log likelihood component in AIC with the log quasi-likelihood (assuming its existence) as

$$\text{QAIC} = -2Q(\widehat{\boldsymbol{\beta}}_{\text{gee}}, \mathbf{R}(\widehat{\boldsymbol{\alpha}}), \widehat{\phi}; \mathcal{D}) + 2k. \tag{2.1}$$

As the correlation structure of longitudinal data is usually complex, it is hard to calculate the log quasi-likelihood component in (2.1), especially for large m_i. To simplify the calculation, Pan (2001a) proposed using the quasi-likelihood under the independence model criterion

$$\text{QIC}(\mathbf{R}) = -2Q(\widehat{\boldsymbol{\beta}}_{\text{gee}}(\mathbf{R}), \mathbf{I}, \widehat{\phi}; \mathcal{D}) + 2\text{trace}(\widehat{\boldsymbol{\Omega}}_{\text{I}}\widehat{\mathbf{V}}_{\text{gee}}).$$

Here, $\widehat{\boldsymbol{\beta}}_{\text{gee}}(\mathbf{R})$ and the sandwich estimator $\widehat{\mathbf{V}}_{\text{gee}}$ are obtained with the working correlation $\mathbf{R}$, while $Q(\widehat{\boldsymbol{\beta}}_{\text{gee}}(\mathbf{R}), \mathbf{I}, \widehat{\phi}; \mathcal{D})$ is reached with working independence, likewise $\widehat{\boldsymbol{\Omega}}_{\text{I}}$, the inverse of the sandwich estimator of $\widehat{\boldsymbol{\beta}}_{\text{gee}}(\mathbf{I})$. Then QIC picks the model with the smallest QIC value.

By using the independent working correlation and ignoring its complex, though significant, role in the expected K-L distance during its derivation, QIC becomes feasible. This simplification lacks of asymptotic properties in theory. We propose another quasi-likelihood-based model selection criterion for longitudinal data incorporating the GEE approach that is theoretically well-behaved in large sample contexts and numerically superior to its competitors even in moderate-size samples.

3. AIC-Type Variable Selection Criterion Incorporating GEE Approach

Claeskens and Hjort (2008) pointed out that among all the candidate models, when the true model is at a fixed distance from the narrow model, with a large sample size the dominating bias always suggests the full model. We propose a variable selection criterion for longitudinal data incorporating the GEE approach in a local misspecification framework, as in Hjort and Claeskens (2003).

3.1. Local misspecification framework

Consider the longitudinal data introduced in Section 2.1. We start with the full model where the covariates can be grouped into two categories: p certain covariates that will be included in the final model, and q uncertain ones that we are not sure about. The corresponding unknown coefficients are therefore composed of the certain coefficients $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ and the uncertain coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$, denoted as $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$. Any submodel S can be written as a special case of the full model: $\boldsymbol{\beta}_{\mathrm{S}} = (\boldsymbol{\theta}, \boldsymbol{\gamma}_{\mathrm{S}}, \mathbf{0}_{\mathrm{S}^c})$, where $\boldsymbol{\gamma}_{\mathrm{S}}$ is a $q_{\mathrm{S}} \times 1$ subvector of $\boldsymbol{\gamma}$ and $\mathbf{0}_{\mathrm{S}^c}$ is a $q_{\mathrm{S}^c} \times 1$ subvector of $q \times 1$ vector $\mathbf{0}$ with $\mathrm{S} \subset \{1, \ldots, q\}$. Let $\mathcal{N}$ denote the *empty* set. When $\mathrm{S} = \mathcal{N}$, the narrow model, $\boldsymbol{\beta}_{\mathcal{N}} = (\boldsymbol{\theta}, \mathbf{0})$, includes only the certain covariates. The true model is defined under the local misspecification framework as in Hjort and Claeskens (2003):

$$\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) = \big(\boldsymbol{\theta}_0, \frac{\boldsymbol{\delta}}{\sqrt{n}}\big).$$

Here $\boldsymbol{\gamma}_0 = 0$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)^\top$ measures how far the true model is from the narrow model in directions $1, \ldots, q$ of order $O(1/\sqrt{n})$; some δ_i's can be 0. Under this scenario, the size of the squared model biases and the model variances can reach $O(1/n)$, the highest possible large sample approximation.

In data analysis, we need to choose a narrow model, which should include highly significant covariates and other covariates of interest. This could be done on theoretical grounds or based on a pre-fit of the full model, with interests and experience included, theoretical and numerical evidences suggest that the particular choice of the narrow model only slightly influences the results.

To simplify the discussion in the context of the GEE approach, we ignore the treatment of the nuisance parameters $\boldsymbol{\alpha}$ and ϕ and assume the consistency of $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)$ and $\widehat{\phi}(\boldsymbol{\beta})$, and the boundedness of $\partial \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)/\partial \phi$ as in Liang and Zeger (1986). Thus, the quasi-score of the full model, evaluated at $(\boldsymbol{\theta}_0, \mathbf{0})$, can be written as

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix} = \begin{bmatrix} \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\gamma}; \mathcal{D})}{\partial \boldsymbol{\theta}} \\ \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\gamma}; \mathcal{D})}{\partial \boldsymbol{\gamma}} \end{bmatrix}_{\boldsymbol{\theta} = \boldsymbol{\theta}_0, \boldsymbol{\gamma} = \mathbf{0}}.$$

The corresponding $(p+q) \times (p+q)$ quasi-likelihood information matrix is denoted by

$$\boldsymbol{\Sigma} = \mathrm{Var}_N(\mathbf{U}) = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01} \\ \boldsymbol{\Sigma}_{10} & \boldsymbol{\Sigma}_{11} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{00} & \boldsymbol{\Sigma}^{01} \\ \boldsymbol{\Sigma}^{10} & \boldsymbol{\Sigma}^{11} \end{bmatrix},$$

where $\boldsymbol{\Sigma}^{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{00}^{-1}\boldsymbol{\Sigma}_{01})^{-1}$. Let $\boldsymbol{\pi}_{\mathrm{S}}$ be the $q_{\mathrm{S}} \times q$ projection matrix mapping $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}_{\mathrm{S}}$ with q_{S} the size of S, $\boldsymbol{\pi}_{\mathrm{S}}\boldsymbol{\gamma} = \boldsymbol{\gamma}_{\mathrm{S}}$; here $\boldsymbol{\pi}_{\mathrm{S}} = (\boldsymbol{I}_{q_{\mathrm{S}}} : \mathbf{0})$ or a column

permutation thereof. The quasi-score of the submodel S, evaluated at $(\boldsymbol{\theta}_0, \mathbf{0})$, can be written as

$$\mathbf{U}_{\mathrm{S}} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_{2,\mathrm{S}} \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 \\ \boldsymbol{\pi}_{\mathrm{S}} \mathbf{U}_2 \end{bmatrix}.$$

The corresponding quasi-likelihood information matrix has a $(p+q_{\mathrm{S}}) \times (p+q_{\mathrm{S}})$ dimension

$$\boldsymbol{\Sigma}_{\mathrm{S}} = \begin{bmatrix} \boldsymbol{\Sigma}_{00} & \boldsymbol{\Sigma}_{01}\boldsymbol{\pi}_{\mathrm{S}}^{\top} \\ \boldsymbol{\pi}_{\mathrm{S}}\boldsymbol{\Sigma}_{10} & \pi_{\mathrm{S}}\boldsymbol{\Sigma}_{11}\boldsymbol{\pi}_{\mathrm{S}}^{\top} \end{bmatrix} \quad \text{and} \quad \left(\boldsymbol{\Sigma}_{\mathrm{S}}^{11}\right)^{-1} = \boldsymbol{\pi}_{\mathrm{S}}\left(\boldsymbol{\Sigma}^{11}\right)^{-1}\boldsymbol{\pi}_{\mathrm{S}}^{\top}.$$

3.2. Quasi-likelihood-based ΔAIC

Let $(\widehat{\boldsymbol{\theta}}_{\mathrm{S}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{S}})$ be the GEE estimates under the submodel S. The AIC value of the submodel S is

$$-2\sum_{i=1}^{n} \log f(\mathbf{y}_i, \widehat{\boldsymbol{\theta}}_{\mathrm{S}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{S}}) + 2|\mathrm{S}|,$$

where $|\mathrm{S}|$ is the number of parameters in the submodel S. Similarly, the quasi-likelihood-based AIC value of the submodel S can be calculated as

$$\mathrm{QAIC}_{n,\mathrm{S}} = -2\sum_{i=1}^{n} Q(\widehat{\boldsymbol{\theta}}_{\mathrm{S}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{S}}; \mathbf{y}_i) + 2|\mathrm{S}|.$$

Here, QAIC is generally difficult to implement, due to the complex correlation structure of longitudinal data, especially the part with integration involving the inverse of working covariance matrix in the quasi-likelihood component. Nevertheless, every submodel includes the certain parameter $\boldsymbol{\theta}$, of which the narrow model is composed. By subtracting QAIC value of the narrow model from QAIC value of every submodel, we can avoid calculating log quasi-likelihood directly. We propose the AIC-type quasi-likelihood-based model selection criterion for longitudinal data incorporating the GEE approach as

$$\Delta\mathrm{AIC}_{n,\mathrm{S}} = \mathrm{QAIC}_{n,\mathrm{S}} - \mathrm{QAIC}_{n,\mathcal{N}}.$$

Result gives the specific form and the large sample behavior of $\Delta\mathrm{AIC}_{n,\mathrm{S}}$, with $\stackrel{d}{=}$ denoting equality in distribution and $\stackrel{d}{\to}$ denoting convergence in distribution.

