Research Article

Statistics in Medicine

Received 20 January 2015, Accepted 5 November 2015 Published online 2 December 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6821

A determinant-based criterion for working correlation structure selection in generalized estimating equations

Ajmery Jaman,[a*†] Mahbub A. H. M. Latif,[a] Wasimul Bari[b] and Abdus S. Wahed[c]

In generalized estimating equations (GEE), the correlation between the repeated observations on a subject is specified with a working correlation matrix. Correct specification of the working correlation structure ensures efficient estimators of the regression coefficients. Among the criteria used, in practice, for selecting working correlation structure, Rotnitzky-Jewell, Quasi Information Criterion (QIC) and Correlation Information Criterion (CIC) are based on the fact that if the assumed working correlation structure is correct then the model-based (naive) and the sandwich (robust) covariance estimators of the regression coefficient estimators should be close to each other. The sandwich covariance estimator, used in defining the Rotnitzky-Jewell, QIC and CIC criteria, is biased downward and has a larger variability than the corresponding model-based covariance estimator. Motivated by this fact, a new criterion is proposed in this paper based on the bias-corrected sandwich covariance estimator for selecting an appropriate working correlation structure in GEE. A comparison of the proposed and the competing criteria is shown using simulation studies with correlated binary responses. The results revealed that the proposed criterion generally performs better than the competing criteria. An example of selecting the appropriate working correlation structure has also been shown using the data from Madras Schizophrenia Study.

Keywords: bias-corrected sandwich covariance estimator; correlation information criterion; model-based covariance estimator; Rotnitzky-Jewell criteria

1. Introduction

Longitudinal study is one of the principle research strategies employed in biomedical and social science research [1, 2], where each subject is followed over a period of time, and repeated observations of the response and relevant covariates are recorded. Responses measured on the same subject are usually assumed to be correlated and, therefore, a proper modeling approach is needed that accounts for within subject correlation. The commonly used generalized linear models (GLM) cannot be extended for correlated responses because of intractability of discrete multivariate distributions. Liang and Zeger [3] proposed a non-likelihood-based method, namely, generalized estimating equations (GEE), which provide a regression methodology for the marginal analysis of correlated responses. GEE account for within subject or cluster correlation and similar to GLM, it can be applied for both discrete and continuous responses. Instead of completely specifying the associated multivariate distribution, GEE require to specify only the lower-order moments. A user-defined working correlation matrix is required instead of true correlation matrix to specify the within subject correlations. GEE result in consistent estimators of the regression coefficients and also of their variances under weak assumptions about the true correlation among the observations within a subject [3]. However, the efficiency of the consistent estimators depends on the correct specification of the working correlation structure, and a working correlation structure that closely approximates the true underlying pattern results in better precision [4].

[a]*Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh*
[b]*Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1-000, Bangladesh*
[c]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.*
**Correspondence to: Ajmery Jaman, Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka-1000, Bangladesh.*
[†]*E-mail: ajaman@isrt.ac.bd*

Therefore, developing a criterion to select an appropriate working correlation structure in GEE based on the available data has been an active field of research over the past two decades.

Akaike's information criterion (AIC) [5] is the most widely used model selection criterion, which is based on the likelihood function and can only be applied for likelihood-based method. Pan [6] extended the AIC for non-likelihood-based method, such as GEE, by replacing the likelihood function with the corresponding quasi-likelihood function constructed under the working independence structure. Pan's modified version of the AIC is known as the quasi-information criterion (QIC). The fact that QIC requires estimation of the quasi-likelihood function, which is independent of the assumed correlation structure, substantially weakens the performance of QIC in selecting the appropriate working correlation structure in GEE [7]. Hin and Wang [7] modified the QIC by ignoring the quasi-likelihood term, and the resulting criterion is known as the correlation information criterion (CIC). In comparison with QIC, CIC showed a better performance in selecting the correct correlation structure in GEE [7].

Shults *et al*. [8] studied the Rotnitzky-Jewell (RJ) criteria [9] in selecting the appropriate working correlation structure in GEE. The RJ criteria are defined as functions of the model-based and the sandwich covariance estimators of the regression coefficients, both of which are estimated under the assumed working correlation structure. The motivation behind these criteria is that these two covariance estimators should be similar when the working correlation structure is close to the true structure. Shults *et al*. [8] also compared the RJ criteria with the Shults-Chaganty (SC) criterion [10], which is based on weighted error sum of squares. In a simulation study conducted by Shults *et al*. [8], the RJ criteria outperformed the SC criterion in selecting the true working correlation structure.

All the criteria discussed earlier (except the SC criterion) are defined as functions of the sandwich covariance estimator, which is biased downward [11] and generally has a larger variability than the corresponding model-based estimator [12]. Use of the negatively-biased covariance estimator can result in the hypothesis tests of the regression coefficients that are too liberal and confidence intervals on the regression parameters that are too narrow, specially for smaller samples [13, 14]. Few alternatives of the sandwich covariance estimator are available in the literature that estimate the covariance matrix of regression estimators more precisely by removing the inherent bias in sandwich estimator [11, 12]. In this article, we propose a new criterion for appropriate working correlation structure selection in GEE based on the bias-corrected sandwich covariance estimator proposed by Mancl and DeRouen [11]. The proposed criterion chooses an appropriate correlation structure in GEE by comparing the bias-corrected sandwich estimator with the model-based covariance estimator of the marginal model parameter estimates. The proposed criterion performs better than the RJ criteria because it is influenced by all the elements of a particular matrix common to both the proposed and RJ criteria, whereas the RJ criteria only involve the diagonal elements of that matrix.

In the next section, we give a brief description of the GEE approach to introduce the notation used in this article. The description of the competing and the proposed criteria is given in Section 2.1 and Section 3, respectively. In Section 4, we present results from a simulation study that evaluates the performance of the proposed criterion by comparing it with the other existing criteria (QIC, CIC, and RJ) in terms of the selection of true underlying structure for correlated binary responses in marginal regression models. In Section 5, we compare the existing and the proposed criteria using a real data set from Madras Longitudinal Schizophrenia Study. Finally, a general discussion is provided in Section 6.

