Journal of Advances in Mathematics and Computer Science

Volume 39, Issue 8, Page 43-56, 2024; Article no.JAMCS.120449

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Parsimonious Selection of a Working Correlation Matrix in Generalized Estimating Equations

Robert Nyamao Nyabwanga [a*], Kepher Makambi [b], Fred Monari [a] and Lewis Keter [a]

[a] *Department of Mathematics and Actuarial Science, Kisii University, Kenya.*

[b] *Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Cancer Center, USA.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/jamcs/2024/v39i81920

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/120449

Received: 25/05/2024
Accepted: 30/07/2024
Published: 13/08/2024

Original Research Article

Abstract

The quasi-likelihood information criteria (QIC) developed based on the Kullback-Leibler cross-entropy principles is famously used in generalized estimating equations modelling to select a working correlation structure that is vital in improving efficiency of estimates. However, many studies have shown that its

**Corresponding author: E-mail: nyamaonyabwanga@gmail.com;*

Cite as: *Nyabwanga, Robert Nyamao, Kepher Makambi, Fred Monari, and Lewis Keter. 2024. "Parsimonious Selection of a Working Correlation Matrix in Generalized Estimating Equations". Journal of Advances in Mathematics and Computer Science 39 (8):43-56. https://doi.org/10.9734/jamcs/2024/v39i81920.*

use favors over-parameterized correlation structures. In this paper, we suggest a modification to the penalty term of the original QIC by adding a weighting factor built using the number of correlation and regression parameters as cost components. This is aimed at improving its selection rates of a parsimonious correlation matrix structure. Using a simulation study, the performance of the modified QIC was established to be better than that of the original QIC, EAIC and EBIC. Further, it was found out that as the number of repeated measures and degree of correlation became larger, the proposed method gained more power in choosing the correct matrix. The new method was illustrated using the data for Mother's Stress and Children's Morbidity study.

Keywords: Generalized estimating equations; weighted euclidean squared distance; working correlation matrix; parsimony; quasi-likelihood information criteria.

2010 Mathematics Subject Classification: 62B10, 62J12, 62P10.

1 Introduction

The method of generalized estimating equations (GEE) [1], that has the advantage of only requiring the estimation of the mean and variance for the response variable and a provisional correlation matrix is commonly used in biomedical, ecological, economic and other statistical research and applications in which the within-subject responses are correlated to obtain GEE estimates that are assumed to be robust to misspecification of the correlation matrix [1, 2]. However, it has been established that the assumption of robustness to mis-specification of working correlation structure of the sandwich variance estimator does not hold in all situations [3]. Hence, the asymptotic relative efficiency of the GEE estimator could be enhanced by the correct specification of the working correlation structure [3, 4, 5, 6].

The quasi-likelihood information criterion (QIC) developed by [7], is frequently used in choosing both covariates and a working correlation matrix in GEE modeling. Simulation studies have however established success rates of less than 50% when QIC is used in choosing a working correlation structure if the selection set is not limited to parsimonious correlation structures only as in [7]. For instance, studies by [8], [9] and [10] established that QIC often selects over-parameterized correlation structures at the expense of parsimonious ones hence fails to identify the correct correlation matrix. This results in the use of a mis-specified correlation structure with eventual less efficient GEE estimates [3]. These views were corroborated by [11] who pointed out that the use of over-parameterized correlation structures results in models with higher mean square error (MSE) values compared to cases where parsimonious structures are used.

In efforts to address the shortcomings of QIC, [12] proposed the correlation information criteria (CIC) that uses only the penalty term of QIC while [13] proposed the replacement of the quasi-likelihood with an empirical-likelihood and proposed the empirical Akaike information criteria (EAIC) and the empirical Bayesian information criteria (EBIC) which they applied successfully for the selection of the correlation structure in GEE. Through simulations they established superior performance of their proposed criteria compared to QIC and CIC.

In this study, we integrate a tuning parameter into the bias correction term of [7]'s QIC so as to penalize simultaneously for both the regression and correlation parameters. This resultant criteria is expected to penalize both the correlation structure that estimates many parameters and that which despite being parsimonious result in a poor-fitting model hence picking out from a selection set the best parsimonious structure. The performance of the new method is established using simulation studies and is compared to that of the original QIC, CIC, EAIC and EBIC.

The rest of the paper is organized as follows: Section 2 provides a review of the GEE method while section 3 presents [7]'s QIC and the proposed method. section 4 presents a simulation study investigating the performance of the proposed method compared to other criteria. In section 5, we apply the proposed method to Mother's Stress and Children's Morbidity data. Section 6 provides conclusions of the study.

2 Generalized Estimating Equations

Let y_{it} be the ith individual response on measurement t ($t = 1, \ldots, m_i$) and $i = 1...n$ such that, the ith subject's data consist of a $m_i \times 1$ response vector $y_i = (y_{i1}, .., y_{im_i})^T$ and a $p \times 1$ covariate vector $X_i = (X_{i1}, .., X_{ip})^T$. The density function of y_{it} takes the form of exponential family whose log-likelihood function is

$$\ell(y_{it}; \theta_{it}, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_{it}\theta_{it} - b(\theta_{it})}{\phi} + c(y_{it}, \phi) \right\},$$

where θ_{it} is the canonical parameter, ϕ is the scale parameter and $a(.)$, $b(.)$ and $c(.)$ are known functions such that $b'(\theta_{it}) = \mu_{it}$ relate to X_{it} through a known link function. That is, $g(\mu_{it}) = X_{it}^T\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\phi b''(\theta_{it}) = Var(y_{it}) = \phi V(\mu_{it})$, where $V(.)$ is the variance function of μ_{it}. As indicated by [14], ϕ is estimated by:

$$\hat{\phi} = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{t=1}^{m_i} \frac{(y_{it} - \mu_{it})^2}{V(\mu_{it})}, \quad N = \sum_{i=1}^{n} m_i \tag{2.1a}$$

and

$$\frac{\partial(\ell(y_{it}; \theta_{it}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \cdot (\frac{\partial \mu_i}{\partial \beta_j}) \tag{2.1b}$$