Theorem 1. *Under the Assumptions given in the Appendix, as n goes to infinity,*

$$\begin{aligned} \Delta AIC_{n,s} &\stackrel{d}{=} -n\widehat{\boldsymbol{\gamma}}^{\top}\left(\boldsymbol{\Sigma}^{11}\right)^{-1}\boldsymbol{\pi}_s^{\top}\boldsymbol{\Sigma}_s^{11}\boldsymbol{\pi}_s\left(\boldsymbol{\Sigma}^{11}\right)^{-1}\widehat{\boldsymbol{\gamma}} + 2\left|\frac{S}{\mathcal{N}}\right| \\ &\stackrel{d}{\to} -\chi^2_{|S/\mathcal{N}|}(\boldsymbol{\lambda}_s) + 2\left|\frac{S}{\mathcal{N}}\right| \end{aligned}$$

with the non-centrality parameter $\boldsymbol{\lambda}_s = n\boldsymbol{\gamma}_o^{\top}\left(\boldsymbol{\Sigma}^{11}\right)^{-1}\boldsymbol{\pi}_s^{\top}\boldsymbol{\Sigma}_s^{11}\boldsymbol{\pi}_s\left(\boldsymbol{\Sigma}^{11}\right)^{-1}\boldsymbol{\gamma}_o$. *The number of degrees of freedom,* $|S/\mathcal{N}|$, *is the number of covariates in the candidate model S not in the narrow model.*

In the large sample context, the behavior of $\Delta\text{AIC}_{n,\text{S}}$ is fully dictated by the full model's GEE estimator $\widehat{\boldsymbol{\gamma}}$, and the limiting behaviors of all $\Delta\text{AIC}_{n,\text{S}}$ in principle determine the limits of all model selection probabilities through

$$\text{P}(\Delta\text{AIC selects submodel S} \mid \widehat{\boldsymbol{\gamma}}) \to \text{P}(\Delta\text{AIC selects submodel S} \mid \boldsymbol{\gamma}_0).$$

As shown in the proof of Theorem 1, in the Appendix, by subtracting, the complex component in QAIC is canceled out and the remaining terms involve only the uncertain parameters and the quasi-likelihood information matrix, both of which can be consistently estimated by the GEE approach. In particular, the estimators of $\boldsymbol{\Sigma}^{11}$ and $\boldsymbol{\Sigma}_{\text{S}}^{11} = \{\boldsymbol{\pi}_{\text{S}}(\boldsymbol{\Sigma}^{11})^{-1}\boldsymbol{\pi}_{\text{S}}^{\top}\}^{-1}$ can be obtained based on the sandwich estimator $\widehat{\boldsymbol{\Sigma}}_{\text{gee}}$. Similar to the traditional AIC, the model with the *smallest* ΔAIC value is selected as the final model.

Remark 1. There is no likelihood ratio tests available incorporating GEE approach for hypothesis testing Lipsitz and Fitzmaurice (2009, p.55). Nevertheless, the availability of quasi-likelihood suggests quasi-likelihood ratio tests. Consider the hypotheses:

$$\text{H}_0 : \boldsymbol{\gamma} = \mathbf{0} \quad \text{vs} \quad \text{H}_a : \boldsymbol{\gamma} \neq \mathbf{0}.$$

The null model can be viewed as a narrow model with only the certain parameter vector $\boldsymbol{\theta}$, while the alternative model is the full model. The quasi-likelihood ratio test statistic between the alternative and null models can be written as

$$\begin{aligned}
\text{QLR}_n &= 2\big[Q(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}; \mathcal{D}) - Q(\widehat{\boldsymbol{\theta}}_{\mathcal{N}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{N}}; \mathcal{D})\big] \\
&= -\text{QAIC}_{n,\text{F}} + 2|\text{F}| + \text{QAIC}_{n,\mathcal{N}} - 2|\mathcal{N}| \\
&= -\Delta\text{AIC}_{n,\text{F}} + 2\Big|\frac{\text{F}}{\mathcal{N}}\Big| \\
&\stackrel{d}{=} n\widehat{\boldsymbol{\gamma}}^{\top}(\boldsymbol{\Sigma}^{11})^{-1}\widehat{\boldsymbol{\gamma}}.
\end{aligned}$$

This shares the same form of quadratic type as the Wald test statistic.

4. Focused Information Criterion Incorporating GEE Approach

The criterion ΔAIC selects a model as an overall good fit on the basis of the observation data only; it may not necessarily be a good choice for estimating a special parameter. Hjort and Claeskens (2003) proposed a focused information criterion to minimize the limiting risk of the focus parameter's estimator. Here we consider a similar idea. Assume a focus parameter can be written as the function of the model parameters, denoted by $\boldsymbol{\zeta} = \boldsymbol{\zeta}(\boldsymbol{\theta}, \boldsymbol{\gamma})$, which has continuous partial derivatives in the neighborhood of $(\boldsymbol{\theta}_0, \mathbf{0})$. Denote the corresponding GEE estimator under the submodel S by $\widehat{\boldsymbol{\zeta}}_{\text{S}}$, and let

$$\boldsymbol{\omega} = \boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{00}^{-1}\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\theta}} - \frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\gamma}}, \boldsymbol{\tau}_0^2 = \left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\theta}}\right)^{\top}\boldsymbol{\Sigma}_{00}^{-1}\left(\frac{\partial\boldsymbol{\zeta}}{\partial\boldsymbol{\theta}}\right) \text{ and } \text{D}_{\text{S}} = \boldsymbol{\pi}_{\text{S}}^{\top}\boldsymbol{\Sigma}_{\text{S}}^{11}\boldsymbol{\pi}_{\text{S}}(\boldsymbol{\Sigma}^{11})^{-1},$$

where the derivatives are evaluated at $(\boldsymbol{\theta}_0, \mathbf{0})$.

Theorem 2. *Under the assumptions given in the Appendix, as n goes to infinity,*

$$\sqrt{n}(\widehat{\boldsymbol{\zeta}}_s - \boldsymbol{\zeta}_o) \xrightarrow{d} \boldsymbol{\Omega}_s = \boldsymbol{\Omega}_o + \boldsymbol{\omega}^\top \boldsymbol{\delta} - \boldsymbol{\omega}^\top D_s\, \boldsymbol{\Delta},$$

where $\boldsymbol{\zeta}_o = \boldsymbol{\zeta}(\boldsymbol{\theta}_o, \boldsymbol{\gamma}_o)$. $\begin{bmatrix}\mathbf{M}_1\\ \mathbf{M}_2\end{bmatrix} \sim N_{p+q}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Omega}_o = \left(\frac{\partial \boldsymbol{\zeta}}{\partial \boldsymbol{\theta}}\right)^\top \boldsymbol{\Sigma}_{00}^{-1}\mathbf{M}_1 \sim N_p(\mathbf{0}, \boldsymbol{\tau}_o^2)$, *and* $\boldsymbol{\Delta} = \boldsymbol{\delta} + \boldsymbol{\Sigma}^{11}(\mathbf{M}_2 - \boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{00}^{-1}\mathbf{M}_1) \sim N_q\big(\boldsymbol{\delta}, \boldsymbol{\Sigma}^{11}\big)$.

The limiting variable $\boldsymbol{\Omega}_s$ *is normal with mean* $\boldsymbol{\omega}^\top(I_q - D_s)\,\boldsymbol{\delta}$ *and variance* $\boldsymbol{\tau}_o^2 + \boldsymbol{\omega}^\top \boldsymbol{\pi}_s^\top \boldsymbol{\Sigma}_s^{11} \boldsymbol{\pi}_s \boldsymbol{\omega}$.