2. Generalized estimating equations

Let y_{ij} and x_{ij} be the response and the $p \times 1$ vector of covariates, respectively, at the jth time for the ith subject ($i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, n_i$) in a data set $\mathcal{D}$. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ be the $n_i \times 1$ vector of responses and $x_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})'$ be the $n_i \times p$ matrix of covariates for the subject i. The marginal model for the response y_{ij} requires specifying marginal mean $\mu_{ij} = E(Y_{ij} \mid x_{ij})$ and variance by a generalized linear model [15] as $g(\mu_{ij}) = x_{ij}'\beta$ and $var(Y_{ij}) = \phi v(\mu_{ij})$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of regression coefficients, g is the link function, v is a known variance function, and ϕ is the scale parameter. The covariance matrix of the response is specified as

$$\mathbf{V}_i = \text{cov}(\mathbf{Y}_i) = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}, \tag{1}$$

where $\mathbf{A}_i$ is a $n_i \times n_i$ diagonal matrix with elements $var(Y_{ij}) = \phi v(\mu_{ij})$ as the jth diagonal element, $\mathbf{R}_i(\boldsymbol{\alpha})$ is the correlation matrix among the outcomes measured at different times for the ith subject and α is a

$q-$dimensional vector of unknown parameters that completely specifies within subject correlation. The GEE estimate the vector of regression coefficients β by solving the following estimating equations [3]:

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{S}_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,, \tag{2}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ is the marginal mean vector for the subject i and $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \beta'$. The estimation of β requires specifying the structure of the working correlation matrix $\mathbf{R}_i(\alpha)$. The common choices of working correlation structure include 'exchangeable' or 'compound symmetry' (CS) for which $corr(Y_{ij}, Y_{ik}) = \alpha_0$ for any $j \neq k = 1, \ldots, n_i$ and 'first-order autoregressive' (AR(1)) for which $corr(Y_{ij}, Y_{ik}) = \alpha_0^{|j-k|}$ for any $j, k = 1, \ldots, n_i$. The other choices for working correlation structure includes 'unstructured' for which $corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$ for any $j \neq k$ and 'toeplitz' for which $corr(Y_{ij}, Y_{ik}) = \alpha_d$ for any $j \neq k$ and $d = |j - k|$.

Given the model for marginal mean function μ_{ij} is correct, and when mild regularity conditions hold, the GEE estimators $\hat{\beta}$ asymptotically follow a multivariate normal distribution with mean vector β and covariance matrix [3]

$$V_S = V_M \left[\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{r}_i \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right] V_M, \tag{3}$$

where $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ and $V_M = \left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right)^{-1}$. The matrix V_M is known as the model-based estimator of the covariance of $\hat{\beta}$, and V_S is known as the sandwich covariance estimator, which reduces to V_M if the model specifications for marginal mean and variance are correct.

Prentice and Zhao [16] proposed the method (typically known as the second-order GEE or simply GEE2) of estimating correlation parameter vector α and regression parameter vector β simultaneously that is based on a second set of estimating equations for α. In this study, GEE2 are used for estimating the marginal model parameters because GEE2 are more robust to misidentification of the correlation structure than the usual moment-based estimation of the correlation parameters [17, p. 59]. As mentioned before, the correct choice of the working correlation structure ensures efficient estimators for the marginal model parameters, and a number of criteria have been proposed in the literature for selecting appropriate working correlation structure in GEE setup. In this article, we have considered QIC, CIC, and RJ criteria for comparison because of their relatively good performance and described these criteria in the following sections.

2.1. Existing criteria for correlation structure selection in generalized estimating equations

Quasi-information criterion [6], a popular model selection criterion used in GEE, comprises two terms; one of which is quasi-log-likelihood function. For the given data $\mathscr{D}$ and under the working independence assumption that the observations y_{ij} are independent in $\mathscr{D}$, that is, $\mathbf{R}_i(\alpha) = \mathrm{I}$ for each i, the quasi-log-likelihood function can be expressed as [6]

$$QL(\beta, \phi; \mathbf{I}, \mathscr{D}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} QL(\beta, \phi; (y_{ij}, \mathrm{x}_{ij}))\,. \tag{4}$$

The notation QL is used here to denote the quasi-log-likelihood function instead of Q that was used in Pan's paper [6]. With this definition of the quasi-log-likelihood function, the QIC criterion can be expressed as

$$\mathrm{QIC} = -2\, QL(\beta, \phi; \mathbf{I}, \mathscr{D}) + 2\, trace \left[\left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i \right) \hat{V}_S \right], \tag{5}$$

where the quasi-log-likelihood, $\mathbf{D}_i$, $\mathbf{A}_i$, and $\hat{V}_S$ are to be evaluated at the values of β and ϕ estimated under the assumed working correlation structure. Later on Hin and Wang [7] argued to use the trace term in QIC for choosing appropriate correlation structure for a given GEE model and proposed the CIC

$$\text{CIC} = trace\left[\left(\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{A}_i^{-1} \mathbf{D}_i\right) \hat{V}_S\right], \tag{6}$$

where $\mathbf{D}_i$, $\mathbf{A}_i$, and $\hat{V}_S$ are to be evaluated at the values of β and ϕ obtained under the assumed working correlation structure as evaluated in QIC. In case of both the QIC and CIC criteria, the correlation structure that results in the minimum criterion value is considered as optimal/appropriate.

However, a decade before the proposal of QIC criterion, Rotnitzky and Jewell [9] introduced a set of criteria, which were compared and implemented by Shults *et al.* [8] in their study of comparison between two treatments for major depressive episodes. The RJ criteria compare the model-based (that assumes correct specification) and the sandwich estimators of the covariance matrix of $\hat{\beta}$ under the assumed working correlation structure. If the working correlation structure is close to the true structure, the model-based and the sandwich estimators should be similar, and consequently, both $Q = \hat{V}_M^{-1}\hat{V}_S$ and Q^2 should be close to a $p \times p$ identity matrix. Following this, the three RJ criteria are defined as

$$\begin{aligned} \text{RJ1} &= trace(Q)/p \\ \text{RJ2} &= trace(Q^2)/p \\ \text{DBAR} &= \sum_j (e_j - 1)^2 = \text{RJ2} - 2\text{RJ1} + 1\,, \end{aligned}$$

where e_j are the eigenvalues of Q and p is the dimension of the parameter vector β. With RJ1, the structure corresponding to the minimum value of $|\text{RJ1} - 1|$ is selected; the same rule is applied to RJ2. On the other hand, DBAR criterion chooses the structure corresponding to the minimum absolute value of DBAR.

All the criteria listed earlier involve sandwich covariance estimator $\hat{V}_S$. Because of the negatively biased nature of $\hat{V}_S$, alternative variance covariance estimators have been proposed by Mancl and DeRouen [11] and Kauermann and Carroll [12] by correcting the inherent bias in it. In the next section, a brief description of the bias-corrected sandwich covariance estimator, which was proposed by Mancl and DeRouen [11], is given and based on this estimator a new criterion is proposed for the selection of an appropriate working correlation structure in GEE.

