STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:642–658
Published online 8 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3489

Working-correlation-structure identification in generalized estimating equations

Lin-Yee Hin[1,*,†] and You-Gan Wang[2]

[1]*207, Yat Tung Shopping Center, Tung Chung, Hong Kong*
[2]*CSIRO Mathematical and Information Sciences, CSIRO Long Pocket Laboratories, 120 Meiers Road, Indooroopilly, Qld. 4068, Australia*

SUMMARY

Selecting an appropriate working correlation structure is pertinent to clustered data analysis using generalized estimating equations (GEE) because an inappropriate choice will lead to inefficient parameter estimation. We investigate the well-known criterion of QIC for selecting a working correlation structure, and have found that performance of the QIC is deteriorated by a term that is theoretically independent of the correlation structures but has to be estimated with an error. This leads us to propose a correlation information criterion (CIC) that substantially improves the QIC performance. Extensive simulation studies indicate that the CIC has remarkable improvement in selecting the correct correlation structures. We also illustrate our findings using a data set from the Madras Longitudinal Schizophrenia Study. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: clustered data; correlation modelling; correlation information criterion; covariance; efficiency; generalized estimating equations; model selection; QIC; working correlation structure

1. INTRODUCTION

Appropriate specification of correlation structures in longitudinal data analysis improves estimation efficiency, leading to more reliable statistical inferences [1]. That said, making the appropriate choice of correlation structure is difficult. Traditional model-selection criteria such as Akaike Information Criterion (AIC) are not useful for correlation structure selection because AIC assumes that the response observations are independent [2]. Pan [3] addressed this challenging problem by developing a refined version of AIC, named the 'quasi-log-likelihood under the independence model information criteria' (QIC(R)) for model selection in the generalized estimating equations

*Correspondence to: Lin-Yee Hin, 207, Yat Tung Shopping Center, Tung Chung, Hong Kong.
†E-mail: lyhin@netvigator.com

Received 7 April 2008
Accepted 3 October 2008

(GEE) framework. His criteria can be used for (a) variable selection in the mean function modelling, and (b) working correlation structure selection of potential candidates in the covariance modelling. QIC(R) performs well in both aspects, particularly for covariate selection.

The issue of variable selection in the GEE framework has been addressed by Pan [3], and Cantoni *et al.* [4]. The objective of this paper is to study the misspecification of working correlation structure, conditional on correctly specified mean function, in the same vein as Table 1 in Pan [3]. When the mean function is misspecified, any attempt to select the *true* correlation structure is meaningless because the residuals are distorted by the incorrect mean function. That said, once the mean function of the model is chosen, we still need to choose an appropriate 'working' correlation structure to improve estimation efficiency in the GEE context. Cui [5], Cui and Qian [2], and Kuk [6] used QIC to select the working correlation structure in their study. This paper studies how to improve correct selection of correlation structure.

Hin *et al.* [7] studied the performance profile of QIC(R) in working correlation structure selection under various simulatory conditions, reporting a correct working correlation structure detection rate of 65–98 per cent under different settings. In this paper, we propose a 'correlation information criteria' (CIC(R)) as a modification of QIC(R) to improve its performance profile in working correlation structure modelling. Note that CIC(R) is *not* intended to replace QIC(R) in the variable selection aspect of model selection, but attempts to improve the performance of QIC(R) in working-correlation-structure modelling.

In this section, we outline the theoretical framework of GEE, and define QIC(R). In Section 2, we offer the rationale for the proposal of CIC(R). Simulation reported in Section 3 compares the performance profiles of QIC(R) and CIC(R), and demonstrates that the performance of QIC(R) can be markedly improved by simply removing the first term. In addition, we demonstrate via simulation that mean function misspecification leads to deterioration of QIC(R) and CIC(R) performance profile. Section 4 compares the performance of QIC(R) and CIC(R) when applied to a real data set, and illustrates that the CIC outperforms QIC(R) substantially. Finally, Section 5 discusses some related issues.

Let there be n clusters in the data set, and let t $(t=1,\ldots,m_i)$ index observation times within the ith cluster $(i=1,\ldots,n)$ in this data set $\mathscr{D}$. Data from each cluster are represented by $\mathbf{y}_i$, an $m_i\times 1$ response vector, and $\mathbf{x}_i$, an $m_i\times p$ covariate matrix with the tth row denoted $\mathbf{x}_{it}$. Here p denotes the number of covariates including, if one is present, an intercept term. Let $\boldsymbol{\beta}$ denote a p-vector of regression parameters, and let $g(\cdot)$ denote a differentiable link function. We define $g[E(y_{it})]=\eta_{it}$ where $\eta_{it}=\mathbf{x}_{it}\boldsymbol{\beta}$ for the observation t $(t=1,\ldots,m_i)$ in cluster i, and we use $\mu_i=g^{-1}(\mathbf{x}_i\boldsymbol{\beta})$ to denote the m_i-vector of mean response values for the ith cluster. Let $v(\cdot)$ denote a variance function that defines the mean-variance relation, while $\mathbf{A}_i$ denote the $m_i\times m_i$ diagonal matrix of marginal variances for the ith cluster, with t element $\phi v(\mu_{it})$, where ϕ is the dispersion parameter.

The GEE approach [8] estimates $\boldsymbol{\beta}$ through solving the system of estimating equations

$$\mathbf{U}(\boldsymbol{\beta})=\sum_i \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)=0 \tag{1}$$

where $\mathbf{D}_i$ is the $m_i\times p$ matrix with (stacked) blocks $\partial\mu_i/\partial\boldsymbol{\beta}$, and $\mathbf{V}_i=\mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$. The working correlation matrix for cluster i is denoted by $\mathbf{R}_i(\alpha)$, an $m_i\times m_i$ square matrix depending on some q-vector of correlation parameters α.

For the clustered data set $\mathscr{D}$ defined above, the intracluster correlation structure is $\mathbf{R}_i$. However, in the formulation of QIC(R) (Section 2.2 Pan [3]), the data set $\mathscr{D}$ is assumed to contain independent observations, i.e. $\mathbf{R}_i=\mathbf{I}$. Therefore, we can define a 'working' quasi-log-likelihood as the sum of

DOI: 10.1002/sim

the quasi-log-likelihood from all observations in $\mathscr{D}$ based on the GLM framework as

$$Q(\boldsymbol{\beta}, \phi; I, \mathscr{D}) = \sum_{i=1}^{n} \sum_{t=1}^{m_i} Q(\boldsymbol{\beta}, \phi; (\mathbf{y}_{ij}, \mathbf{x}_{ij})) \tag{2}$$

as defined in Table 9.1 McCullagh and Nelder [9], and Table 4.2 Hardin and Hilbe [10]. This gives the quasi-log-likelihood under the independence model. Note that in Pan [3], the notation $Q(\boldsymbol{\beta}; I, \mathscr{D})$ is used to represent $Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})$.

In this paper, we use $Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})$ instead of $Q(\boldsymbol{\beta}; I, \mathscr{D})$. In addition, we use $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ to represent the GEE estimate of regression coefficients obtained using the hypothesized correlation structure $\mathbf{R}_i = \mathbf{R}_i(\alpha)$, while we use $\hat{\boldsymbol{\beta}}(\mathbf{I})$ to represent the GEE estimate of regression coefficients obtained under working independence $\mathbf{R}_i = \mathbf{I}$. Furthermore, we use $\hat{\phi}(\mathbf{R}_i)$ to represent the GEE estimate of dispersion parameter ϕ obtained using the hypothesized correlation structure $\mathbf{R}_i = \mathbf{R}_i(\alpha)$, while we use $\hat{\phi}(\mathbf{I})$ to represent the GEE estimate of dispersion parameter ϕ obtained under working independence $\mathbf{R}_i = \mathbf{I}$.