The limiting mean square errors therefore can be achieved by Theorem 2 as

$$\mathrm{mse}(\boldsymbol{\Omega}_{\mathrm{S}}) = \boldsymbol{\tau}_{\mathrm{o}}^2 + \boldsymbol{\omega}^\top \boldsymbol{\pi}_{\mathrm{S}}^\top \boldsymbol{\Sigma}_{\mathrm{S}}^{11} \boldsymbol{\pi}_{\mathrm{S}} \boldsymbol{\omega} + \left[\boldsymbol{\omega}^\top\left(\mathrm{I}_{\mathrm{q}} - \mathrm{D}_{\mathrm{S}}\right)\boldsymbol{\delta}\right]^2,$$

where the parameters $\boldsymbol{\tau}_0$, $\boldsymbol{\omega}$, $\boldsymbol{\Sigma}_{\mathrm{S}}^{11}$, D_{S} and $\boldsymbol{\delta}$ can be estimated incorporating the GEE approach under the full model. We propose the quasi-likelihood-based focused information criterion for longitudinal data incorporating the GEE approach as

$$\mathrm{QFIC}_{n,\mathrm{S}} = 2\widehat{\boldsymbol{\omega}}^\top \boldsymbol{\pi}_{\mathrm{S}}^\top \widehat{\boldsymbol{\Sigma}}_{\mathrm{S}}^{11} \boldsymbol{\pi}_{\mathrm{S}} \widehat{\boldsymbol{\omega}} + n\big[\widehat{\boldsymbol{\omega}}^\top(\mathrm{I}_{\mathrm{q}} - \widehat{\mathrm{D}}_{\mathrm{S}})\widehat{\boldsymbol{\gamma}}\big]^2. \tag{4.1}$$

In the large sample context, the behavior of QFIC is not only related to the uncertain parameter $\boldsymbol{\gamma}$, but is also influenced by $\boldsymbol{\omega}$, which is determined by the focus parameter $\boldsymbol{\zeta}$. Therefore QFIC chooses a different model depending on the different focus parameters. The one with the *smallest* QFIC value, therefore the smallest estimated mean square error of the focus parameter's estimator, is selected.

5. Frequentist Model Averaging Incorporating GEE Approach

5.1. Frequentist model averaging

Our model selection procedure aims to select a single final model that has the properties we need, either catching the overall information from the data with ΔAIC, or minimizing the mean square error for the focus parameter's estimator with QFIC. The inference based on the final model, however, ignores the uncertainty introduced by the selection procedure and results in a too-optimistic confidence interval. Frequentist model averaging, introduced in Claeskens and Hjort (2003), is an alternative to model selection that can address this problem and provide relatively robust statistical inference.

The quasi-likelihood-based model averaging (QFMA) estimator of the focus parameter $\boldsymbol{\zeta}$ is defined as the weighted average among the estimators based on all the candidate models incorporating the GEE approach,

$$\widehat{\boldsymbol{\zeta}}(\widehat{\boldsymbol{\gamma}}) = \sum_{\mathrm{S}} c(\mathrm{S}|\widehat{\boldsymbol{\gamma}})\widehat{\boldsymbol{\zeta}}_{\mathrm{S}},$$

where $c(\cdot|\cdot)$ is a weight function satisfying $\sum_{\mathrm{S}} c(\mathrm{S}|\widehat{\boldsymbol{\gamma}}) = 1$ with each term in $[0, 1]$.

Theorem 3. *Under the assumptions given in the Appendix, as n goes to infinity,*

$$\sqrt{n}(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_o) \xrightarrow{d} \boldsymbol{\Omega} = \boldsymbol{\Omega}_o + \boldsymbol{\omega}^\top \boldsymbol{\delta} - \boldsymbol{\omega}^\top \widehat{\boldsymbol{\delta}}(\boldsymbol{\Delta}),$$

where $\widehat{\boldsymbol{\delta}}(\boldsymbol{\Delta}) = \sum_s c(S|\boldsymbol{\Delta}) D_s \boldsymbol{\Delta}$. *The mean and variance of the limiting variable* $\boldsymbol{\Omega}$ *are* $E(\boldsymbol{\Omega}) = \boldsymbol{\omega}^\top \boldsymbol{\delta} - \boldsymbol{\omega}^\top E\big[\widehat{\boldsymbol{\delta}}(\boldsymbol{\Delta})\big]$ *and* $\mathrm{Var}\,(\boldsymbol{\Omega}) = \boldsymbol{\tau}_o^2 + \boldsymbol{\omega}^\top var\big[\widehat{\boldsymbol{\delta}}(\boldsymbol{\Delta})\big]\boldsymbol{\omega}$, *respectively.*

Motivated by Theorem 3 and Hjort and Claeskens (2003), we modify the traditional confidence interval for the focus parameter $\boldsymbol{\zeta}$ based on the model averaging estimator $\widehat{\boldsymbol{\zeta}}$ as

$$\mathrm{low}_n = \widehat{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\omega}}^\top \big[\widehat{\boldsymbol{\gamma}} - \frac{1}{\sqrt{n}}\widehat{\boldsymbol{\delta}}(\widehat{\boldsymbol{\gamma}})\big] - \frac{\boldsymbol{z}_k \widehat{\boldsymbol{\tau}}}{\sqrt{n}},$$
$$\mathrm{up}_n = \widehat{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\omega}}^\top \big[\widehat{\boldsymbol{\gamma}} - \frac{1}{\sqrt{n}}\widehat{\boldsymbol{\delta}}(\widehat{\boldsymbol{\gamma}})\big] + \frac{\boldsymbol{z}_k \widehat{\boldsymbol{\tau}}}{\sqrt{n}},$$

where $\boldsymbol{z}_k$ is the kth standard normal quantile and $\widehat{\boldsymbol{\tau}}/\sqrt{n}$ is the consistent estimator of the standard deviation for $\widehat{\boldsymbol{\zeta}}$ under the full model that can be written as $\boldsymbol{\tau}/\sqrt{n} = n^{-1/2}\big(\boldsymbol{\tau}_0^2 + \boldsymbol{\omega}^\top \boldsymbol{\Sigma}^{11}\boldsymbol{\omega}\big)^{1/2}$. By shifting the center of CI from $\widehat{\boldsymbol{\zeta}}$ by the amount $\widehat{\boldsymbol{\omega}}^\top \big[\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\delta}}(\widehat{\boldsymbol{\gamma}})/\sqrt{n}\big]$ and widening CI as $\boldsymbol{\tau}/\sqrt{n}$ instead of $\boldsymbol{\tau}_{\mathrm{S}}/\sqrt{n}$, thereby respecting the uncertainty of the selection process, the coverage probability is consistent with the nominal level.

Theorem 4. *Under the assumptions given in the Appendix, as n goes to infinity,*

$$Pr(low_n \le \boldsymbol{\zeta}_o \le up_n) \xrightarrow{d} 2\Phi(\boldsymbol{z}_k) - 1,$$

where $\Phi(\cdot)$ *is the standard normal distribution function.*

Theorem 4 can be easily proven by simultaneous convergence in distribution:

$$\big\{\sqrt{n}\big(\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0\big), \widehat{\boldsymbol{\gamma}}\big\} \xrightarrow{d} \big\{\boldsymbol{\Omega}_0 + \boldsymbol{\omega}^\top \boldsymbol{\delta} - \boldsymbol{\omega}^\top \widehat{\boldsymbol{\delta}}(\boldsymbol{\Delta}), \boldsymbol{\Delta}\big\}.$$

5.2. The choice of weight functions

The model averaging estimator can be connected to the model selection estimators by taking specific weight functions. Thus the submodel $\mathrm{S}_{\Delta\mathrm{AIC}}$, selected by ΔAIC, corresponds to an indicator function as its weight function, the hard core weight function,

$$\widehat{\boldsymbol{\zeta}}_{\Delta\mathrm{AIC}} = \sum_{\mathrm{S}} \mathbf{I}(\mathrm{S} = \mathrm{S}_{\Delta\mathrm{AIC}})\widehat{\boldsymbol{\zeta}}_{\mathrm{S}} = \widehat{\boldsymbol{\zeta}}_{\mathrm{S}_{\Delta\mathrm{AIC}}}.$$