3. Proposed criterion

The sandwich covariance estimator $\hat{V}_S$ depends on the residuals $\hat{\mathbf{r}}_i = y_i - \hat{\mu}_i$, because $cov(\mathbf{Y}_i)$ in V_S is estimated by $\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'$, in practice (see Equation (3)). Using a first-order Taylor series expansion of $\hat{\mathbf{r}}_i$ about β, it can be shown that

$$E\left(\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'\right) \approx (I_{n_i} - H_{ii})cov(\mathbf{Y}_i)(I_{n_i} - H_{ii})' + \sum_{m\neq i} H_{im}cov(\mathbf{Y}_m)H_{im}'\,, \tag{7}$$

where $H_{im} = D_i\left(\sum_{l=1}^{N} \mathbf{D}_l'\mathbf{V}_l^{-1}\mathbf{D}_l\right)^{-1}\mathbf{D}_m'\mathbf{V}_m^{-1}$ and I_{n_i} is an identity matrix of the same dimension as H_{ii}. From Equation (7), it is clear that $\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'$ is not an unbiased estimator for $cov(\mathbf{Y}_i)$. The bias involved in estimating $cov(\mathbf{Y}_i)$ by $\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'$ also introduces some bias in the sandwich estimator $\hat{V}_S$. A statistical analysis based on this biased sandwich estimator may lead to invalid conclusions regarding the marginal model parameters, specially when we have small number of clusters in our data [13, 14].

In order to derive a tractable approximation to the bias, Mancl and DeRouen [11] assumed that the contribution to the bias of the sum in expression (7) is negligible. By definition, the elements of H_{im} are between zero and one, usually close to zero, so it may be reasonable to assume that the summation makes only a small contribution to the bias. Note that the matrix H_{ii} is an expression for the leverage of the ith subject [18]. Now considering that the expected value in Equation (7) can be approximated as

$$E\left(\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'\right) \approx (I_{n_i} - H_{ii})cov(\mathbf{Y}_i)\left(I_{n_i} - H_{ii}\right)'\,, \tag{8}$$

the bias-corrected sandwich covariance estimator becomes [11]

$$\hat{V}_{S_{bc}} = V_M\left[\sum_{i=1}^{N} D_i'V_i^{-1}\left(I_{n_i} - H_{ii}\right)^{-1}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'\left(I_{n_i} - H_{ii}\right)^{-1}V_i^{-1}D_i\right]V_M\,. \tag{9}$$

The RJ1 and RJ2 criteria described in Section 2.1 involve only the diagonal elements of Q and Q^2, respectively, on the basis of the fact that the off-diagonal elements of an identity matrix are zero. But Q may not be an exact identity matrix, and hence, the off-diagonal elements are expected to be close to zero, not exactly equal to zero. Hence, we felt the importance of using a function that involves all the elements of Q for choosing between the working correlation structures for a given GEE model.

There are several optimality criteria available in the literature, such as, (i) D-optimality, which performs optimization based on the generalized variance of parameter estimators, or simply, the determinant function; (ii) A-optimality, which performs optimization based on the total variance of parameter estimators, or equivalently, the trace of the variance-covariance matrix of parameter estimators; and (iii) E-optimality, which performs optimization based on the variance of least well-estimated linear combination of parameters, for example, the maximum eigenvalue of the variance-covariance matrix of parameter estimators [19]. Among these criteria, D-optimality is the most widely investigated and most commonly used criterion, and has an advantage over the other two criteria as it does not depend on scale of variables while the other two criteria depend [19]. Moreover, in a study of comparison of the robustness properties among A, D, and E optimal designs, it has been shown using polynomial regression model that D-optimality criterion provides better design efficiency compared with the other criteria subject to small departures from the true regression model [20]. So, while the RJ criteria used A-optimality method, we focus on D-optimality method and propose to use the following determinant function to discriminate between working correlation structures:

$$\det(Q) = \det\left(\hat{V}_M^{-1}\hat{V}_S\right) = \det\left(\hat{V}_M^{-1}\right)\ \det\left(\hat{V}_S\right) = \det\left(\hat{V}_S\right)/\det\left(\hat{V}_M\right), \tag{10}$$

where $\det(Q)$ denotes the determinant of matrix Q. It is observed that the determinant of sandwich estimator $\hat{V}_S$ tends to be larger than that of model-based estimator $\hat{V}_M$ in most cases. Moreover, $\hat{V}_S$ reaches the minimum under the true correlation structure [7] and as such the ratio of the determinants in Equation (10) also reaches minimum when the working correlation structure is close to the true correlation structure. Hence, if we use the determinant function as mentioned in Equation (10), we should choose the working correlation structure that results in the minimum value of the $\det(Q)$. In addition, we have checked that the use of the bias-corrected sandwich estimator $\hat{V}_{S_{bc}}$ (Equation (9)) improves the performance of RJ criteria in terms of improving the percentage selection of the correct correlation structure in a simulation study. So, we replace the usual sandwich estimator $\hat{V}_S$ by the bias-corrected sandwich estimator $\hat{V}_{S_{bc}}$ in Equation (10) and propose a determinant-based criterion (DBC), which chooses the correlation structure that corresponds to the minimum value of

$$\text{DBC} = det(Q_{bc}), \tag{11}$$

where $Q_{bc} = \hat{V}_M^{-1}\hat{V}_{S_{bc}}$. Because DBC is influenced by all the elements of Q, this criterion is expected to perform better than the RJ criteria in selecting the best correlation structure in GEE setup. The performance of the proposed DBC criterion is evaluated and compared with the existing criteria by a simulation study with different simulation scenarios, which is shown in the following section.

4. Simulation study

One of the objectives of this study is to compare the proposed criterion (DBC) with the existing criteria (RJ1, RJ2, DBAR, QIC, and CIC) in selecting the correct working correlation structure in GEE using simulation studies. In the simulations, response y_{ij} corresponding to jth time point in the ith cluster is generated from Bernoulli(μ_{ij}), where μ_{ij} is specified as a function of covariates x_1 and x_2 as

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij}, \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_i. \tag{12}$$

In the previous model, the covariates x_1 and x_2 are binary, and both are generated from Bernoulli(0.5). The difference between the covariates is that x_1 is cluster-specific (takes same values over all the observations within the same cluster) and x_2 is observation-specific (takes different values for different observations within the same cluster). For all the simulations, responses are generated with the values of regression coefficients fixed at $\beta_0 = -1.6$, $\beta_1 = 0.38$, and $\beta_2 = 0.35$. Depending on the combinations of correlation

structures under which the responses are simulated, the following two cases are considered in the simulation study:

Case A: Independence (Indep), exchangeable (or CS) and AR(1) correlation structures
Case B: Indep, CS, AR(1), toeplitz (Toep) and unstructured (UN) correlation structures.

For a specific subject i, the pair-wise correlation corresponding to observations (j,j') of the working correlation matrix under unstructured correlation structure is defined using the expression $\alpha^{|T_{ij}-T_{ij'}|^{\lambda}}$, where T indicates the time, which reduces to exchangeable and AR(1) structures for $\lambda = 0$ and $\lambda = 1$, respectively. For the simulations, $\lambda = 0.5$ and $T \in \{1, 2, \ldots, n_i\}$ are used. For toeplitz correlation structure, the first $(n_i - 1)$ elements of $\alpha^{|T_{ij}-T_{ij'}|^{\lambda}}$ are used as the correlation parameter vector for the ith cluster. The simulation study is conducted for both balanced and unbalanced cluster situations, which are described in the following two sections.