Define $\mathbf{V}_r$ as the asymptotic covariance of $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$. According to Liang and Zeger [8], $\mathbf{V}_r$ can be consistently estimated by

$$\hat{\mathbf{V}}_r = \sum_{i=1}^{n} (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \operatorname{cov}\{\mathbf{U}(\boldsymbol{\beta})\} (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \tag{3}$$

where $\boldsymbol{\beta}$ values in $\mathbf{D}$ and $\mathbf{V}$ are all evaluated at $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$, and that $\operatorname{cov}\{\mathbf{U}(\boldsymbol{\beta})\}$ can be consistently estimated by the outer products of gradients estimator

$$\operatorname{cov}\{\mathbf{U}(\boldsymbol{\beta})\} = \sum_{i=1}^{n} \mathbf{U}(\boldsymbol{\beta}) \mathbf{U}(\boldsymbol{\beta})' = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} E\{(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)'\} \mathbf{V}_i^{-1} \mathbf{D}_i$$

a $p \times p$ matrix, which is the sum of the outer product of the $p \times 1$ vector of empirical estimating function for all n clusters. Note that $\hat{\mathbf{V}}_r$ is free from the over dispersion parameter $\hat{\phi}$ although it appears in $\mathbf{V}_i$. The model-based variance matrix for $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ can be estimated by

$$\hat{\Omega} = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi = \hat{\phi}(\mathbf{R}_i)} \tag{4}$$

For convenience, we will also write $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ as $\hat{\boldsymbol{\beta}}$. The QIC(R) is defined as [3]

$$\mathrm{QIC(R)} \equiv -2Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D}) + 2\operatorname{tr}(\hat{\Omega}_I \hat{\mathbf{V}}_r) \tag{5}$$

where

$$Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D}) = Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{R}_i)}$$

$$\hat{\Omega}_I = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i \Big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi = \hat{\phi}(\mathbf{R}_i)} \quad \text{and} \quad \hat{\mathbf{V}}_r = \mathbf{V}_r|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{R}_i)}$$

Note that all three entities $Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D})$, $\hat{\Omega}_I$, and $\hat{\mathbf{V}}_r$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\mathbf{R}_i)$.

DOI: 10.1002/sim

2. A DERIVATIVE OF QIC(R)–CIC(R)

The QIC(R) can be decomposed into

$$\mathrm{QIC}_P(R) = T_1(R) + T_{2P}(R) \tag{6}$$

$$T_1(R) = -2Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi=\hat{\phi}(\mathbf{R}_i)} \tag{7}$$

and

$$T_{2P}(R) = 2\,\mathrm{tr}[(\hat{\Omega}_I|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi=\hat{\phi}(\mathbf{R}_i))\hat{\mathbf{V}}_r|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i)}] \tag{8}$$

$T_1(R)$ denotes the sum of quasi-log-likelihood for the $\sum_{i=1}^{n} m_i$ observations in the data set $\mathscr{D}$. Note that $E(T_1(R)) = -2\sum_i E\{Q(\beta_T, \phi; I, y_i)\}$ is free from both R and R_T, R_T being the true intracluster correlation structure. As we will see, however, $E(T_{2P}(R))$ does depend on the R and R_T matrix. It therefore makes sense to ignore T_1 when comparing different correlation models.

In order to better understand the two terms, $T_1(R)$ and $T_{2P}(R)$, for model selection, we examine $\mathrm{QIC}_P(R)$, a version of approximate expected Kullback–Leibler distance, using the covariance penalty theory described by Efron [11].

Making use of the *q class* error measures $Q(y_{it}, \mu_{it})$, which are the difference between a concave function $q(y_{it})$ and its tangent through a point $(\hat{\mu}_{it}, q(\hat{\mu}_{it}))$ as proposed by Efron [12], we define $q(y_{it}) = -2Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D})$, making $Q(y_{it}, \mu_{it})$ the deviance, i.e. twice the Kullback–Leibler distance. Comparing $\mathrm{QIC}_P(R)$ and equation (3.26) Efron [12] reveals that $T_1(R)$ is the *apparent error rate*, i.e. the observed inaccuracy of the fitted model applied to the original data points, $T_{2P}(R)$ is the covariance penalty term.

For mean function modelling, $\mathrm{QIC}_P(R)$ can be used as a model selection measure, as error or inadequacy of mean function modelling will be reflected in the expected apparent error, $T_1(R)$. However, for covariance function modelling, the efficiency impairment of parameter estimation can better be reflected in the expected covariance penalty term, $T_{2P}(R)$, as this entity can be viewed as a ratio of the robust covariance against the model-based covariance. If we use $\mathrm{QIC}_P(R)$ for covariance structure selection, the random error incurred during parameter estimation of $\hat{\boldsymbol{\beta}}$ that is of the same order of magnitude as $T_{2P}(R)$, causing the information reflected through $T_{2P}(R)$ to be masked by the random error of $T_1(R)$. Therefore, we propose using $T_{2P}(R)$ alone for covariance structure modelling.

In addition, $T_1(R)$ relates to the quasi-log-likelihood for independent observations, and hence does not contain information about the hypothesized intracluster correlation structure. On the contrary, $T_{2P}(R)$ contains information of the hypothesized correlation structure via $\hat{\mathbf{V}}_r$. Therefore, it is logical to expect that $T_1(R)$ would not be informative in intracluster correlation structure identification.

For these reasons, we propose a simple modification of the QIC(R) as follows:

$$\mathrm{CIC(R)} = \mathrm{tr}\left[\left(\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi=\hat{\phi}(\mathbf{R}_i)}\right) \hat{\mathbf{V}}_r \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i)}\right] \tag{9}$$

DOI: 10.1002/sim

Note that $T_{2P}(R) = 2 \times \mathrm{CIC(R)}$. Connecting with (3), we can express (9) as

$$\mathrm{CIC(R)} = \mathrm{tr}\left[\sum_{i=1}^{n} (\mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i)^{-1} (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathrm{var}(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D})_i^{-1} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i), \phi=\hat{\phi}(\mathbf{R}_i)} \right]$$

which has a limit as $n \to \infty$

$$\lim_{n\to\infty} \mathrm{CIC(R)} = \mathrm{tr}(\Omega_I V_R)$$

in which Ω_I^{-1} and V_R are the limits of $n^{-1} \sum_{i=1}^{n} (\mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i)^{-1}$ and

$$n^{-1} (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathrm{var}(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1}$$

evaluated at the true parameter values, $\boldsymbol{\beta} = \boldsymbol{\beta}_T$. In an ideal situation when the correlation is correctly specified, $\mathbf{V} = \mathrm{var}(\mathbf{y}_i)$, and we have,

$$\lim_{n\to\infty} \mathrm{CIC}(R_T) \to \mathrm{tr}(\Omega_I V_{R_T}) = \mathrm{tr}\left[\sum_{i=1}^{n} (\mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i)(\mathbf{D}_i' \,\mathrm{var}^{-1}(Y_i) \mathbf{D}_i)^{-1} \right] \tag{10}$$

It is well known that V_R reaches the minimum when $\mathbf{R} = \mathbf{R}_T$ (Gauss–Markov Theorem), the trace term $\mathrm{tr}(\Omega_I \mathbf{V}_r)$ therefore also reaches the minimum at $\mathbf{R}_T$, i.e. in general, $\mathrm{CIC(R)} \geqslant \mathrm{CIC}(R_T)$.