Likewise, for the model selected by QIC, $\widehat{\boldsymbol{\zeta}}_{\mathrm{QIC}} = \sum_{\mathrm{S}} \mathrm{I}(\mathrm{S} = \mathrm{S}_{\mathrm{QIC}})\widehat{\boldsymbol{\zeta}}_{\mathrm{S}} = \widehat{\boldsymbol{\zeta}}_{\mathrm{S}_{\mathrm{QFIC}}}$, and for the model selected by QFIC, $\widehat{\boldsymbol{\zeta}}_{\mathrm{QFIC}} = \sum_{\mathrm{S}} \mathrm{I}(\mathrm{S} = \mathrm{S}_{\mathrm{QFIC}})\widehat{\boldsymbol{\zeta}}_{\mathrm{S}} = \widehat{\boldsymbol{\zeta}}_{\mathrm{S}_{\mathrm{QFIC}}}$. Buckland, Burnham and Augustin (1997) suggest that the choice of the weights in model

averaging estimators should be proportional to $\exp(f_{\rm S} - |{\rm S}|)$, where $f_{\rm S}$ is the maximized log-likelihood under the submodel S. Thus weights are proportional to $\exp(Q_{\rm S} - |{\rm S}|)$ with $Q_{\rm S}$ the quasi-likelihood of the submodel S for longitudinal data incorporating the GEE approach and $|{\rm S}|$ the number of parameters in the submodel S. A direct calculation (Buckland, Burnham and Augustin (1997)) indicates that the corresponding smoothed ΔAIC and QIC weights can be represented as

$$\frac{\exp\big(-\Delta{\rm AIC}_{n,{\rm S}}/2\big)}{\sum_{\rm T}\exp\big(-\Delta{\rm AIC}_{n,{\rm T}}/2\big)} \quad \text{and} \quad \frac{\exp\big(-{\rm QIC}_{n,{\rm S}}/2\big)}{\sum_{\rm T}\exp\big(-{\rm QIC}_{n,{\rm T}}/2\big)}.$$

It can also be beneficial to consider the information carried by QFIC by using the smoothed QFIC weights, similar to that suggested in Claeskens and Hjort (2003) and Claeskens and Hjort (2008):

$$\frac{\exp\Big(-(\kappa/2)[({\rm QFIC}_{n,{\rm S}})/(\widehat{\boldsymbol{\omega}}^{\top}\widehat{\boldsymbol{\Sigma}}^{11}\widehat{\boldsymbol{\omega}})]\Big)}{\sum_{\rm T}\exp\Big(-(\kappa/2)[({\rm QFIC}_{n,{\rm T}})/(\widehat{\boldsymbol{\omega}}^{\top}\widehat{\boldsymbol{\Sigma}}^{11}\widehat{\boldsymbol{\omega}})]\Big)}, \qquad \text{for some } \kappa \geq 0, \tag{5.1}$$

with κ bridging the weights from the uniform (κ close to 0) to the hard-core (large κ).

6. Simulation Studies

We investigated the performance of the proposed quasi-likelihood-based ΔAIC, QFIC, and QFMA for longitudinal data incorporating the GEE approach. We compared ΔAIC and the traditional QIC in terms of frequency in selecting the true model. The post-model selection procedures using QFIC, ΔAIC, and QIC, denoted as P-QFIC, P-ΔAIC, and P-QIC, were compared with their smoothed versions, denoted as S-QFIC, S-ΔAIC, and S-QIC, in terms of the coverage probabilities (CPs) of the estimated 95% confidence intervals and the estimated mean square errors (MSEs) for the focus parameters. As a reference, the inference based on the full model (Full) is reported as well.

We considered discrete and continuous responses with $n = 50$ and/or $n = 100$ subjects where each had $m = 3$ visits.

Example 1. To compare the performance of ΔAIC and QIC in selecting the true model, we used the same model setting as in Pan (2001a). We did not include QFIC in this as it depends on the focus parameter of interest.

We took binary response with $E(y_{ij}|x_{1,ij}, x_{2,ij}, x_{3,ij}, x_{4,ij}) = \mu_{ij}$ and ${\rm logit}(\mu_{ij}) = \beta_0+\beta_1 x_{1,ij}+\beta_2 x_{2,ij}+\beta_3 x_{3,ij}+\beta_4 x_{4,ij}$, where $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,m\}$ and covariates generated from $x_{1,ij} \overset{i.i.d.}{\sim}$ Bernoulli(1/2), $x_{2,ij} = j-1$ and $x_{3,ij}, x_{4,ij} \overset{i.i.d.}{\sim}$ Uniform$(-1,1)$, where $x_{3,ij}$ and $x_{4,ij}$ were also independent of $x_{1,ij}$. The coefficients were $\beta_0 = 0.25 = -\beta_1 = -\beta_2$ and $\beta_3 = \beta_4 = 0$. The model with intercept,

Table 1. Simulation studies — candidate models.

Model	Covariates	Model	Covariates
M1 - Full	Intercept, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$	M5	Intercept, $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$
M2	Intercept, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$	M6	Intercept, $\mathbf{x}_1$, $\mathbf{x}_3$
M3	Intercept, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$	M5	Intercept, $\mathbf{x}_1$, $\mathbf{x}_4$
M4 - True	Intercept, $\mathbf{x}_1$, $\mathbf{x}_2$	M8 - Narrow	Intercept, $\mathbf{x}_1$

$\mathbf{x}_1$ and $\mathbf{x}_2$ was the true model and the narrow model included the intercept and $\mathbf{x}_1$ only. The final model was selected from the candidate models listed in Table 1.

We used the Copulas package, developed by Yan (2007), to generate two types of correlation structure within each response: exchangeable and autoregressive with a correlation coefficient $\rho = 0.5$, denoted by EX(0.5) and AR(0.5). Under these two scenarios, based on 1,000 simulation replicates, the frequencies of the candidate models selected by ΔAIC and QIC incorporating the GEE approach, with correlation structures IN, EX and AR, are listed in Table 2.

Generally, Table 2 shows better performance of ΔAIC than QIC in terms of higher frequencies of selecting the true model. In particular, under the correct working correlation EX for the EX(0.5) scenario and AR for the AR(0.5) scenario, ΔAIC works observably better than QIC. Under the independent working correlation, QIC turns out to be comparable with ΔAIC. These patterns also show the bias of QIC, introduced by simplifying with working independence model and ignoring the complex part in the deriving process. The narrow model contains only the intercept and $\mathbf{x}_1$ with $(\beta_0, \beta_1) = (2, 1)$, while the coefficients of the two nonzero covariates are $(\beta_2, \beta_3) = (2, -2)/\sqrt{mn}$. With $m = 3$ and $n = 50$ or 100, β_2 is about 0.16 or 0.12, indicating that the signals of x_2 and x_3 are weaker, and resulting in the narrow model being selected with a high proportion.

The first simulation study assumed a simple correlation among the observations. While in many longitudinal studies, it is impossible to know its structure. We generated longitudinal data, 30% of which was from EX(0.5), 30% from AR(0.5), and the rest of which had the correlation structure,

$$\mathbf{R}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & 0.4 & 0.1 \\ 0.4 & 1 & 0.7 \\ 0.1 & 0.7 & 1 \end{bmatrix}.$$

Again, ΔAIC and QIC were applied to this scenario and the results of 1,000 simulation replicates are shown in Table 3.

Table 3 with $n = 50$ shows that QIC works better under IN, while ΔAIC works better under EX and AR. For n=100, ΔAIC works better although it is close to QIC under IN.

Table 2. Simulation I - frequency of model selected by ΔAIC and QIC.

			True Correlation EX(0.5)							
n	Criteria	R	Full	M2	M3	True	M5	M6	M7	Narrow
50	ΔAIC	IN	19	80	81	375	10	72	75	288
		EX	17	86	80	371	11	85	77	273
		AR	23	88	78	367	11	75	81	277
	QIC	IN	20	77	80	364	10	73	74	302
		EX	28	83	91	343	13	80	80	282
		AR	20	81	88	354	14	75	78	290
100	ΔAIC	IN	21	101	108	542	7	29	31	161
		EX	17	107	105	544	8	27	25	167
		AR	15	105	110	540	8	34	22	166
	QIC	IN	20	102	111	541	6	31	31	158
		EX	24	117	119	515	9	32	28	156
		AR	19	107	113	535	9	33	27	157
			True Correlation AR(0.5)							
			Full	M2	M3	True	M5	M6	M7	Narrow
50	ΔAIC	IN	17	68	66	335	14	69	67	364
		EX	17	80	70	322	17	77	78	339
		AR	15	83	64	323	18	80	84	333
	QIC	IN	19	66	69	333	14	75	68	356
		EX	23	76	70	322	16	74	76	343
		AR	20	76	73	315	20	80	78	338
100	ΔAIC	IN	17	100	113	473	12	38	35	212
		EX	14	101	107	480	7	48	37	206
		AR	20	87	113	486	12	50	35	197
	QIC	IN	16	98	115	475	11	41	35	209
		EX	20	109	123	452	13	44	35	204
		AR	20	101	121	462	14	43	33	206

Example 2. We compared model selection and model averaging approaches with continuous and binary responses. Here the continuous response variables are $\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{x}_{1,i} + \beta_2 \mathbf{x}_{2,i} + \beta_3 \mathbf{x}_{3,i} + \boldsymbol{\epsilon}_i$, with $i = 1, \ldots, n$. The covariates $\mathbf{x}_{1,i} = (x_{1,i1}, x_{1,i2}, x_{1,i3})^\top$, $\mathbf{x}_{2,i} = (x_{2,i1}, x_{2,i2}, x_{2,i3})^\top$ and $\mathbf{x}_{3,i} = (x_{3,i1}, x_{3,i2}, x_{3,i3})^\top$ were independently generated from a multivariate normal distribution with mean $(1, 1, 1)^\top$ and identity covariance matrix. The error term $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3})^\top$ was generated independently of the covariates from a three-dimensional normal distribution with mean $\mathbf{0}$, marginal variance $\mathbf{1}$, and three correlation matrices: EX(0.5), AR(0.5), and MIX as in Example 1. The narrow model here contains only the intercept and $\mathbf{x}_1$ with $(\beta_0, \beta_1) = (2, 1)$, while the coefficients of the other covariates were $(\beta_2, \beta_3) = (2, -2)/\sqrt{mn}$. Submodels are given in Table 4.