4.1. Balanced clusters

At first, we compare the existing and the proposed criteria by generating data with balanced clusters, that is, with the same number of observations in each cluster. In this case, we choose $n_i = 5$ for all clusters. The simulation results with balanced clusters are given in Table I and Table II.

Table I presents the percentages of samples (out of 2000 replications) for which different working correlation structures were selected by each of the selection criteria for Case A, where only three correlation structures were considered as candidates. The results are presented for $\alpha = 0.3$ and 0.5, and for sample sizes $N = 50$ and 100. The results show that for Case A, when the true intra-cluster correlation structure (ICS) is independence, the RJ, QIC, and CIC criteria make the correct selection 13–25% of the time for different combinations of α and N. But the DBC criterion makes the correct choice 41–44% of the time in that case. Moreover, the RJ, QIC, and CIC criteria incorrectly select either CS or AR(1) structure more often than the true independence structure. When the true ICS is CS, the correct selections of CS structure are 82–90% for RJ criteria, 59% for QIC criterion and 70% for CIC criterion for low α and small N. But for high α and large N, RJ criteria select the true CS structure almost all the time while QIC and CIC make the correct choice 81% and 93% of the time, respectively. As compared with these criteria, the DBC criterion makes the correct choice exactly 100% of the time when the sample size is large and makes it 98% or 99% when the sample size is small. Further, when the true ICS is AR(1), the RJ criteria select the true structure 36–43% of the time with selection of CS structure for the remaining samples. The QIC criterion correctly chooses the AR(1) structure 62–74% of the time while the CIC criterion makes the correct choice 95% of the time for high intra-cluster correlation and large N, and makes it 77–88% for the other three scenarios. On the other hand, the proposed DBC criterion correctly chooses the AR(1) structure 56–58% of the time, which is greater than the percentage selection by RJ criteria.

Table II shows the percentage selections (out of 2000 replications) of different working correlation structures by each of the criteria for Case B, where all the five correlation structures were considered as candidates. Similar to Case A, the results are presented for $\alpha = 0.3$ and 0.5, and for sample sizes $N = 50$ and 100. For Case B with independent observations, the correct selections of independence structure are 4–15% for RJ criteria, 6% or 7% for QIC criterion, 2% or 3% for CIC criterion, and 38–42% for DBC criterion. In fact, the selection of the other four working correlation structures is mixed for RJ and DBC criteria while both the QIC and CIC criteria mostly select the unstructured correlation structure instead of selecting the correct independence structure. When the true correlation structure is CS, the correct selections of CS structure are 33–41% for RJ criteria, at most 10% for QIC criterion, at most 2% for CIC criterion, and 33–37% for the proposed DBC criterion when α=0.3 and 42–44% when α=0.5. Besides that RJ and DBC criteria incorrectly select either the unstructured or the toeplitz correlation structure while QIC and CIC select the unstructured correlation structure, in most cases. Further, if the true intra-cluster correlation structure is AR(1), correct correlation structure detections are 13–25% of the time for RJ criteria, 13–18% for QIC criterion, at most 5% for CIC criterion, and 39–45% for DBC criterion. The selection of the other competing correlation structures, in this case, is mixed for RJ and DBC criteria while QIC and CIC criteria select the UN structure most of the time. However, if we have data with the toeplitz or the unstructured correlation structure as the true ICS and if we consider the Case B situation, then none of the competing criteria as well as the proposed one performs well in identifying the true correlation structure. Although the CIC criterion makes the correct choice of unstructured correlation structure 89–93% of the time, this is due do the bias of CIC criterion towards unstructured correlation structure as it always chooses unstructured correlation whenever it is included in the candidate structures.

Table I. Percentage selection of working correlation structures by different correlation structure selection criteria out of 2000 independent replications for Case A with n_i fixed at 5 for all subjects.

			Working Correlation Structures					
			$N = 50$			$N = 100$		
True α	True ICS	Criterion	Indep	CS	AR(1)	Indep	CS	AR(1)
$\alpha = 0.3$	Indep	RJ1	**19**	59	22	**18**	60	22
		RJ2	**20**	57	23	**19**	59	22
		DBAR	**14**	64	22	**13**	68	19
		QIC	**22**	38	40	**25**	36	39
		CIC	**17**	40	43	**20**	38	42
		DBC	**44**	31	25	**41**	33	26
	CS	RJ1	0	**82**	18	0	**94**	6
		RJ2	0	**90**	10	0	**97**	3
		DBAR	0	**88**	12	0	**94**	6
		QIC	28	**59**	13	23	**72**	5
		CIC	8	**70**	22	3	**85**	12
		DBC	0	**98**	2	0	**100**	0
	AR(1)	RJ1	4	60	**36**	0	62	**38**
		RJ2	2	60	**38**	0	62	**38**
		DBAR	0	60	**40**	0	58	**42**
		QIC	16	22	**62**	12	19	**69**
		CIC	7	17	**76**	3	10	**87**
		DBC	0	42	**58**	0	43	**57**
$\alpha = 0.5$	Indep	RJ1	**17**	61	22	**18**	60	22
		RJ2	**19**	58	23	**19**	60	21
		DBAR	**16**	61	23	**13**	70	17
		QIC	**24**	38	38	**23**	37	40
		CIC	**19**	40	41	**18**	40	42
		DBC	**44**	29	27	**42**	33	25
	CS	RJ1	0	**88**	12	0	**97**	3
		RJ2	0	**93**	7	0	**98**	2
		DBAR	0	**89**	11	0	**97**	3
		QIC	25	**69**	6	17	**81**	2
		CIC	3	**80**	17	0	**93**	7
		DBC	0	**99**	1	0	**100**	0
	AR(1)	RJ1	0	61	**39**	0	62	**38**
		RJ2	0	60	**40**	0	61	**39**
		DBAR	0	57	**43**	0	61	**39**
		QIC	13	20	**67**	9	17	**74**
		CIC	1	10	**89**	0	4	**96**
		DBC	0	44	**56**	0	44	**56**

For Case A, QIC criterion incorrectly chooses the independence structure 17–28% and 9–16% of the time when the true structures are CS and AR(1), respectively. On the other hand, QIC incorrectly chooses the independence structure 6–13% of the time for different simulation scenarios for Case B. Among the other criteria, CIC criterion is found to incorrectly select the independence structure at most 8% of the time for low α and small N when the true structure is AR(1), while RJ1 and RJ2 make this incorrect choice at most 4% and 2% of the time, respectively. But the proposed DBC criterion never chooses the independence structure when the true ICS is other than independence and it is true for all simulation scenarios. Performances of the competing criteria, namely, RJ, QIC, CIC, and DBC are summarized in Table III for various simulation scenarios. The proposed DBC criterion outperforms the RJ criteria in selecting the correct correlation structure almost in every simulation scenario. Moreover, the proposed criterion also performs better than the QIC and CIC criteria for all the simulation scenarios but two

Table II. Percentage selection of working correlation structures by different correlation structure selection criteria out of 2000 independent replications for Case B with n_i fixed at 5 for all subjects.