Rewriting (10), we have

$$\mathrm{tr}\left[\sum_{i=1}^{n} (\mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i)(\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \right] = p + \mathrm{tr}\left[\sum_{i=1}^{n} \{\mathbf{D}_i' (\mathbf{A}_i^{-1} - \mathbf{V}_i^{-1}) \mathbf{D}_i\} (\mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1} \right] \tag{11}$$

In the special case when all observations are independent, $\mathbf{A}_i^{-1} = \mathbf{V}_i^{-1}$, the second term in (11) vanishes, CIC(R) reduces to p, and $\mathrm{QIC}_P(R)$ reduces to a form similar to the AIC.

The results above concur with the view expressed in Section 2.4 Pan [3] that: (a) $T_{2P}(R)$ is vital for working correlation structure selection and (b) the limiting value of $\mathrm{tr}(\Omega_I \hat{V}_r)$, i.e. p, when the model-based covariance and the robust covariance are asymptotically equal *cannot* be used to replace the quadratic term in correlation structure selection. From (11) above, the model-based covariance and robust covariance can be equal in the QIC(R) construct only when $\mathbf{A}_i = \mathbf{V}_i$, i.e. $\mathbf{R}_i = \mathbf{I}$, i.e. when the observations in the data set are independent.

In the next section, we will compare, via simulation, the performance profile of CIC(R) with QIC(R) to investigate whether $\mathrm{tr}(\Omega_I \hat{V}_r)$ alone confers better performance profile than QIC(R).

Hardin and Hilbe [10] have a slightly different interpretation of the QIC(R),

$$\mathrm{QIC(R)} = -2Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D}) + 2\,\mathrm{tr}(\hat{\Omega}_I \hat{\mathbf{V}}_r) \tag{12}$$

where $Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D})$ and $\hat{\mathbf{V}}_r$ are the same as (5), but $\hat{\Omega}_I$ is slightly different,

$$\hat{\Omega}_I = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{I}), \phi=\hat{\phi}(\mathbf{I})}$$

The version of QIC(R) defined by Hardin and Hilbe [10] stated in (12) can be decomposed as

$$\mathrm{QIC}_{HH}(R) = T_1(R) + T_{2HH}(R) \tag{13}$$

DOI: 10.1002/sim

and

$$T_{2HH}(R) = 2\,\mathrm{tr}\left[\left(\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{I}),\phi=\hat{\phi}(\mathbf{I})}\right) \hat{\mathbf{V}} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i)}\right] \tag{14}$$

where $\hat{V}_r$ is constructed under the hypothesized correlation structure $\mathbf{R}_i$ in that $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$.

According to the notation in Hardin and Hilbe [10], $Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, \mathscr{D})$ and $\hat{\mathbf{V}}_r$ are evaluated using $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ and $\hat{\phi}(\mathbf{I})$. However, $\hat{\Omega}_I$ is evaluated using $\hat{\boldsymbol{\beta}}(\mathbf{I})$ and $\hat{\phi}(\mathbf{I})$ instead of $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ and $\hat{\phi}(\mathbf{R}_i)$ as in Pan [3]. The induced difference between $T_{2P}(R)$ and $T_{2HH}(R)$ is $O_p(\hat{\boldsymbol{\beta}}(\mathbf{R}_i) - \hat{\boldsymbol{\beta}}(\mathbf{I}))$, which is $O_p(n^{-1/2})$. We therefore expect little effect on CIC(R) in using either $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ or $\hat{\boldsymbol{\beta}}(\mathbf{I})$ to evaluate $\hat{\Omega}_I$.

3. SIMULATION STUDIES

The simulation study in this section consists of two parts. Part I: we assess the performance profile of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R) in correlation structure selection when the mean function is *correctly* specified. Part II: we investigate the performance profile of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R) in correlation structure selection when the mean function is *misspecified*.

Part I: We assess and compare the performance of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R) in detecting the correct working intracluster correlation structure between exchangeable and AR(1) working correlation structures. Simulation settings considered allow performance comparison between continuous and discrete responses.

We conduct a simulation study with 1000 independent realizations for each of the following settings:

(i) Gaussian response, exchangeable or AR(1) intracluster correlation structure, $\alpha = 0.4$, and $\mu_{it} = 0.3x_{1t} + 0.3x_{2t}$.
(ii) Binary response, exchangeable or AR(1) intracluster correlation structure, $\alpha = 0.4$, and $\mathrm{logit}(\mu_{it}) = 0.62x_{1t} + 0.62x_{2t}$.

Each independent realization contains 100 balanced clusters ($n = 100$) of size 5 ($m = 5$), where x_{1t}, x_{2t} are observation-level covariates generated at random from uniform distribution U[0, 1]. We repeat the simulation for all the settings above for a smaller sample size ($n = 30, m = 5$). All results are shown in Table I for comparison.

All computations are performed using *R* version 2.5.0 [13], with GEE fitting performed using the `yags` library [14]. Gaussian random variable generation was performed using `MASS` library [15], and the binary random variables generated using `bindata` library [16].

$\mathrm{QIC}_P(R)$ demonstrates a detection rate of approximately 70 per cent when the true intracluster correlation structure is either exchangeable or AR(1), and the response variable being either Gaussian or binary. This performance is comparable to that originally reported by Pan [3].

$\mathrm{QIC}_{HH}(R)$ demonstrates a detection rate similar, but not identical, to that of the version of QIC(R) proposed by Pan [3]. This can be explained by the fact that Hardin and Hilbe [10] used $\hat{\boldsymbol{\beta}}(\mathbf{I})$ to evaluate the model-based covariance matrix as in (14), contrary to Pan [3] who used $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ to evaluate the model-based covariance matrix as in (8).

CIC(R) demonstrates a detection rate of approximately 95 per cent when the true intracluster correlation structure is either exchangeable or AR(1), and the response variable being either

DOI: 10.1002/sim

Gaussian or binary. This performance is better than the version of QIC(R) proposed by Pan [3], and that by Hardin and Hilbe [10], respectively.

The fact that $T_{2P}(R) = 2 \times \text{CIC(R)}$ and yet the detection rate of $\text{QIC}_P(R)$ is lower than that of CIC(R) raises the suspicion that $T_1(R)$ offsets the detection rate conferred by $T_{2P}(R)$. Therefore, we examine

(a) the performance of $T_{2P}(R)$ and $T_{2HH}(R)$ in detecting the correct intracluster correlation structure,
(b) the marginal distribution of $\text{QIC}_P(R)$ and $\text{QIC}_{HH}(R)$ as well as their components, and
(c) the power of $\text{QIC}_P(R)$, $\text{QIC}_{HH}(R)$ and CIC(R) in identifying the true intracluster correlation structure in the presence of a wrongly specified one.

We assess the performance of QIC(R) and $T_{2HH}(R)$ in correlation structure identification (Table I). The detection rate of $T_{2P}(R)$ is higher than $\text{QIC}_P(R)$ and that of $T_{2HH}(R)$ is higher than $\text{QIC}_{HH}(R)$. In addition, the detection rate of CIC(R) is very similar to that of $T_{2HH}(R)$. Since $T_{2P}(R) = 2 \times \text{CIC(R)}$, the detection rates and power of $T_{2P}(R)$ and CIC(R) are the same.