A focus parameter, $\zeta = -2\beta_0 + 2\beta_1 - 0.5\beta_2 + 0.5\beta_3$, was considered in this study. The models were fitted incorporating the GEE approach. The simulation

Table 3. Simulation II — frequency of model selected by ΔAIC and QIC - mixed true correlation.

n	Criteria	R	Full	M2	M3	True	M5	M6	M7	Narrow
50	ΔAIC	IN	20	73	67	308	14	81	75	362
		EX	13	71	70	316	16	87	61	366
		AR	19	70	69	313	11	89	70	359
	QIC	IN	21	68	68	317	19	87	68	352
		EX	26	76	76	303	24	92	68	335
		AR	23	76	72	305	24	89	72	339
100	ΔAIC	IN	12	94	92	497	9	56	36	204
		EX	15	90	98	496	8	58	35	200
		AR	15	83	95	506	8	54	32	207
	QIC	IN	14	91	93	496	10	54	36	206
		EX	27	92	94	478	14	58	37	200
		AR	24	100	94	476	15	57	33	201

Table 4. Candidate models — when Y is continuous.

Model	Covariates	Model	Covariates
Full Model	Inte., $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$	Submodel 2	Inte., $\mathbf{x}_1$, $\mathbf{x}_3$
Submodel 1	Inte., $\mathbf{x}_1$, $\mathbf{x}_2$	Narrow Model	Inte., $\mathbf{x}_1$

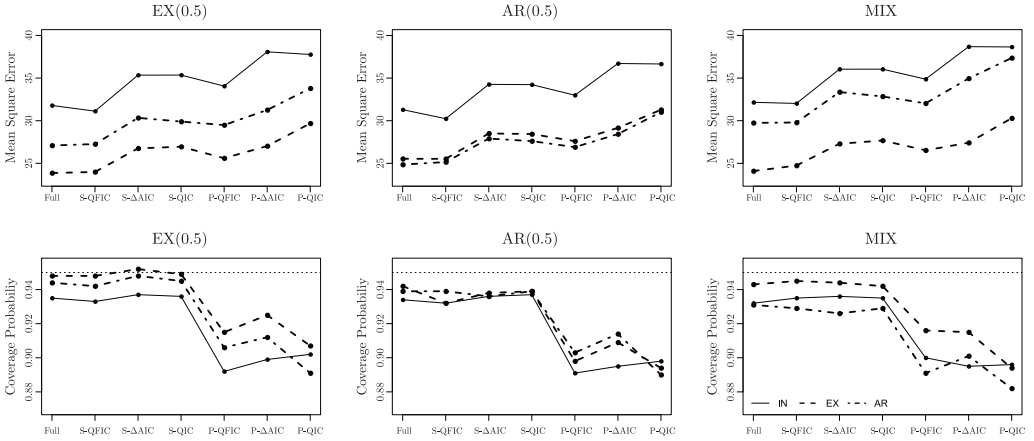

Figure 1. Simulation results (continuous response) for three scenarios: exchangeable, autoregressive and mixture with six selection criterion. Upper panel: MSE; lower panel: coverage probability. (Solid line: independent; broken line: exchangeable; dot-broken line: autoregressive.)

results, based on 1,000 replications, are presented in Figure 1.

With the influence of uncertain coefficients, regardless of working correlation

structures, MSE plots in the upper panel of Figure 1 consistently show the better performance, in term of relatively smaller MSE values, of model averaging, and also the better performance of the model selection criterion QFIC. The S-ΔAIC behavior was similar to that of S-QIC, while P-ΔAIC worked better than P-QIC. The full model provides unbiased estimates at the cost of increased variability, therefore relatively larger MSE. GEE estimates under the correct working correlation structures of EX and AR always have the smallest MSE values for all model selection or model averaging procedures. This is consistent with the true correlation's efficiency pointed out in Liang and Zeger (1986) under MIX, AR gives the smallest MSE value, which may be due to the relatively closer correlation structure of EX to the true correlation. In all three scenarios, IN results in the largest MSE.

The three CP plots in the bottom panel of Figure 1 indicate the better performance of modified CIs based on the model averaging procedure compared with traditional CIs.

In practice, the number of visits of each subject can vary, especially in clinical studies. The model selection and model averaging approaches we have proposed can be applied to these situation. We conducted a simulation study to illustrate their application to the situation of unequal number of visits.

We generated 100 samples, 30 samples with 4 visits, 30 samples with 2 visits, and 40 samples with 3 visits. The rest of the model was the same with continuous response variables $\mathbf{y}$ generated from the linear combination of $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ and a focus parameter, $\zeta = -2\beta_0 + 2\beta_1 - 0.5\beta_2 + 0.5\beta_3$. For this MIX case, 30 samples with 4 visits were generated with AR correlation, 30 samples with 2 visits were generated with EX correlation, and 40 samples with 4 visits with UN correlation. The result are shown in Figure 2. The trend in Figures 1 and 2 are very similar in terms of mean square error and coverage probability for the six selection criteria.

To implement ΔAIC, QFIC and QFMA, we need a narrow model containing "important" covariates. For possible misspecification assessment, we conducted a simulation study. We used the simulation parameters and scenarios of Example 2, except for intercept and x_2 in the narrow model instead of intercept and x_1. The focus parameter stays the same. The estimations of mean square error and coverage probability of the six approaches are shown in Figure 3.

There we can see that, even when misspecified, the trends do not change very much. So the simulation suggests that the choice of the narrow model only slightly influences the results.

Example 3. We considered binary responses generated as in Example 1 with coefficients combinations $(\beta_0, \beta_1) = (3, -3)$ and $(\beta_2, \beta_3, \beta_4) = (1, 1, -1)/\sqrt{mn}$. The narrow model therefore contains the intercept and $\mathbf{x}_1$, and the candidate models are listed in Table 5.

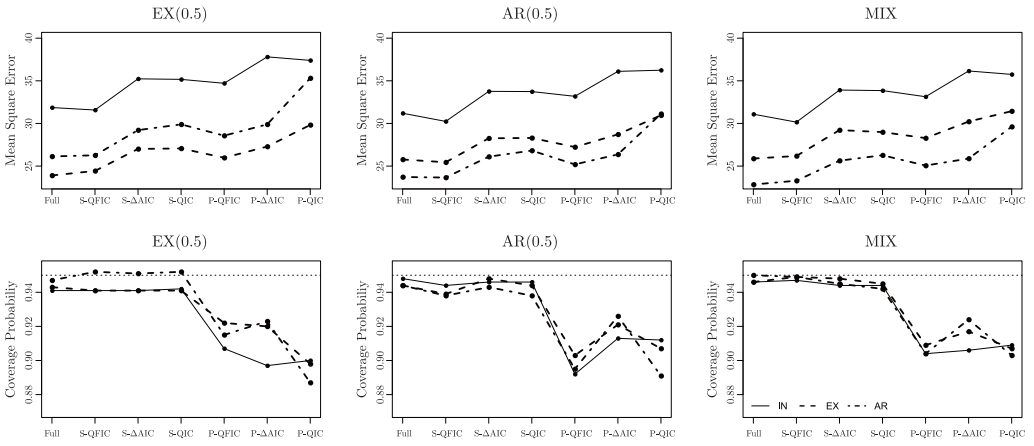

Figure 2. Simulation results (continuous response with different number of visits) for three scenarios: exchangeable, autoregressive and mixture with six selection criterion. Upper panel: MSE; lower panel: coverage probability.