			Working Correlation Structures									
			$N = 50$					$N = 100$				
True α	True ICS	Criterion	Indep	CS	AR(1)	Toep	UN	Indep	CS	AR(1)	Toep	UN
$\alpha = 0.3$	Indep	RJ1	**12**	25	15	20	28	**15**	23	16	19	27
		RJ2	**14**	24	16	20	26	**14**	22	18	19	27
		DBAR	**6**	21	7	22	44	**4**	24	7	23	42
		QIC	**6**	8	8	18	60	**6**	7	6	16	65
		CIC	**2**	4	3	9	82	**2**	3	2	8	85
		DBC	**42**	10	23	10	15	**38**	11	23	10	18
	CS	RJ1	0	**33**	13	25	29	0	**33**	5	26	36
		RJ2	0	**36**	7	27	30	0	**35**	2	28	35
		DBAR	0	**34**	7	28	31	0	**34**	4	26	36
		QIC	11	**7**	3	18	61	12	**9**	0	16	63
		CIC	1	**1**	2	4	92	0	1	1	3	95
		DBC	0	**37**	0	31	32	0	**33**	0	28	39
	AR(1)	RJ1	3	28	**22**	22	25	0	32	**24**	20	24
		RJ2	1	30	**23**	22	24	0	32	**25**	20	23
		DBAR	0	32	**14**	20	34	0	37	**13**	19	31
		QIC	8	6	**13**	17	56	6	8	**13**	15	58
		CIC	1	1	**4**	7	87	0	1	**3**	6	90
		DBC	0	22	**45**	15	18	0	22	**44**	15	19
	Toep	RJ1	0	30	25	**21**	24	0	28	23	**23**	26
		RJ2	0	33	19	**22**	26	0	30	17	**25**	28
		DBAR	0	35	11	**20**	34	0	37	9	**21**	33
		QIC	10	7	9	**13**	61	8	9	6	**14**	63
		CIC	1	2	3	**4**	90	0	2	2	**4**	92
		DBC	0	33	16	**22**	29	0	37	5	**25**	33
	UN	RJ1	1	29	23	22	**25**	0	28	23	22	**27**
		RJ2	0	32	18	25	**25**	0	31	17	23	**29**
		DBAR	0	33	11	23	**33**	0	36	9	22	**33**
		QIC	10	8	8	14	**60**	8	8	7	15	**62**
		CIC	1	1	3	5	**90**	0	1	3	3	**93**
		DBC	0	33	13	24	**30**	0	35	6	25	**34**
$\alpha = 0.5$	Indep	RJ1	**14**	23	17	20	26	**13**	23	17	19	28
		RJ2	**14**	22	18	21	25	**14**	23	18	19	26
		DBAR	**5**	22	8	23	42	**5**	25	6	22	42
		QIC	**7**	8	9	17	59	**7**	7	7	15	64
		CIC	**2**	3	3	8	84	**3**	3	2	7	85
		DBC	**39**	10	25	10	16	**38**	11	23	9	19
	CS	RJ1	0	**34**	8	28	30	0	**36**	2	27	35
		RJ2	0	**36**	5	31	28	0	**38**	1	27	34
		DBAR	0	**36**	8	28	28	0	**41**	2	26	31
		QIC	12	**10**	2	20	56	13	**10**	0	19	58
		CIC	0	**1**	2	5	92	0	**2**	0	3	95
		DBC	0	**44**	0	32	24	0	**42**	0	28	30
	AR(1)	RJ1	0	32	**21**	24	23	0	32	**22**	21	25
		RJ2	0	34	**22**	23	21	0	31	**23**	22	24
		DBAR	0	38	**16**	21	25	0	40	**16**	19	25
		QIC	9	7	**18**	19	47	7	8	**18**	18	49
		CIC	0	1	**5**	11	83	0	0	**3**	8	89
		DBC	0	30	**39**	16	15	0	31	**40**	13	16
	Toep	RJ1	0	32	18	**26**	24	0	32	14	**28**	26
		RJ2	0	35	14	**27**	24	0	34	10	**29**	27
		DBAR	0	38	12	**24**	26	0	41	10	**23**	26
		QIC	10	9	6	**19**	56	10	11	2	**22**	55
		CIC	0	2	3	**6**	89	0	1	2	**5**	92
		DBC	0	43	4	**29**	24	0	43	1	**29**	27
	UN	RJ1	0	32	19	25	**24**	0	33	14	27	**26**
		RJ2	0	34	14	26	**26**	0	36	10	28	**26**
		DBAR	0	38	13	22	**27**	0	40	11	24	**25**
		QIC	10	9	5	20	**56**	10	10	3	22	**55**
		CIC	0	2	3	6	**89**	0	1	2	4	**93**
		DBC	0	43	4	29	**24**	0	44	1	29	**26**

Table III. Best criteria among RJ1, RJ2, DBAR, CIC and DBC for various simulation scenarios of balanced cluster situation.

		$N = 50$		$N = 100$	
	True ICS	Case A	Case B	Case A	Case B
$\alpha = 0.3$	Indep	DBC	DBC	DBC	DBC
	CS	DBC	DBC	DBC	RJ2
	AR(1)	CIC	DBC	CIC	DBC
	Toep		RJ2, DBC		RJ2, DBC
	UN		CIC		CIC
$\alpha = 0.5$	Indep	DBC	DBC	DBC	DBC
	CS	DBC	DBC	DBC	DBC
	AR(1)	CIC	DBC	CIC	DBC
	Toep		DBC		RJ2, DBC
	UN		CIC		CIC

exceptions: (i) For Case A, the CIC criterion shows a superior performance in selecting the true AR(1) correlation structure than any other criteria considered. QIC criterion also shows a better performance in this case compared with DBC in selecting the correct AR(1) structure not surprisingly being dominated by the trace term in it; and (ii) For Case B, the CIC criterion is the best for the selection of true unstructured correlation structure, although this is due to the bias of CIC criterion towards unstructured correlation structure, and the same thing happens to QIC as it is closely related to CIC.

4.2. Unbalanced clusters

In practice, we may not always have balanced clusters in the data. For example, observations could be missing after certain time points in some clusters. This could happen when it is no more possible to follow up some subjects up to the end of a longitudinal study or could be due to withdrawal from the study. In such case, we have longitudinal data that have unbalanced clusters. So in this section, we check how performances of the competing criteria get affected in the presence of unbalanced clusters in the data.