From (7), we have $T_1(R) = -2Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i)}$. Figure 1(a) shows that the distributions of $-2Q(\boldsymbol{\beta}, \phi; I, \mathscr{D})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_i)}$ when $\mathbf{R}_i$ is working independence, exchangeable, or AR(1) are very

Table I. Frequencies of the intracluster correlation structure identified using three different criteria, $\text{QIC}_P(R)$, $\text{QIC}_{HH}(R)$, and CIC(R), from 1000 independent replications in each setting. The magnitude of intracluster correlation α is 0.4 in all settings.

	Gaussian response											
	True ICS is Ex						True ICS is AR(1)					
	$n=30$			$n=100$			$n=30$			$n=100$		
	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)
$\text{QIC}_P(R)$	135	**617**	248	84	**687**	229	173	256	**571**	81	199	**720**
$\text{QIC}_{HH}(R)$	139	**610**	251	86	**684**	230	178	254	**568**	81	200	**719**
CIC(R)	15	**809**	176	0	**962**	38	26	153	**821**	1	40	**959**
$T_{2HH}(R)$	17	**809**	174	0	**963**	37	28	159	**813**	1	40	**959**
	Binary response											
	True ICS is Ex						True ICS is AR(1)					
	$n=30$			$n=100$			$n=30$			$n=100$		
	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)
$\text{QIC}_P(R)$	74	**723**	203	72	**721**	207	105	189	**706**	100	175	**725**
$\text{QIC}_{HH}(R)$	85	**716**	199	89	**702**	209	109	197	**694**	107	173	**720**
CIC(R)	13	**805**	182	0	**956**	44	25	142	**833**	0	28	**972**
$T_{2HH}(R)$	26	**780**	194	1	**933**	66	36	142	**822**	0	32	**968**

Note: True ICS denotes true intracluster correlation structure (ICS) of simulated data. Ind, Ex, and AR(1) refer to ICS being independence, exchangeable and AR(1) respectively. $\text{QIC}_P(R)$ is QIC(R) calculated according to the interpretation by Pan [3]. $\text{QIC}_{HH}(R)$ is QIC(R) calculated according to the interpretation by Hardin and Hilbe [10]. $T_{2P}(R)$ is the second term in $\text{QIC}_P(R)$ where $T_{2P}(R) = 2 \times \text{CIC(R)}$. $T_{2HH}(R)$ is the second term in $\text{QIC}_{HH}(R)$. n is the number of balanced clusters (each of size 5) in the simulated data set generated in each replication. The frequencies in bold are the frequencies of correct ICS identification.

DOI: 10.1002/sim

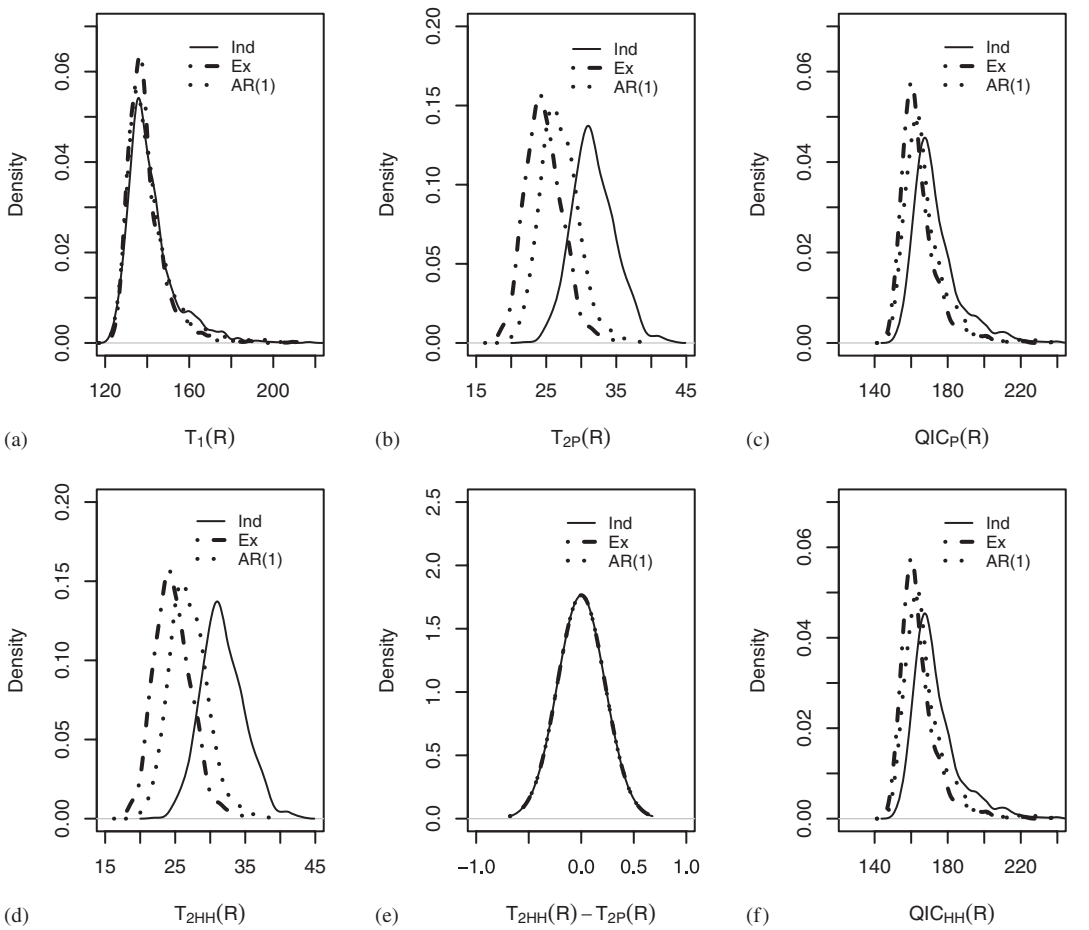

Figure 1. Binary response, true intracluster correlations is exchangeable: Kernel density estimates of the distributions when $\mathbf{R}_i$ is working independence, exchangeable, and AR(1): (a) $T_1(R)$; (b) $T_{2P}(R)$; (c) $\mathrm{QIC}_P(R)$; (d) $T_{2HH}(R)$; (e) $T_{2HH}(R)-T_{2P}(R)$ and (f) $\mathrm{QIC}_{HH}(R)$; (1000 independent replications for $n=100$, $m=5$).

similar regardless of the true intracluster correlation structure. That said, they are not identical (Table II).

The further apart the marginal distributions of $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ under three different hypothesized correlation structures, the higher their respective detection rate and power in identifying the true correlation structure from the wrongly specified one. While $T_{2P}(R)$ and $T_{2HH}(R)$ demonstrate clear distribution differentiation under different hypothesized working correlation structures (working independence exchangeable, and AR(1)), $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ do not. Instead, the pattern of marginal distribution of $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ resembles that of T_1. This is further evidence that T_1 offsets the detection rate and power of $T_{2P}(R)$ and $T_{2HH}(R)$. Kernel density estimate for the marginal distributions of $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ and their components illustrating this point, when the true intracluster correlation structure is exchangeable and

DOI: 10.1002/sim

Table II. Summary statistics for $T_1(R)$ evaluated assuming working correlation structure being working independence, exchangeable, or AR(1) from 1000 independent replications in each setting.