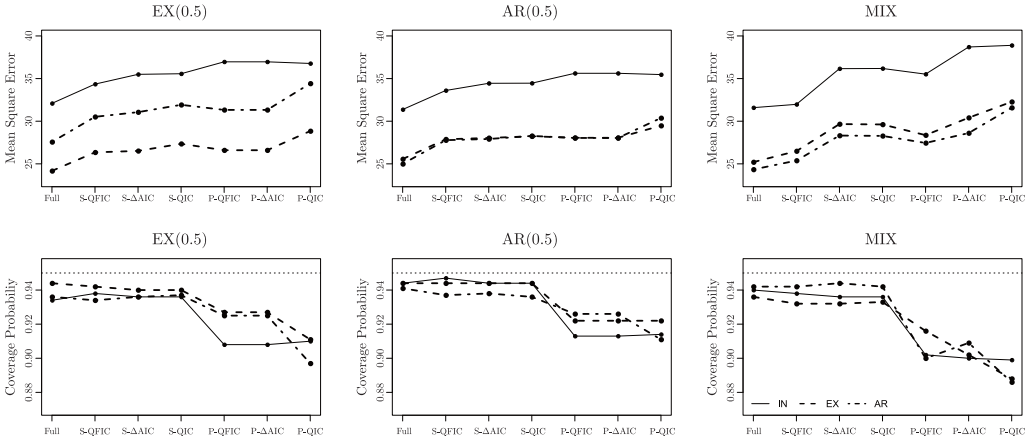

Figure 3. Simulation results (continuous response with misspecified narrow model) for three scenarios: exchangeable, autoregressive and mixture with six selection criterion. Upper panel: MSE; lower panel: coverage probability.

Table 5. Candidate models – when Y is binary.

	Covariates		Covariates		Covariates		Covariates
1	Inte, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$	3	Inte, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_4$	5	Inte, $\mathbf{x}_1$, $\mathbf{x}_3$, $\mathbf{x}_4$	7	Inte, $\mathbf{x}_1$, $\mathbf{x}_4$
2	Inte, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$	4	Inte, $\mathbf{x}_1$, $\mathbf{x}_2$	6	Inte, $\mathbf{x}_1$, $\mathbf{x}_3$	8	Inte., $\mathbf{x}_1$

We focused on $\zeta = 2\beta_1 + 2\beta_2 + 0.5\beta_3 + 0.5\beta_4 + 0.5\beta_5$. The simulation results, based on 1,000 replications, are presented in Figure 4. The patterns of binary longitudinal data there are similar to those in the continuous case.

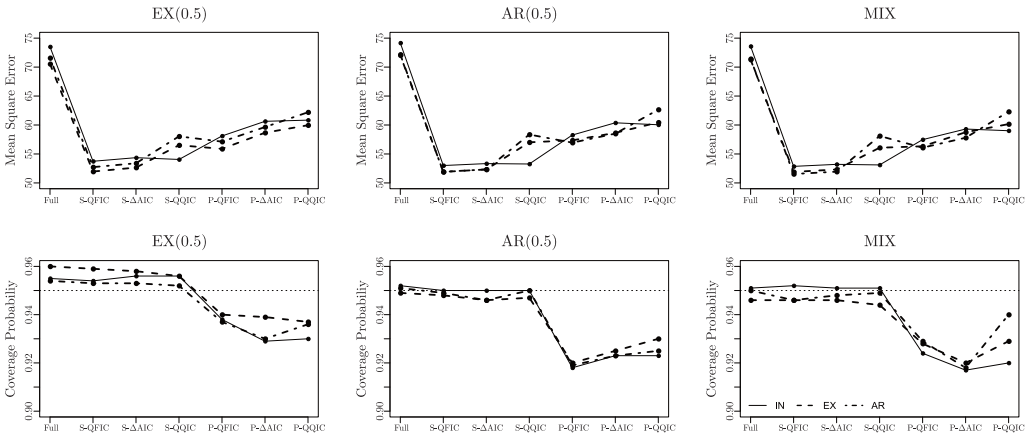

Figure 4. Simulation results (binary response) for three scenarios: exchangeable, autoregressive and mixture with six selection criterion. Upper panel: MSE; lower panel: coverage probability.

In our data example in Section 7, we have 12 candidate covariates and we identify 4 significant covariates. Instead of an exhaustive search, Claeskens, Croux and van Kerckhoven (2006) introduced the backward elimination approach that can significantly reduce the number of candidate models we need to consider. The detail of this approach are introduced in Section 7. Here we just mimic the situation of Section 7's data example.

We generated 100 samples with $m = 3$. The binary response data was generated through the linear combination of 15 candidate covariates, with coefficients $\beta = 2, -2, 2, -2, (1, -1, 0.5, -0.5, 0.1, -0.1)/\sqrt{(mn)}, 0, 0, 0, 0, 0$. We focus on the parameter estimation of $\zeta = -2 \times cd4 + 2 \times cd8 - 2 \times age + 2$. We only considered the using independent working correlation structure and the backward elimination approach The results are shown in Figure 5.

Similar trends were observed. Model averaging version of the estimates have smaller MSEs than the model selection version and QFIC have smaller MSE than ΔAIC and QIC. The modified CIs have closer to 95% coverage probability than did the traditional CIs. We found that our proposed methods could work well incorporating the backward elimination approach for large numbers of covariates.

In summary, for both binary and continuous longitudinal data studies, the MSE and CP plots consistently show the advantage of QFMA, compared with ΔAIC and QFIC.

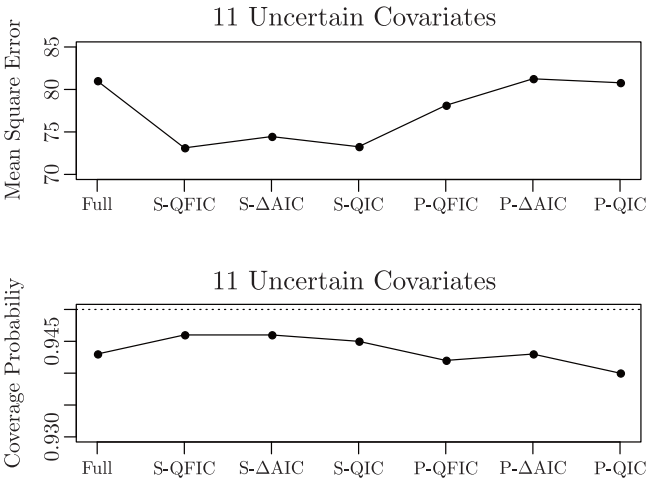

Figure 5. Simulation results (binary response with 12 candidate covariates) for three scenarios: exchangeable, autoregressive and mixture with six selection criterion. Upper panel: MSE; lower panel: coverage probability.

7. An Example

We applied our model selection criteria ΔAIC and QFIC, and the model averaging procedure QFMA incorporating the GEE approach, to the AIDS Clinical Trials Group protocol A5055 longitudinal study. A5055 was a Phase I/II, randomized, open-label, 24-week comparative study of the PK, tolerability, safety and antiretroviral effects of two regimens of indinavir (IDV), ritonavir (RTV), and two nucleoside analogue reverse transcriptase inhibitors on HIV-1-infected patients who failed protease inhibitor containing antiretroviral therapies.

In this study, 42 patients were randomized to one of two regimens and were visited at entry, weeks 1, 2 and 4 and every 4 weeks thereafter through week 24 of follow-up. Plasma for HIV-1 RNA testing was conducted at each visit, providing a binary response (rna: 0=negative and 1=positive). A series of potential explanatory variables were collected at the same time, including age, CD4 cell counts (cd4), CD8 cell counts (cd8), Phenotypic determination of antiretroviral drug resistance (ic50), the trough level of IDV and RTV concentration in plasma (icmin, rcmin), the IDV and RTV concentration in plasma measured after 12h from dose taken (ic12h, rc12h), the maximum IDV and RTV concentration in plasma (icmax, rcmax), the area under the plasma concentration-time curve for IDV and RTV (iauc, rauc) and pill counts for monitoring adherence (iadh, radh). A more detailed description and some analysis is in Wu et al. (2005), Huang, Liang and Wu (2008), and Acosta et al. (2004). We aim to identify pertinent covariates in order to better predict the antiretroviral treatment response for a new patient.

Table 6. Data example- Full model estimation under IN, EX and AR working correlations.