In Section 4.1, we have compared the existing and the proposed criteria by generating data having clusters of size five (i.e. $n_i = 5$). Now, we consider that in some clusters, we have observations fewer than five. To generate these data, we randomly selected n_i from the set {3, 4, and 5} for the ith cluster. We have conducted the simulation with all the settings as used for balanced cluster situation. The simulation results are given in Table IV and Table V for Cases A and B, respectively. In both tables, we can see that the DBC criterion has higher detection rates of the correct correlation structures compared to the RJ criteria in almost all the simulation scenarios. Moreover, DBC is also better than QIC and CIC criteria for all the simulations except for true AR(1) structure in Case A and for true unstructured correlation structure in Case B. Hence, the conclusion regarding the performance comparison among the existing and the proposed criteria is same for both balanced and unbalanced cluster situations.

To check how performances of the existing and the proposed criteria change quantitatively in presence of unbalanced clusters in the data, we looked at the differences of percentage selections of correct correlation structures between balanced and unbalanced cluster situations. At first, we compare Table I with Table IV for percentage differences between balanced and unbalanced situations for Case A. When the true structure is independence, the differences are close to zero for all the criteria for different combinations of α and N. This is not surprising because if the observations are independent, then no matter how many observations are there in each cluster, the GEE approach (hence all the criteria) will take the data as if there are no clusters and will give the same result. When the true structure is CS, the detection rate of the correct CS structure reduces 4–11% for RJ criteria, 7–12% for QIC criterion, and 8–14% for CIC criterion in unbalanced cluster situations compared with balanced situations. But for DBC criterion, the reduction rate is 3–5% for $N = 50$ and nearly zero for $N = 100$. When the true structure is AR(1), we observe either no difference or a slight increase of 1–5% in the correct choice of AR(1) structure for RJ criteria and an increase of 2–5% for DBC criterion. On the other hand, we see a reduction up to 6% and 7% for QIC and CIC criterion, respectively.

Further for Case B, there is no notable difference in the percentage selections of correct correlation structures between balanced and unbalanced cluster situations (compare Table II with Table V) for all

Table IV. Percentage selection of working correlation structures by different correlation structure selection criteria out of 2000 independent replications for Case A with n_i randomly drawn from the set {3, 4 and 5} for every ith subject.

			Working Correlation Structures					
			$N = 50$			$N = 100$		
True α	True ICS	Criterion	Indep	CS	AR(1)	Indep	CS	AR(1)
$\alpha = 0.3$	Indep	RJ1	**19**	58	23	**19**	59	22
		RJ2	**20**	56	24	**19**	58	23
		DBAR	**15**	61	24	**13**	66	21
		QIC	**22**	42	36	**23**	39	38
		CIC	**18**	44	38	**17**	42	41
		DBC	**47**	29	24	**42**	32	26
	CS	RJ1	0	**71**	29	0	**84**	16
		RJ2	0	**81**	19	0	**89**	11
		DBAR	0	**81**	19	0	**87**	13
		QIC	31	**52**	17	26	**63**	11
		CIC	16	**60**	24	8	**75**	17
		DBC	0	**93**	7	0	**99**	1
	AR(1)	RJ1	5	59	**36**	1	59	**40**
		RJ2	2	60	**38**	0	59	**41**
		DBAR	1	55	**44**	0	57	**43**
		QIC	24	20	**56**	18	18	**64**
		CIC	12	19	**9**	5	15	**80**
		DBC	0	40	**60**	0	40	**60**
$\alpha = 0.5$	Indep	RJ1	**18**	60	22	**16**	62	22
		RJ2	**20**	58	22	**18**	61	21
		DBAR	**15**	60	25	**14**	65	21
		QIC	**23**	41	36	**24**	39	37
		CIC	**18**	44	38	**19**	43	38
		DBC	**43**	30	27	**43**	32	25
	CS	RJ1	0	**79**	21	0	**91**	9
		RJ2	0	**86**	14	0	**94**	6
		DBAR	0	**81**	19	0	**89**	11
		QIC	30	**57**	13	23	**71**	6
		CIC	12	**66**	22	3	**85**	12
		DBC	0	**96**	4	0	**100**	0
	AR(1)	RJ1	0	59	**41**	0	59	**41**
		RJ2	0	57	**43**	0	57	**43**
		DBAR	0	52	**48**	0	57	**43**
		QIC	17	17	**66**	15	16	**69**
		CIC	4	14	**82**	1	9	**90**
		DBC	0	39	**61**	0	41	**59**

the criteria except QIC and CIC in one situation. When the true correlation structure is unstructured in Case B, the performance of both QIC and CIC criteria worsen greatly in presence of unbalanced clusters in the data.

5. Analysis of Madras longitudinal study data

As an example, Madras Longitudinal Schizophrenia Study data [21] are considered in this paper for comparing the existing and the proposed criteria in selecting the appropriate correlation structure in a GEE setup. The data are collected from 86 subjects on six common schizophrenia symptoms, which were classified into positive symptoms (hallucinations, delusions, and thought disorders) and negative

Table V. Percentage selection of working correlation structures by different correlation structure selection criteria out of 2000 independent replications for Case B with n_i randomly drawn from the set {3, 4 and 5} for every ith subject.