	Gaussian response			
	True ICS is Ex		True ICS is AR(1)	
	$n=30$ Mean (SD)	$n=100$ Mean (SD)	$n=30$ Mean (SD)	$n=100$ Mean (SD)
Working independence	145.80 (20.80)	497.42 (40.18)	147.46 (19.49)	496.02 (34.61)
Exchangeable	146.58 (21.10)	498.23 (40.42)	147.76 (19.62)	496.30 (34.70)
AR(1)	146.38 (21.07)	497.97 (40.33)	147.98 (19.70)	496.48 (34.72)
	Binary response			
	True ICS is Ex		True ICS is AR(1)	
	$n=30$ Mean (SD)	$n=100$ Mean (SD)	$n=30$ Mean (SD)	$n=100$ Mean (SD)
Working independence	50.04 (14.66)	142.03 (12.57)	47.94 (10.29)	140.12 (10.29)
Exchangeable	47.41 (12.29)	139.79 (9.92)	47.30 (9.93)	139.66 (9.73)
AR(1)	48.45 (13.92)	140.65 (10.95)	46.64 (9.08)	138.98 (9.19)

the response variable is binary, is displayed in Figure 1. Similar patterns are observed for cases when the response variable is Gaussian, and when the intracluster correlation structure is either exchangeable or AR(1) (figures not shown). The distributions of $T_{2HH}(R)-T_{2P}(R)$ centers around zero as shown in Figure 1(f). The fact that this difference is much smaller than the $T_{2HH}(R)$ and $T_{2P}(R)$ values demonstrate our earlier contention that evaluation of $\hat{\Omega}_T$ using $\hat{\boldsymbol{\beta}}(\mathbf{R}_i)$ and $\hat{\phi}(\mathbf{R}_i)$, or $\hat{\boldsymbol{\beta}}(\mathbf{I})$ and $\hat{\phi}(\mathbf{I})$ only amount to a small difference of $O_p(n^{-1/2})$ in T_2.

To investigate the impact of $T_1(R)$ on the power to differentiate the true intracluster correlation structure from an incorrectly specified one, we compare the distributions of

(a) $\mathrm{QIC}_P(R_1)-QIC_P(R_T)$,
(b) $\mathrm{QIC}_{HH}(R_1)-\mathrm{QIC}_{HH}(R_T)$,
(c) $\mathrm{CIC}(R_1)-\mathrm{CIC}(R_T)$, and
(d) $T_{2HH}(R_1)-T_{2HH}(R_T)$

where R_1 is the incorrectly specified intracluster correlation structure and R_T is the true intracluster correlation structure, which is either exchangeable or AR(1). The power of $\mathrm{QIC}_P(R)$ to differentiate the true intracluster correlation structure from a wrongly specified correlation structure is the area under the corresponding curve bounded by $\mathrm{QIC}_P(R_1)-\mathrm{QIC}_P(R_T)>0$. The same applies to $\mathrm{QIC}_{HH}(R)$ and CIC(R).

For both Gaussian and binary responses with exchangeable or AR(1) correlation structures, $T_1(R)$ demonstrates power of 50 per cent or less (Figure 2), which is weak for a correlation detection tool. On the contrary, the power of CIC(R) is much higher as shown in Figure 3, which compares the power of CIC(R) and QIC(R) for selecting the correct correlation structure when that true correlation structure is either exchangeable or AR(1) for binary responses. As we can see, $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ have very similar performance in identifying different intracluster

DOI: 10.1002/sim

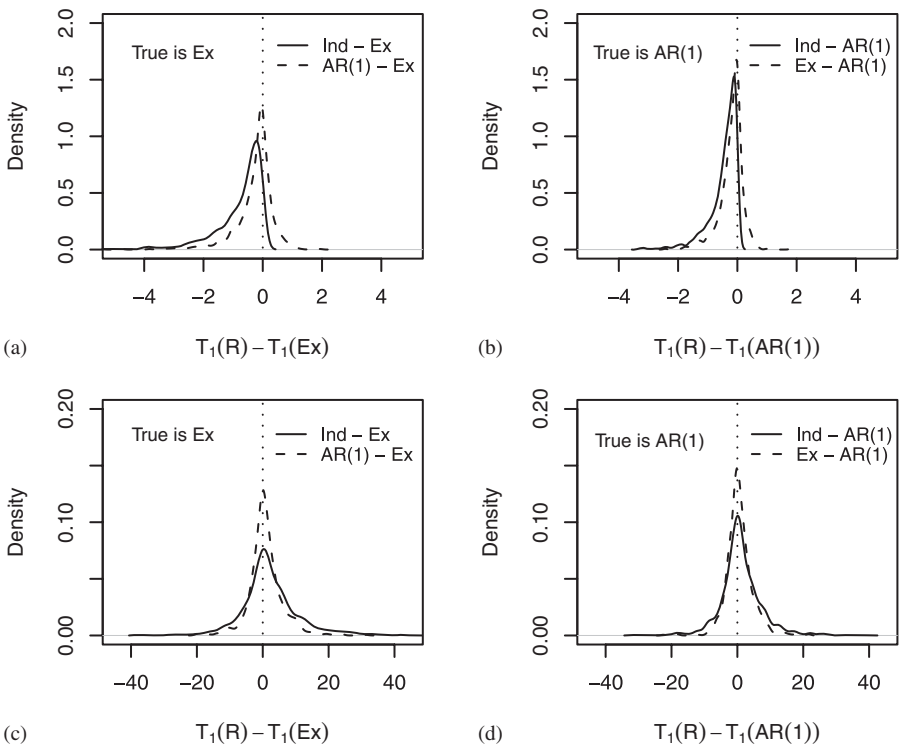

Figure 2. Marginal distribution patterns of $[T_1(R_1)-T_1(R_T)]$ where R_1 is the incorrectly specified correlation and $T_1(R_1)=-2Q(\boldsymbol{\beta},\phi;I,\mathscr{D})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_1)}$, while R_T is the true correlation, and $T_1(R_T)=-2Q(\boldsymbol{\beta},\phi;I,\mathscr{D})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\mathbf{R}_T)}$. For the left panel, R_T is exchangeable (Ex); for the right panel, R_T is AR(1). The vertical line indicates zero difference. The area under curve for positive difference can be regarded as the power of differentiating the true correlation structure from the incorrectly specified correlation structure (1000 independent replications for $n=100$, $m=5$).

correlation structures. However, the performance of CIC(R) is much better. The distribution patterns of $\mathrm{CIC}(R_1)-\mathrm{CIC}(R_T)$ (Figure 3) and $T_{2HH}(R_1)-T_{2HH}(R_T)$ (not displayed here) are similar. The pattern for Gaussian responses is very similar and hence not presented here. This shows that the presence of $T_1(R)$ reduces the power of $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$. Also, similar patterns are also observed for smaller sample sizes ($n=30, m=5$).