	IN		EX		AR	
Factor	Estimate	p-value	Estimate	p-value	Estimate	p-value
int.	3.95e+01	0.0124	3.92e+01	0.0144	3.70e+01	0.0226
cd4	1.04e-02	0.0011	1.06e-02	0.0011	1.06e-02	0.0013
cd8	-1.73e+00	0.0043	-1.75e+00	0.0037	-1.70e+00	0.0032
age	-8.07e-02	0.0064	-8.00e-02	0.0071	-7.77 e-02	0.0087
icmax	-2.30e+00	0.0384	-2.29e+00	0.0410	-2.23e+00	0.0373
iauc	9.19e-02	0.0796	9.12e-02	0.0851	8.51e-02	0.0917
ic50	2.13e-01	0.0908	2.08e-01	0.0975	1.89e-01	0.0919
rcmin	1.75e-01	0.1076	1.77e-01	0.1097	1.85e-01	0.1091
rc12h	7.88e-01	0.1278	8.09e-01	0.1223	7.99e-01	0.1392
ic12h	-6.34e-04	0.2284	-6.49e-04	0.2220	-6.13e-04	0.2423
rcmax	-0.71e+00	0.2632	-1.55e+00	0.2826	-1.36e+00	0.3715
iadh	-4.80e+00	0.2914	-4.68e+00	0.3037	-3.91e+00	0.3424
radh	2.24e+00	0.6473	2.07e+00	0.6736	1.25e+00	0.7865
icmin	7.55e-05	0.7900	8.33e-05	0.7704	12.70e-05	0.6463
rauc	-3.28e-03	0.8530	-3.92e-03	0.8277	-6.69e-03	0.7248

We first fit the full model by considering all 14 possible covariates with the full model

$$\begin{aligned}\text{logit}(\mu_{ij}) =&\beta_0 + \beta_1 \text{cd4}_{ij} + \beta_2 \text{cd8}_{ij} + \beta_3 \text{age}_{ij} + \beta_4 \text{ic50}_{ij} + \beta_5 \text{radh}_{ij} + \beta_6 \text{iadh}_{ij} \\ &+ \beta_{13} \text{rauc}_{ij} + \beta_{14} \text{iauc}_{ij} + \beta_7 \text{rcmin}_{ij} + \beta_8 \text{icmin}_{ij} + \beta_9 \text{rcmax}_{ij} \\ &+ \beta_{10} \text{icmax}_{ij} + \beta_{11} \text{rc12h}_{ij} + \beta_{12} \text{ic12h}_{ij},\end{aligned}$$

with $i = 1, \ldots, 42$, $j = 1, \ldots, t_i$ and μ_{ij} the conditional expectation of rna_{ij}. Due to the complicated correlation structure within each patient's serial observations, the marginal logistic regression model was fit incorporating the GEE approach under three different working correlation structures: IN, EX and AR. By the order of the covariates' significance, the results are listed in Table 6 in terms of the corresponding coefficients' estimates and p-values. In particular, under the EX working correlation structure, the estimate of nuisance parameter α was 0.01, while under the AR working correlation structure, the estimate of nuisance parameter α was 0.15.

From Table 6, IN and EX give similar coefficients' estimates and p-values, and quite different from those under AR, but all the results indicate the same highly significant covariates: int., cd4, cd8 and age. We took these as certain and we ran the selection and averaging procedures among the 11 uncertain ones. The backward elimination approach was used. It starts with the full model, deletes one covariate at each step based on certain model selection criterion, and

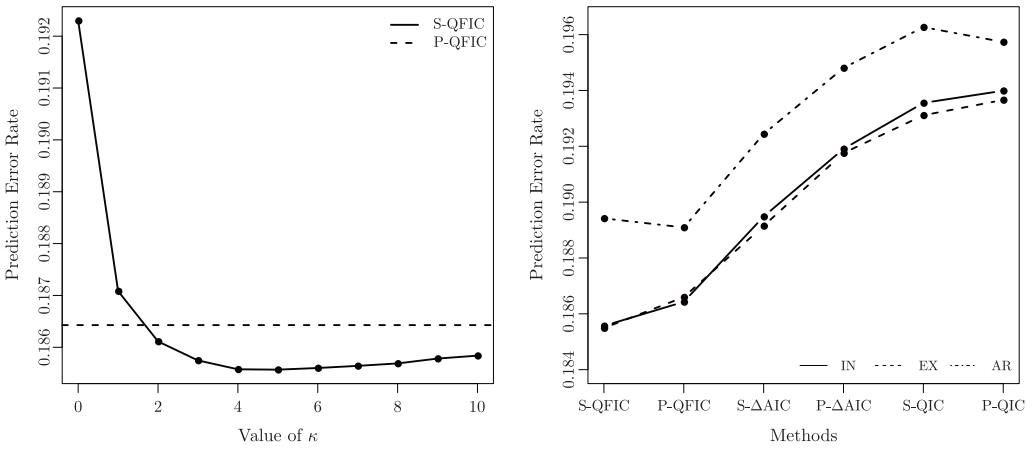

Figure 6. Prediction errors for the data example using the backward elimination approach.

ends up with 12 nested candidate models, among which the model selection and averaging procedures are processed.

We examined six model selection and averaging procedures' predictive powers by using cross-validation experiment: S-QFIC, S-ΔAIC, S-QIC, P-QFIC, P-ΔAIC, and P-QIC. With the correlation structures of patients observations, the leave-one-patient-out was chosen over of the leave-one-observation-out. The prediction error rates were evaluated by the percentage of wrong predictions among 1,000 replicates; these are plotted in Figure 6.

The parameter κ bridges the QFIC-based weights from uniform to hard-core. The left panel in Figure 6 gives the prediction error rates of S-QFIC using κ values ranging from 0 to 10. When $\kappa = 0$, the estimate is the arithmetic mean of 12 estimates from the candidate models and results in the largest error rate. The dashed line gives the error rate obtained by using QFIC selection procedure, equivalent to S-QFIC that assigns weight 1 to the best model in the sense of QFIC and 0 to other models. Here the prediction error rate of S-QFIC dramatically decreases as κ takes values from 0 to 1, is less than that of P-QFIC when $\kappa = 2$, reaches a minimum when κ is about 5, and converges to the error rate of P-QFIC when $\kappa \to \infty$.

The right panel in Figure 6 plots the prediction error rates based on S-QFIC with $\kappa = 5$, S-ΔAIC, S-QIC, P-QFIC, P-ΔAIC, and P-QIC incorporating the GEE approach with IN, EX and AR working correlation structures.

One could also use the forward selection approach. Starting from the null model, it adds the variable that yields the lowest value for the model selection criterion, to the currently "best" model, repeated until the full model is obtained. It also ends up with 12 nested candidate models. We examined six model selection

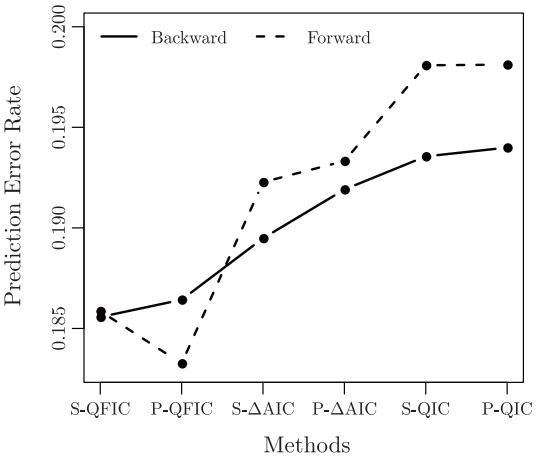

Figure 7. Prediction errors for the data example using forward selection approach.

and averaging procedures' predictive powers (S-QFIC, S-ΔAIC, S-QIC, P-QFIC, P-ΔAIC, and P-QIC) by using cross-validation. The prediction error rate is plotted in Figure 7, based on 1,000 replicates. Here, we took the weight parameter $\omega = 5$ that gave the smaller prediction error rate using the backward approach. We also plotted prediction error rates from the backward elimination approach with $\omega = 5$, using only the IN working correlation structure.

From the graph, the backward approach did provide a slightly smaller error rate than forward approach, which may due to the selection of $\omega = 5$. The general trend is similar regarding the six prediction error rates.

To demonstrate the final models selected by QFIC for different estimation interests, we chose as focus parameters the coefficients of three significant covariates cd4 , cd8 and age. The backward elimination selection was processed incorporating the GEE approach using the IN working correlation structure based on ΔAIC, QIC, and QFIC. The 12 nested candidate models are listed in Tables 7 and 8 along with the values of model selection criterion and the focus parameters' estimates.

Regardless of the focus parameters, ΔAIC and QIC resulted in the same best model among the same 12 nested models. The model selection criterion QFIC, selected three different final models among the 12 nested models based on the different focus parameters.

We need a narrow model to include highly significant covariates and other covariates of interest. However, theoretical and numerical evidence suggests that the choice of the narrow model only slightly influences the results.

8. Conclusion and Remarks

Table 7. Data example — 12 nested model selected by ΔAIC.