			Working Correlation Structures									
			$N = 50$					$N = 100$				
True α	True ICS	Criterion	Indep	CS	AR(1)	Toep	UN	Indep	CS	AR(1)	Toep	UN
$\alpha = 0.3$	Indep	RJ1	**12**	28	16	21	23	**13**	25	16	19	27
		RJ2	**14**	28	17	20	21	**13**	26	18	19	24
		DBAR	**7**	25	12	22	34	**5**	25	8	22	40
		QIC	**8**	10	11	22	49	**8**	9	9	18	56
		CIC	**3**	5	6	15	71	**3**	4	3	10	80
		DBC	**44**	10	21	11	14	**41**	11	22	9	17
	CS	RJ1	0	**33**	19	23	25	0	**31**	13	27	29
		RJ2	0	**34**	13	27	26	0	**35**	7	29	29
		DBAR	0	**34**	11	25	30	0	**37**	7	26	30
		QIC	13	**12**	4	22	49	12	**14**	2	20	52
		CIC	2	**5**	3	14	76	1	**5**	1	11	82
		DBC	0	**38**	2	29	31	0	**37**	0	28	35
	AR(1)	RJ1	5	30	**26**	20	19	1	32	**24**	20	23
		RJ2	1	30	**26**	23	20	0	32	**25**	20	23
		DBAR	0	30	**17**	25	28	0	35	**15**	20	30
		QIC	11	7	**13**	22	47	9	5	**14**	19	53
		CIC	2	3	**6**	16	73	1	2	**4**	11	82
		DBC	0	20	**46**	16	18	0	19	**45**	14	22
	Toep	RJ1	1	30	25	**22**	22	0	27	24	**24**	25
		RJ2	1	31	23	**24**	21	0	30	20	**25**	25
		DBAR	0	32	13	**26**	29	0	33	12	**23**	32
		QIC	14	9	9	**20**	48	10	9	7	**18**	56
		CIC	2	4	5	**14**	75	0	3	4	**9**	84
		DBC	0	29	19	**26**	26	0	28	10	**29**	33
	UN	RJ1	1	29	26	22	**22**	0	29	24	23	**24**
		RJ2	1	30	24	23	**22**	0	29	20	26	**25**
		DBAR	0	30	15	24	**31**	0	33	13	23	**31**
		QIC	13	8	9	19	**51**	10	8	8	19	**55**
		CIC	1	4	5	13	**77**	0	3	4	9	**84**
		DBC	0	28	20	24	**28**	0	28	11	28	**33**
$\alpha = 0.5$	Indep	RJ1	**13**	27	16	22	22	**14**	24	18	20	24
		RJ2	**14**	27	18	21	20	**16**	23	18	20	23
		DBAR	**7**	23	10	25	35	**5**	24	9	24	38
		QIC	**8**	11	11	23	47	**7**	9	9	19	56
		CIC	**3**	5	5	16	71	**3**	4	4	10	79
		DBC	**42**	11	22	12	13	**39**	12	22	10	17
	CS	RJ1	0	**40**	14	25	21	0	**39**	6	29	26
		RJ2	0	**42**	9	27	22	0	**42**	4	29	25
		DBAR	0	**41**	11	27	21	0	**43**	6	28	23
		QIC	16	**16**	4	21	43	13	**19**	1	21	46
		CIC	2	**7**	5	16	70	0	**8**	1	14	77
		DBC	0	**44**	2	28	26	0	**46**	0	30	24
	AR(1)	RJ1	0	36	**24**	22	18	0	33	**23**	24	20
		RJ2	0	35	**27**	22	16	0	34	**26**	22	18
		DBAR	0	34	**22**	23	21	0	37	**19**	21	23
		QIC	12	7	**20**	24	37	10	7	**20**	21	42
		CIC	1	3	**10**	21	65	0	1	**7**	16	76
		DBC	0	29	**41**	15	15	0	25	**43**	18	14
	Toep	RJ1	0	36	22	**21**	21	0	33	20	**25**	22
		RJ2	0	39	18	**23**	20	0	35	16	**27**	22
		DBAR	0	40	16	**22**	22	0	38	16	**24**	22
		QIC	12	10	10	**23**	45	12	11	5	**24**	48
		CIC	1	4	6	**17**	72	0	3	3	**12**	82
		DBC	0	38	11	**27**	24	0	39	3	**32**	26
	UN	RJ1	0	34	22	24	**20**	0	32	19	26	**23**
		RJ2	0	36	19	25	**20**	0	36	16	26	**22**
		DBAR	0	38	17	23	**22**	0	38	14	24	**24**
		QIC	14	12	9	23	**42**	12	11	6	24	**49**
		CIC	1	6	6	17	**70**	0	2	3	13	**80**
		DBC	0	41	11	27	**21**	0	39	4	31	**26**

symptoms (flat affect, apathy, and withdrawal), and the level of each symptom was measured every month during the first year following the subject's hospitalization for schizophrenia. Following Diggle *et al.* [21], the interest is on the binary outcome 'thought disorders' (0=absence and 1=presence), for which a large fraction of the subjects (about 65%) are symptomatic at the time of hospitalization (at month 0). To evaluate whether the course of recovery differs for different subjects, Diggle *et al.* [21] considered the following marginal logistic regression model for $\mu_{ij} = E(Y_{ij})$:

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1\,\text{Month} + \beta_2\,\text{Age} + \beta_3\,\text{Sex} + \beta_4\,\text{Month}\times\text{Age} + \beta_5\,\text{Month}\times\,\text{Sex}, \tag{13}$$

where $Y_{ij} \in \{0, 1\}$ denotes presence/absence of thought disorder for the *ith* subject at the *jth* month, Month $\in \{0, 1, \ldots, 11\}$ denotes the month since the follow-up starts, Age denotes age-at-onset (1= 20 years or less than 20 years, and 0=greater than 20 years), and Sex denotes gender of the subject (0=male and 1=female). Because the sample size is not large enough to estimate 66 different correlation parameters when the working correlation matrix has unstructured form, the model is fitted under independence, CS and AR(1) correlation structures only.

For logistic regression models, the correlation parameter must satisfy the Prentice constraint ($L_w \leqslant \alpha \leqslant U_w$) to ensure pair-wise joint probabilities of responses to be non-negative, where the lower and upper boundary values L_w and U_w are defined as functions of μ_{ij} [22]. The usefulness of the Prentice constraints to aid in selecting working correlation structure is based on the observation that the boundary values L_w and U_w, which are functions of the consistent estimators of regression coefficients [3], are consistent even if the working correlation structure is misspecified. But the estimators of correlation parameters may not be consistent under misspecified working correlation structure [23], and hence, the boundary values can be violated asymptotically.

Table VI shows the values of competing criteria, estimates of the correlation parameter and the boundary values under different correlation structures, which are obtained by fitting the model (13). It shows that all the competing criteria except QIC and CIC select AR(1) as the most appropriate correlation structure for analyzing Madras data, whereas QIC and CIC select the independence and the CS correlation structure, respectively. Moreover, the estimates of correlation parameter does not lie within the Prentice bounds for CS correlation structure, but it does for AR(1) correlation structure. We can conclude that AR(1) correlation structure is the most appropriate for analyzing Madras data. Table VI also shows the

Table VI. The criterion value, the estimates of intra-cluster correlation, and the regression coefficient estimates (with standard errors in parenthesis) from the analysis of Madras schizophrenia data for different working correlation structures.

	Working correlation structures		
	Indep	CS	AR(1)
Criterion value:			
|RJ1 − 1|	2.006	0.794	**0.021**
|RJ2 − 1|	9.197	3.428	**0.091**
|DBAR|	5.184	1.841	**0.048**
QIC	**953**.374	957.090	955.011
CIC	18.038	**15**.178	18.064
DBC	842.371	16.668	**1.707**
Correlation coefficient:			
$\hat{\alpha}$		0.273	0.635
Bound ($\hat{L}_w$, $\hat{U}_w$)		(−0.014, 0.064)	(−0.017, 0.784)
Regression coefficients:			
Month			−0.233[a] (0.055)
Age			0.619 (0.459)
Sex			−0.130 (0.419)
Month×Age			−0.096 (0.084)
Month×Sex			−0.157[c] (0.088)
Const.			0.542[c] (0.291)

a : p-value < 0.001; b : p-value < 0.050; c : p-value < 0.100

fit of the model (13) under the most appropriate correlation structure AR(1). The decline over time in the proportion of subjects suffering from thought disorder is statistically significant, and the proportion decreases approximately 20% per month among the older men, approximately 30% per month among the older women, approximately 28% per month among the young men, and approximately 40% per month among the young women during the first year of hospitalization. There is no strong indication that women are less likely to be diagnosed with thought disorder and the rate of recovery depends on the age-at-onset. Interaction effect between month and sex was of primary interest because if it differed significantly from zero and had a negative estimated value, this would indicate that the rate of recovery is faster among women. This interaction effect has negative estimated value but not significant at 5% level of significance.