Part II: At the suggestion of both reviewers, we investigated the performance of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R) in detecting the correct working intracluster correlation structure in the presence of mean function misspecification. We do so for the scenario when the response variable is Gaussian, true correlation structure is AR(1) with a magnitude of 0.4, and the true response generating marginal mean function is $\mu_{it}=\beta_1 x_{1t}+\beta_2 x_2$, with $\beta_1=1$ and three different values for β_2, 0.5, 1, and 1.5. Here x_{1t} is an observation-level covariate, x_2 is a cluster-level covariate, both generated at random from random Gaussian distribution. We also introduce an irrelevant observation level random variable $x_{3t}=3\times \mathrm{U}[0,1]$, where U[0, 1] is a random uniform variable, in that x_{3t} is not used to generate the response variable in the simulated data sets. We investigate

DOI: 10.1002/sim

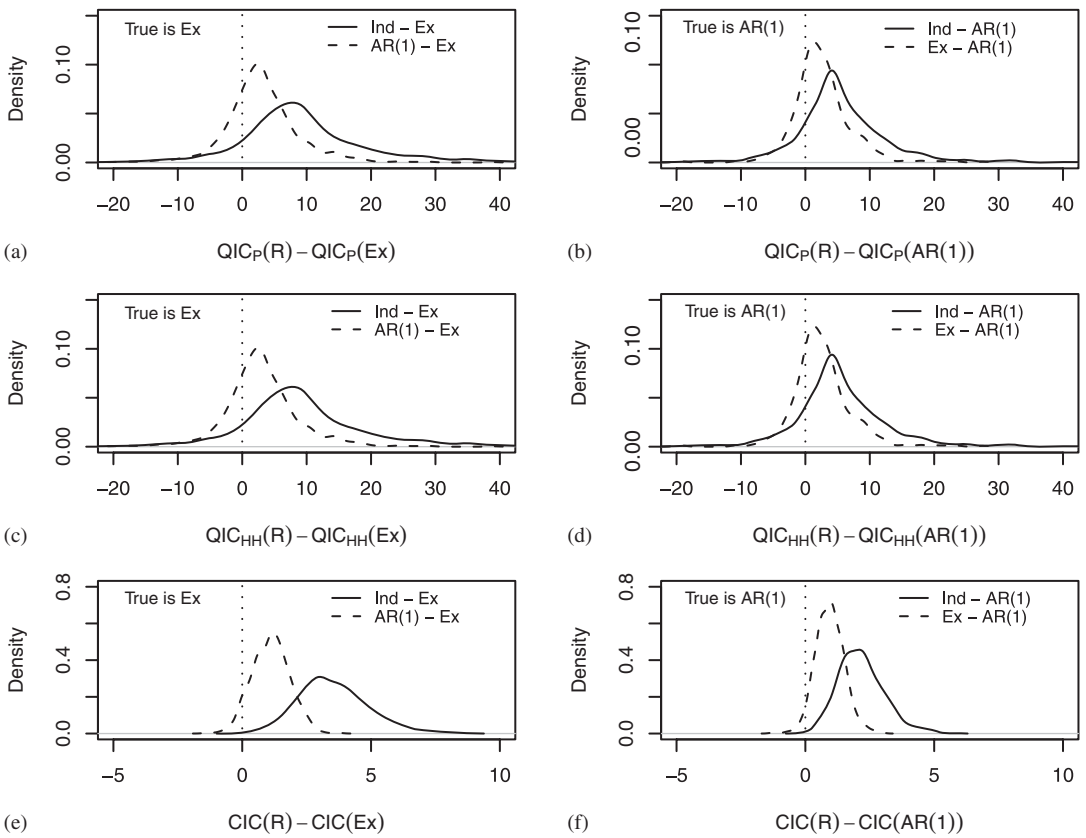

Figure 3. Binary responses. Kernel density estimates of the distribution patterns of $\mathrm{QIC}_P(R_1)-\mathrm{QIC}_P(R_T)$, $\mathrm{QIC}_{HH}(R_1)-\mathrm{QIC}_{HH}(R_T)$, and $\mathrm{CIC}(R_1)-\mathrm{CIC}(R_T)$ where R_1 is incorrectly specified while R_T is the true correlation. For the left panel, R_T is exchangeable (Ex); for the right panel, R_T is AR(1). The vertical line indicates zero difference. The area under curve for positive difference can be regarded as the power of differentiating the true correlation structure from the incorrectly specified correlation structure (1000 replications for $n=100, m=5$).

the performance of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R) when the mean function is specified by including the random variables in the following combinations: (1) x_{1t}, (2) x_2, (3) x_{1t} and x_2, i.e. the true mean, (4) x_{1t} and x_{3t}, (5) x_2 and x_{3t}, or (6) x_{1t}, x_2 and x_{3t}. For each of the six hypothetical mean function specification under the three different settings, 1000 simulation runs are generated, and the results tabulated in Table III. The results when the true mean function is modelled using x_{1t} and x_2 are also displayed side by side for ease of comparison. Results for mean function combinations $\{x_2\}$. $\{x_{1t}, x_2\}$, $\{x_2, x_{3t}\}$, and $\{x_{1t}, x_2, x_{3t}\}$ are not displayed in Table III when true mean function is $\mu_{it}=x_{1t}+x_2$, or $\mu_{it}=x_{1t}+1.5x_2$ because their results are not affected by the β_2 values when the mean model contains x_2.

Results in Table III show that in the presence of covariate mis-specification, performance of CIC(R) and $T_{2HH}(R)$ deteriorates. The more important the covariate, i.e. the more it contributes to the marginal mean generating process, its omission from mean function modelling will lead to a

DOI: 10.1002/sim

Table III. Frequencies of the intracluster correlation structure identified using the three correlation selection criteria, $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R), from 1000 independent replications in each setting of 100 clusters ($n=100$) of size 5.

	$\mu_{it}=x_{1t}+0.5x_2$															
	x_{1t}			x_2		x_{1t},x_2			x_{1t},x_{3t}			x_2,x_{3t}			x_{1t},x_2,x_{3t}		
	Ind	Ex	AR(1)	Ind/Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)
$\mathrm{QIC}_P(R)$	151	293	**556**	480	**520**	125	187	**688**	119	299	**582**	251	314	**435**	91	225	**684**
$\mathrm{QIC}_{HH}(R)$	151	293	**556**	480	**520**	127	187	**686**	121	300	**579**	254	313	**433**	94	227	**679**
CIC(R)	2	130	**868**	371	**629**	8	88	**904**	0	88	**912**	71	322	**607**	1	61	**938**
$T_{2HH}(R)$	2	130	**868**	371	**629**	8	88	**904**	0	85	**915**	72	322	**606**	1	61	**938**
	$\mu_{it}=x_{1t}+x_2$															
	x_{1t}			x_2		x_{1t},x_2			x_{1t},x_{3t}			x_2,x_{3t}			x_{1t},x_2,x_{3t}		
	Ind	Ex	AR(1)	Ind/Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)
$\mathrm{QIC}_P(R)$	244	308	**448**	—	—	—	—	—	271	332	**397**	—	—	—	—	—	—
$\mathrm{QIC}_{HH}(R)$	244	308	**448**	—	—	—	—	—	272	331	**397**	—	—	—	—	—	—
CIC(R)	0	210	**790**	—	—	—	—	—	0	190	**810**	—	—	—	—	—	—
$T_{2HH}(R)$	0	210	**790**	—	—	—	—	—	0	191	**809**	—	—	—	—	—	—
	$\mu_{it}=x_{1t}+1.5x_2$															
	x_{1t}*			x_2		x_{1t},x_2			x_{1t},x_{3t}†			x_2,x_{3t}			x_{1t},x_2,x_{3t}		
	Ind	Ex	AR(1)	Ind/Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)	Ind	Ex	AR(1)
$\mathrm{QIC}_P(R)$	357	304	**333**	—	—	—	—	—	451	261	**285**	—	—	—	—	—	—
$\mathrm{QIC}_{HH}(R)$	357	304	**333**	—	—	—	—	—	451	261	**285**	—	—	—	—	—	—
CIC(R)	0	292	**702**	—	—	—	—	—	0	276	**721**	—	—	—	—	—	—
$T_{2HH}(R)$	0	292	**702**	—	—	—	—	—	0	276	**721**	—	—	—	—	—	—

True intracorrelation is AR(1) of magnitude 0.4. Covariates x_{1t} and x_2 are used to generate the true mean. The irrelevant observation-level covariate x_{3t} is introduced as a contamination variable in mean modelling. Six candidate models using different combinations of covariates are investigated. Frequencies of correct correlation structure selection are displayed in bold.