Factors	1	2	3	4	5	6	7	8	9	10	11	12
icmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆					
iauc	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	
ic50	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		
rcmin	⋆	⋆	⋆	⋆	⋆	⋆						
rc12h	⋆	⋆	⋆	⋆	⋆							
ic12h	⋆	⋆	⋆	⋆								
rcmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆				
iadh	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆			
radh	⋆	⋆	⋆									
icmin	⋆	⋆										
rauc	⋆											
ΔAIC(e-00)	-45.7	-47.7	-49.6	-51.3	-50.8	-51.1	-48.9	-40.9	-32.2	-18.0	-7.7	0.0
$\widehat{\boldsymbol{\beta}}_1$(e-03)	10.45	10.41	10.42	10.37	9.98	8.63	8.36	9.05	9.29	9.36	9.38	9.47
$\widehat{\boldsymbol{\beta}}_2$(e-00)	-1.73	-1.70	-1.66	-1.68	-1.62	-1.39	-1.32	-1.00	-1.12	-1.14	-0.86	-0.87
$\widehat{\boldsymbol{\beta}}_3$(e-02)	-8.07	-8.08	-8.16	-8.06	-9.27	-8.43	-9.07	-6.47	-5.39	-5.54	-5.57	-4.44

Based on the quasi-score function, we have proposed two variable selection criteria, one is the true model-oriented and the other is the parameter of interest-oriented, and a model averaging procedure for longitudinal data incorporating the GEE approach, and have derived asymptotic properties for the proposed procedures. Simulation studies and data analysis have shown their superiorities in terms of smaller mean square error and closer to 95% coverage probability and smaller prediction error rate.

The key point of the proposed ΔAIC is to consider the difference between the candidate model and a narrow model to avoid calculating the integration involved in the quasi-likelihood by executing the Taylor expansion. The resulting criterion can be easily implemented by fitting the full model with a penalty term. Although our criterion is built up under the AIC framework, we can also analogously define a BIC-type quasi-likelihood-based model selection criterion (referenced as ΔBIC) for longitudinal data incorporating the GEE approach by just changing the penalty term.

There are two issues regarding model selection for longitudinal data incorporating the GEE approach: variable selection and working correlation selection. However, currently ΔAIC and QFIC are limited to variable selection. More work needs to do for the selection of working correlation structures.

In the study of weight choice for the model averaging procedure, we have noted the effect of κ on the weights. When the performances among all candidate models are quite different, a large value of κ is preferable to stretch the weights' differences. When all candidate models behave alike, a small κ is chosen to shrink the weights' difference. More research is needed for the theoretical properties of κ.

Table 8. Data example - 12 nested models.

The selected one (shaded) by QFIC for CD4												
Factors	1	2	3	4	5	6	7	8	9	10	11	12
icmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆				
iauc	⋆	⋆	⋆	⋆	⋆							
ic50	⋆	⋆	⋆	⋆	⋆	⋆						
rcmin	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	
rc12h	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆			
ic12h	⋆	⋆	⋆	⋆								
rcmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		
iadh	⋆	⋆	⋆									
radh	⋆											
icmin	⋆	⋆	⋆	⋆	⋆	⋆	⋆					
rauc	⋆	⋆										
QFIC_1(e-04)	2.956	2.847	2.797	2.773	2.700	2.119	2.047	2.048	2.175	6.560	8.500	15.50
$\widehat{\boldsymbol{\beta}}_{cd4}$(e-03)	10.45	10.41	10.37	10.25	9.96	10.80	10.17	10.35	10.35	9.41	9.87	9.47

The selected one (shaded) by QFIC for CD8												
Factors	1	2	3	4	5	6	7	8	9	10	11	12
icmax	⋆	⋆	⋆	⋆	⋆							
iauc	⋆	⋆	⋆	⋆	⋆	⋆						
ic50	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	
rcmin	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆			
rc12h	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆				
ic12h	⋆	⋆	⋆	⋆								
rcmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		
iadh	⋆	⋆	⋆	⋆	⋆	⋆	⋆					
radh	⋆	⋆										
icmin	⋆	⋆	⋆									
rauc	⋆											
QFIC(e-00)	2.67	2.23	1.94	1.86	1.85	1.86	1.87	1.98	4.15	6.26	9.00	18.91
$\widehat{\boldsymbol{\beta}}_{cd8}$(e-00)	-1.73	-1.70	-1.73	-1.68	-1.62	-1.50	-1.51	-1.49	-1.16	-1.02	-1.12	-0.87

The selected one (shaded) by QFIC for age												
Factors	1	2	3	4	5	6	7	8	9	10	11	12
icmax	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆				
iauc	⋆	⋆	⋆	⋆	⋆							
ic50	⋆	⋆	⋆	⋆								
rcmin	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆			
rc12h	⋆	⋆	⋆	⋆	⋆	⋆	⋆					
ic12h	⋆	⋆										
rcmax	⋆	⋆	⋆									
iadh	⋆	⋆	⋆	⋆	⋆	⋆						
radh	⋆											
icmin	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	
rauc	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆	⋆		
QFIC_3(e-02)	2.78	2.45	2.32	1.78	1.71	1.52	1.39	0.92	1.42	2.12	5.30	2.25
$\widehat{\boldsymbol{\beta}}_{age}$(e-02)	-8.07	-7.97	-8.94	-7.78	-7.35	-6.28	-6.35	-5.67	-5.55	-6.27	-5.71	-4.44

In high-dimensional settings with a large number of uncertain parameters, averaging on all possible candidate models is practically infeasible. A backward or a forward selection procedure is preferable to reduce computational burden. A further investigation on such procedures is warranted.

Supplementary Material

The supplementary material presents assumptions and the proofs of Theorems 1−4.

Acknowledgements

The authors thank the Editors, a former Editor, an associate editor and two referees for their valuable suggestions and comments that have substantially improved an earlier version of this paper. Zou's research was partially supported by National Natural Science Foundation of China (Grant nos. 11471324 and 11331011) and a grant from the Beijing High-level Talents Program. Liang's research was partially supported by NSF grant DMS-1418042, and by Award Number 11529101, made by National Natural Science Foundation of China.

References

Acosta, E., Wu, H., Hammer, S., Yu, S., Kuritzkes, D., Walawander, A., Eron, J., Fichtenbaum, C., Pettinelli, C., Neath, D., Ferguson, E., Saah, A. and Gerber, J. (2004). Comparison of two indinavir/ritonavir regimens in the treatment of HIV-infected individuals. *J. Acquir Immune Defic Syndr* **37**, 1358-66.

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

Buckland, S., Burnham, K. and Augustin, N. (1997). Model selection: An integral part of inference. *Biometrics* **53**, 603-618.

Burnham, K. and Anderson, D. (1998). *Model Selection and Inference: a Practical Information-theoretical Approach*. Springer, New York.

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

Claeskens, G., Croux, C. and van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics* **62**, 972-979.

Claeskens, G. and Hjort, N. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, 900-916.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*, Cambridge University Press, Cambridge.

Danilov, D. and Magnus, J. R. (2004a). Forecast accuracy after pretesting with an application to the stock market. *J. Forecast.* **23**, 251-274.

Danilov, D. and Magnus, J. R. (2004b). On the harm that ignoring pretesting can cause. *J. Econom.* **122**, 27-46.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc.* **57**, 45-70.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59**, 126-132.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econom. Theory* **21**, 60-68.

Hjort, N. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, 879-899.

Huang, Y., Liang, H. and Wu, H. L. (2008). Identifying predictors for anti-HIV treatment response: mechanism-based differential equation models versus empirical semiparametric regression models. *Statist. Medicine* **27**, 4722-4739.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34**, 2554-2591.

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

Lipsitz, S. and Fitzmaurice, G. (2009). Generalized estimating equations for longitudinal data analysis. *Longitudinal Data Analysis*, CRC Press, 43-78, Boca Raton, FL.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics* **15**, 661-675.

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

Pan, W. (2001b). Model selection in estimating equations. *Biometrics* **57**, 529-534.

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

Shen, X., Huang, H.-C. and Ye, J. (2004). Adaptive model selection and assessment for exponential family models. *Technometrics* **46**, 306-317.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. Roy. Statist. Soc.* **71**, 177-190.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817-838.

Wu, H., Huang, Y., Acosta, E., Rosenkranz, S., Kuritzkes, D., Eron, J., Perelson, A. and Gerber, J. (2005). Modeling long-term HIV dynamics and antiretroviral response: effects of drug potency, pharmacokinetics, adherence, and drug resistance., *J. Acquir Immune Defic Syndr* **39**, 272-83.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *J. Statist. Software* **21**, 1-21.

Global Biostatistical Science, Amgen Inc., Thousand Oaks, California 91320, U.S.A.

E-mail: huiy@amgen.com

Department of Mathematics, Shandong University of Technology, Zibo 255000, China.

E-mail: mathlinpeng@163.com

School of Mathematical Science, Capital Normal University, Beijing 100037, China.

E-mail: ghzou@amss.ac.cn

Department of Statistics, George Washington University, Washington, D.C. 20052, U.S.A.

E-mail: hliang@gwu.edu

(Received September 2013; accepted January 2016)