6. Discussion

Selecting an appropriate working correlation structure in GEE is important for obtaining efficient estimates of the marginal regression model parameters. In this article, we have proposed a new criterion DBC for selecting an appropriate working correlation structure in GEE and compared it with the existing criteria using simulation study involving both balanced and unbalanced cluster situations. As noted in the summary of balanced cluster situations given in Table III, for almost all the simulation scenarios, the proposed DBC criterion performs better than the RJ criteria in terms of relatively high percentage selection of the correct correlation structure. Moreover, for clusters of size 50 and low correlation (0.3), the proposed criterion shows a superior performance over all the other criteria in selecting the correct exchangeable correlation structure (as it chooses exchangeable structure almost all the time) when the competing working correlation structure set includes the independence, exchangeable and AR(1) structures. But for clusters of size 100 or high correlation (0.5), the RJ criteria perform equally well as the proposed criterion for true exchangeable structure. However, as for the correct selection of AR(1) structure, the proposed criterion is better than only the RJ criteria for all the situations, while the CIC criterion is proved to be the best. But if we include toeplitz and the unstructured correlation structure in the competing set, the proposed criterion performs even better than the CIC criterion in selecting all the correct correlation structure except the unstructured correlation structure. Even when we have data with unbalanced clusters, DBC outperforms the RJ, QIC, and CIC criteria in the same simulation scenarios as mentioned in case of balanced cluster situations, where DBC proved the best. However, in case of unbalanced cluster situations compared with the balanced situations, we have observed a notable reduction in the percentage selection of correct CS structure for all the criteria except DBC when we choose the appropriate correlation structure from only three competing structures, Indep, CS, and AR(1). For DBC criterion, the reduction is nearly zero with large number of samples reflecting its more robust nature compared with the other criteria when the true underlying pattern is CS.

Note that there are no methods currently available in the literature other than the criteria discussed in this paper to choose a proper correlation structure in GEE. Proposed criterion and the others reviewed in this paper provide objective criteria to choose correlation structures. These criteria do not perform well mostly when true correlation structure is independence (e.g. first and fourth blocks in Table I). The fact that in this scenario, none of the proposed criteria can choose the correct independence structure is not a problem, as longitudinal/clustered data are rarely independent. But notice that when the true structure is something other than independence (e.g. Blocks 2,3, 5, and 6 in Table I), these criteria are useful. When the true correlation structure is CS, DBC provides 98–100% correct selection. When the true correlation structure is AR(1), CIC makes the correct choice 76–96% of the time. Therefore, if DBC does not choose CS as the appropriate correlation structure then we can confidently eliminate it from our consideration, and so on.

In Section 5, we have fitted the marginal model for Madras Schizophrenia Study data using GEE with different working correlation structures, separately; calculated the value of the correlation structure selection criteria (RJ, QIC, CIC, and DBC); and checked for the violation of Prentice constraints on correlation (that must be satisfied in case of correlated binary data for the joint probabilities to be non-negative) for each of the correlation structures considered. Among the selection criteria, only the CIC criterion is found to select the exchangeable correlation structure for which Prentice constraints are violated. But the proposed as well as the RJ criteria select the AR(1) structure that satisfies Prentice constraints and as such AR(1) can be considered as the most appropriate working correlation structure for analyzing Madras Schizophrenia data with GEE setup. However, one point can be noticed here that several of the

criteria including DBC provided the same choice of correlation structure for the Madras Schizophrenia data. But this is not unexpected as it is one real dataset for which we do not know the true correlation structure. In fact in the simulation studies, there are few datasets for which several criteria reached same conclusion (e.g. for true AR(1) correlation structure RJ1, RJ2, DBAR, and DBC reached the same conclusion in 29% samples out of 2000).

Overall, the importance of this research lies in that it proposes a new criterion which is clearly better than the RJ criteria. This study also reveals that even if the CIC criterion shows a superior performance in case of detecting the true AR(1) correlation structure in simulation study, for Madras Schizophrenia data (for which AR(1) structure is reasonably a good choice) CIC chooses the exchangeable structure for which Prentice constraints are violated.

7. Software

All computations are performed using *R* version 2.15.2, with GEE implemented via *geepack* library. Correlated binary response was generated using *binarySimCLF* library which implement the method described by Qaqish [24].

Acknowledgements

We would like to thank the anonymous associate editor and the two anonymous reviewers for their constructive comments that helped us a great deal to improve this paper.

References

1. Goldstein H, Sc B. *The Design and Analysis of Longitudinal Studies: Their Role in the Measurement of Change*. Academic Press: London, 1979.
2. Nesselroade JR, Baltes PB. *Longitudinal Research in the Study of Behavior and Development*. Academic Press: New York, 1979.
3. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**(1):13–22.
4. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation* 1994; **23**(4):939–951.
5. Akaike H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 1973, 267–281.
6. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**(1):120–125.
7. Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 2009; **28**(4):642–658.
8. Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T. A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine* 2009; **28**(18):2338–2355.
9. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990; **77**(3):485–497.
10. Shults J, Chaganty NR. Analysis of serially correlated data using quasi-least squares. *Biometrics* 1998; **54**(4):1622–1630.
11. Mancl LA, DeRouen TA. A covariance estimator for gee with improved small-sample properties. *Biometrics* 2001; **57**(1):126–134.
12. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; **96**(456):1387–1396.
13. Paik MC. Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics-Simulation and Computation* 1988; **17**(4):1155–1171.
14. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with gaussian error. *Statistics in Medicine* 1996; **15**(16):1793–1806.
15. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman and Halls: London, 1989.
16. Prentice R, Zhao L. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 1991; **47**(3):825–839.
17. Shults J, Hilbe JM. *Quasi-Least Squares Regression*. CRC Press: Boca Raton, 2014.
18. Preisser JS, Qaqish BF. Deletion diagnostics for generalised estimating equations. *Biometrika* 1996; **83**(3):551–562.
19. Atkinson A, Donev A, Tobias R. *Optimum Experimental Designs, with SAS*. Oxford University Press: New York, 2007.
20. Wong WK. Comparing robust properties of a, d, e and g-optimal designs. *Computational Statistics & Data Analysis* 1994; **18**(4):441–448.
21. Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data*. Oxford University Press: New York, 2002.
22. Prentice R. Correlated binary regression with covariates specific to each binary observation. *Biometrics* 1988; **44**(4): 1033–1048.
23. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* 1995; **82**(2):407–410.

24. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 2003; **90**(2):455–463.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web-site. *R* codes that can be used for computing the value of the proposed DBC criterion for a given GEE model can be found in the online version of this article at the publisher's website, or can alternatively be downloaded from the first author's web page "http://www.isrt.ac.bd/ajaman".