*994 realized replications.

†997 realized replications.

DOI: 10.1002/sim

greater extent of performance deterioration of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, and CIC(R). As β_2 increases from 0.5 to 1.5 in the true response generating marginal mean, the detection rate of CIC(R) for models with misspecified mean $\mu_{it}=x_{1t}$ decreases from 86.8 to 70.2 per cent, while the detection rate of QIC(R) decreases concurrently from 55.6 to 33.3 per cent. Similarly, as β_2 increases from 0.5 to 1.5 in the true response generating marginal mean, the detection rate of CIC(R) for models with misspecified mean $\mu_{it}=x_{1t}+x_{3t}$ decreases from 91.2 to 72.1 per cent, while the detection rate of QIC(R) decreases concurrently from 58.2 to 28.5 per cent, respectively. However, in the cases of using a misspecified mean function, even if the true correlation structure is identified, it is not clear if it will help improving the estimation efficiency. We will further discuss this in Discussion.

4. AN EXAMPLE

We use the data set from the Madras Longitudinal Schizophrenia Study available from Diggle *et al.* [17] as an example. This study investigates the course of positive and negative psychiatric symptoms over the first year after initial hospitalization for schizophrenia. The response variable (Y) is binary, indicating the presence of thought disorder. The covariates are (a) Month—duration since hospitalization (in months), (b) Age—age of patient at the onset of symptom (1 represents Age <20; 0 otherwise), (c) Gender—gender of patient (1 = female; 0 = male), (d) Month$\cdot$Age—Month$\cdot$Age—Interaction term between variables Month and Age, and (e) Interaction term between variables Month and Gender.

We fit three GEE models with the same canonical link relationship

$$\mathrm{logit}[Y_{it}]=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\mathrm{Month}+\boldsymbol{\beta}_2\mathrm{Age}+\boldsymbol{\beta}_3\mathrm{Gender}+\boldsymbol{\beta}_4\mathrm{Month}\cdot\mathrm{Age}+\boldsymbol{\beta}_5\mathrm{Month}\cdot\mathrm{Gender}$$

under three different hypothesized working correlation structures, working independence, exchangeable, and AR(1). The parameter estimates from the three models differ markedly (Table IV). In addition, their corresponding standard errors under different hypothesized correlation structures vary, with the AR(1) model producing the smallest standard errors for all the $\boldsymbol{\beta}$ estimates. For each model, we calculate and compare the values of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, CIC(R), as well as $T_1(R)$, $T_{2P}(R)$, and $T_{2HH}(R)$. It is interesting to see that $\mathrm{QIC}_P(R)$ selects the independence model while $\mathrm{QIC}_{HH}(R)$ selects AR(1). The proposed CIC(R) selects AR(1) as the most appropriate model, and the corresponding correlation parameter is 0.632 indicating that the independence model may be inappropriate. It is reassuring to note that $T_{2HH}(R)$ also selects AR(1) as the most appropriate correlation structure. In this example, we can see that the effect of $T_1(R)$ on $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ can be detrimental, although it may not be so in other cases.

From Table IV we can see that, in absolute terms, the difference in magnitude of the $\mathrm{QIC}_p(R)$ between 'working-independence' and AR(1) is actually very small. We can decompose these QIC values into two additive terms: (1) a large constant $E[T_1(R)]$, which is independent of R and (2) a relatively smaller term T_2 that is dependent upon R. As the sample size n increase, the magnitude of $E(T_1)$ increases linearly with n, but the magnitude of T_2 remains unchanged. Therefore, for two competing hypothesized working correlation structures, R_1 and R_2, the absolute difference of their QIC values, $|\mathrm{QIC}_P(R_1)-\mathrm{QIC}_P(R_2)|$, will decrease as the sample size increases due to the increase in the magnitude of $E(T_1)$ as the sample size increases.

DOI: 10.1002/sim

Table IV. Parameter estimates of $\boldsymbol{\beta}$, ϕ, and α from the Madras data using GEE under different hypothesized correlation structures.

Covariates	$\hat{\boldsymbol{\beta}}$ (Robust SE) under Working Independence	$\hat{\boldsymbol{\beta}}$ (Robust SE) under Exchangeable	$\hat{\boldsymbol{\beta}}$ (Robust SE) under AR(1)
Intercept	0.64 (0.30)	0.62 (0.31)	0.54 (0.29)
Month	−0.25 (0.05)	−0.27 (0.06)	−0.23 (0.05)
Age	0.81 (0.49)	1.05 (0.54)	0.62 (0.45)
Gender	−0.38 (0.44)	−0.59 (0.52)	−0.13 (0.41)
Month·Age	−0.13 (0.09)	−0.08 (0.09)	−0.09 (0.08)
Month·Gender	−0.11 (0.09)	−0.13 (0.09)	−0.15 (0.08)
	Hypothesized correlation structure		
	Working Independence	Exchangeable	AR(1)
$\hat{\alpha}$	0	0.27	0.63
$\mathrm{QIC}_P(R)$	**955.32**	964.50	955.95
$\mathrm{QIC}_{HH}(R)$	955.32	957.10	**952.82**
CIC(R)	19.01	19.97	**18.54**
$T_1(R)$	**917.29**	926.56	918.86
$T_{2P}(R)$	38.02	37.94	**37.09**
$T_{2HH}(R)$	38.02	39.80	**35.52**

The corresponding values of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$, CIC(R), and their respective components are tabulated, with the lowest value among the three hypothesized correlation structure in bold. $T_{2P}(R) = 2 \times \mathrm{CIC(R)}$. $T_{2P}(R)$ is also shown for ease of comparison with $T_{2HH}(R)$.

Different estimates of $\hat{\beta}$ will lead to different estimates of $E(T_1)$ (which is free from the true R). Therefore, we recommend comparing T_2 values whose expectation depends on the R used. This is part of the reason why we suggest using T_2 for correlation R selection.

We can look at this from another perspective, $\mathrm{QIC}_p(R)$, being a version of the approximate expected Kullback–Leibler distance, measures the 'distance of the hypothetical model to the approximation of the true model'. The Kullback–Leibler distance is a directed or oriented distance between a hypothetical model and the truth, and that the role of the truth and model are not interchangeable. Therefore, we are uncertain about the geometrical representation of the absolute difference between two such distances, and the geometrical implication it carries towards interpretation of the difference between two such competing models.

It is worth noting that in Table 3 of Pan [3] depicting results of variable selection using QIC calculated under working independence, the absolute difference in the values of QIC is also very small, in the order of first decimal place only.

5. DISCUSSION

Although $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ are slightly different, their performance profiles are similar. In addition, when the first terms are removed from $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$, detection rate and power are improved markedly.

DOI: 10.1002/sim

Contrary to $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$, the detection rates (Table I) of CIC(R) and $T_{2HH}(R)$ improve markedly as the number of clusters (n) in the data set increases. In addition, the power of CIC(R) increases as n increases. However, the power of $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ does not change much as n increases (figures for $n=30, m=5$ not shown). Therefore, CIC(R) and $T_{2HH}(R)$ are better detection tools than $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$.

The fact that CIC(R) demonstrates a sensitivity and specificity of around 95 per cent when the number of clusters in the data set is large ($n=100$) makes it a useful tool for clustered data analysis. Even when the number of clusters in the data set is very small ($n=30$), it still confers a sensitivity and specificity of around 80 per cent.

CIC(R) appears to predict the correct correlation structure more frequently than $\mathrm{QIC}_P(R)$ and $\mathrm{QIC}_{HH}(R)$ for Gaussian response with identity link and binary response with logit link with either exchangeable or AR-1 intracluster correlation structure under balanced-clusters data situations. The simulation study performed in this article only examines the performance profile of CIC(R) in choosing between exchangeable and AR(1) correlation structures. The application of CIC(R) in choosing among other competing structures is an open research question. In our simulation study, the dispersion parameter ϕ is known to be 1 and need not be estimated. For the scenario of Gaussian response with identity link and constant variance, when ϕ needs to be estimated, and estimated performed using the bias corrected generalized Pearson statistic, $T_1(R)$ will reduce to $(K-p)$ which is a constant, and the performance of $\mathrm{QIC}_P(R)$, $\mathrm{QIC}_{HH}(R)$ and CIC(R) will become the same.

Pan [3] proposed two areas of application for QIC(R): (1) using QIC(R) to select the appropriate intracluster correlation structure and (2) using QIC(I), i.e. assuming $\mathbf{R}_i=\mathbf{I}$, to select the appropriate combination of covariates. Our proposed modification is suitable for the first purpose only, i.e. to select the appropriate intracluster correlation structure. It is important to bear in mind that CIC(R) is *not* a replacement of $\mathrm{QIC}_P(R)$ in covariate selection.

In a modelling process involving clustered data in the GEE setting, mean function modelling, including covariate selection, can be done using $\mathrm{QIC}_P(R)$ or $\mathrm{QIC}_{HH}(R)$, and not using CIC(R). After the mean function has been modelled adequately, CIC(R) can be used to aid correlation structure modelling with an aim to improve the efficiency of parameter estimation. The importance of proper marginal mean and variance modelling has been stressed in Cui and Qian [2].

It is worth bearing in mind that since the estimation of correlation magnitude depends upon the residuals, when the mean function is misspecified, the residuals are wrongly estimated, and any attempt to estimate the magnitude of correlation will give a wrong and misleading correlation magnitude, even if the working correlation structure is, by chance, the true correlation structure. The story does not end there. Wang and Lin [18] showed that variance function needs to be modelled correctly as well because in the case of variance function misspecification, it is not guaranteed that efficiency will be improved by choosing the correct correlation structure. The role of QIC(R) and CIC(R) in variance function selection constitutes another research area that we will be addressing in another paper.

The correlation matrix $\mathbf{R}_i$ consists of two components: (1) *structure*: the functional form of correlation across each lag within the cluster and (2) *parameter*(s): the parameter(s) that determine the magnitude of correlation across each lag within the cluster. Therefore, in the absence of correct mean function specification, even if the working correlation *structure* is the same as the true correlation *structure*, the *parameter*(s) of the correlation matrix estimated based on a wrong set of residuals will be wrong, leading to an erroneously estimated correlation matrix that cannot yield an optimal covariance.

DOI: 10.1002/sim

The paradigm of first selecting covariates using QIC(R), then selecting the working correlation structure using CIC(R) in GEE modelling is nicely demonstrated in the two examples described by Cui [5] using data from the National Longitudinal Survey of Labor Market Experience [19]. In both examples, GEE models containing different combinations of covariates of interest are fitted under working independence. In both cases, the full models that contain all the covariates of interest have the lowest QIC(R) values.

Conditional on the full mean function, working independence gives the lowest QIC value, while exchangeable and unstructured correlation gives the two lowest 'trace values', i.e. CIC(R) values. We note from the simulation results in Section 3 that QIC(R) has a tendency to choose working independence as the working correlation structure of choice (Table I), and we suspect this may be the case here.

In the first example where the response is the income, there are nine unique number of years indexing the time the data were collected. In the second example where the response is the status of being a union member, there are 12 such unique number indexing the data set. When exchangeable correlation structure is used, only one correlation parameter is being estimated. However, when unstructured correlation is used, example one involves estimation of $(9^2-9)/2=36$ correlation parameters, while example two involves estimation of 66 correlation parameters. Since CIC(R) does not account for penalty in terms of the number of correlation parameters estimated, direct comparison of correlation structures with different number of correlation parameters using CIC(R) may not be advisable especially when the numbers of correlation parameters are so different. This sees a need for further research into developing a more general correlation structure selection criteria that accounts for different number of correlation parameters involved in these competing structures.

ACKNOWLEDGEMENTS

We are grateful to Professor Louise Ryan for interesting discussions on the example, and the reviewers for their constructive comments that greatly improved the article.

REFERENCES

1. Wang Y-G, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalized estimating equations performance. *Biometrika* 2003; **90**:29–41.
2. Cui J, Qian G. Selection of working correlation structure and best model in GEE analyses of longitudinal data. *Communications in Statistics—Simulation and Computation* 2007; **36**:987–996.
3. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**:120–125.
4. Cantoni E, Flemming JM, Ronchetti E. Variable selection for marginal longitudinal generalized linear models. *Biometrics* 2005; **61**:507–514.
5. Cui J. QIC program and model selection in GEE analyses. *The Stata Journal* 2007; **7**:209–220.
6. Kuk AYC. A generalized estimating equation approach to modelling foetal response in developmental toxicity studies when the number of implants is dose dependent. *Applied Statistics* 2003; **52**:51–61.
7. Hin L-Y, Carey V, Wang Y-G. Criteria for working-correlation-structure selection in GEE: assessment via simulation. *American Statistician* 2007; **61**:360–364.
8. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
9. McCullagh P, Nelder J. *Generalized Linear Models*. Chapman & Hall: London, 1989.
10. Hardin J, Hilbe J. *Generalized Estimating Equations*. Chapman & Hall: Florida, 2003.
11. Efron B. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* 1986; **81**:461–470.

DOI: 10.1002/sim

12. Efron B. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 2004; **99**:619–632.
13. *R. A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007.
14. Carey V. *Yags*: *Yet Another GEE Solver*, *R Package Version 4.0-2*, 2004.
15. Venebles W, Ripley B. *Modern Applied Statistics with S*. Springer: New York, 2002.
16. Leisch F, Weingessel A. *Bindata*: *Generation of Artificial Binary Data*, *R Package Version 0.9-12*, 2005.
17. Diggle P, Heagerty P, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, 2002.
18. Wang YG, Lin X. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics* 2005; **61**:413–421.
19. Center for Human Resource Research. *National Longitudinal Survey of Labor Market Experience*, *Young Women 14–26 Years of Age in 1968*. Ohio State University, 1989.

DOI: 10.1002/sim