Definition 2.1. By replacing the log-likelihood in equation (2.1b) by the log-quasi-likelihood, the GEE estimates $\hat{\beta}_G$ are obtained by iteratively solving the following system of equations using Fisher scoring algorithm [1].

$$U(\hat{\beta}, R(\alpha) = \sum_{i=1}^{n} D_i^T V_i (y_i - \mu_i) = 0, \tag{2.2}$$

where $V_i = A_i^{-1/2} R_i(\alpha) A_i^{-1/2}$, $D_i = \partial\mu_i / \partial\beta^T$, $A_i = Diag\{V(\mu_{i1}, \ldots, V(\mu_{im_i})\}$, and $\mu_i = g^{-1}(X_i^T\beta)$. Considering the model-based covariance matrix $I(\hat{\beta}^I|y)|_{p\times p}$ and the sandwich variance estimate $\Sigma(\beta^R)$, then

$$\sqrt{n}(\hat{\beta}_G - \beta) \sim MVN(0, \Sigma(\beta^R)),$$

where

$$\Sigma(\beta^R) = I(\hat{\beta}^I|y)^{-1} J(\hat{\beta}^R|y) I(\hat{\beta}^I|y), \tag{2.3}$$
$$I(\hat{\beta}^I|y) = \frac{1}{n} \sum_{i=1}^{n} D_i^T V_i^{-1} D_i, \quad J(\hat{\beta}^R|y) = \frac{1}{n} \sum_{i=1}^{n} D_i^T V_i^{-1} (y_i - \mu_i)(y_i - \mu_i)^T D_i$$

Remark 2.1. As n becomes larger, $\sum_{i=1}^{n}(y_i - \mu_i)(y_i - \mu_i)^T/n \to Cov(Y_i)$, and if $R_i(\alpha)$ is correctly specified, $V_i = Cov(Y_i)$ [4] such that

$$J(\hat{\beta}^R|y) - I(\beta^I|y) \xrightarrow{p} 0, \qquad \{I(\hat{\beta}^I|y) - I(\beta^I|y)\} \xrightarrow{p} 0$$

and

$$\{J(\hat{\beta}^R|y)\}\{I(\beta^I|y)\}^{-1} = I_{p\times p} \quad \Rightarrow \quad \Sigma(\hat{\beta}^R) \to I(\hat{\beta}^I|y)^{-1}$$

where I is an identity matrix. This implies that the lower bound on the variance of the unbiased GEE estimators is attained. [6] further established that, if $g^{(-1)}(X_{it}^T\beta)$, $V(\mu_{it})$ and $g(\cdot)$ are correctly specified, $J(\hat{\beta}^R|y)$ in equation (2.3) overrides the impact of using a mis-specified correlation matrix and results to consistent regression estimates for large n. However, a mis-specified structure yields inconsistent correlation parameter estimate ($\hat{\alpha}$) which in turn affects the efficiency of the GEE estimates.

3 QIC and the Proposed Penalty Adjusted QIC

3.1 Quasi-likelihood information criteria

[7] considered the extension of AIC to the GEE framework and came up with the quasi-likelihood information criteria QIC^R defined as

$$QIC^R = -2Q(\hat{\beta}^R|y, I, (\mathbf{Y}, \mathbf{X})) + 2tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\}, \tag{3.1}$$

where β^R is the GEE estimator obtained using a working correlation structure $R(\alpha)$. The optimal correlation matrix is that which minimizes the QIC value [15]

Remark 3.1. The quasi-likelihood term $-2Q(\hat{\beta}^R|y; I, (\mathbf{Y}, \mathbf{X}))$ [16], is evaluated at $\beta = \beta(\hat{R})$ under independence correlation assumption.

Remark 3.2. The quantity $2tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\}$ is the bias correction for over-complexity whereby $\Sigma(\hat{\beta}^R)$ contains information on the postulated correlation structure.

Remark 3.3. The log-quasi-likelihood function and $I(\hat{\beta}^I|y)$ are estimated using the independent working correlation thus makes QIC^R efficient when the correlation parameters estimated are fewer [17]. However, when the correlation matrix estimates many parameters just like in the case of the unstructured correlation in which $m(m-1)/2$ correlation parameters are estimated, the corresponding traces that measure model complexity tend to be under-estimated. Hence, there is a higher likelihood of selecting a more complex structure compared to a parsimonious one [8]. This implies that the ability of QIC^R to sufficiently penalize for complexity declines as the number of correlation parameters q approaches $m(m-1)/2$. This explains the assertions by [8] that QIC^R was biased towards selecting the unstructured correlation matrix and caution by [11] on the use of QIC^R in choosing a correlation structure for longitudinal data since it has a higher likelihood of picking the most complex matrix as the best regardless of what the true structure is hence a violation to the Occam's razor principle. This is also in line with recommendations by [18] that estimation of correlation parameters should be penalized when selecting a working structure since the use of complex correlation structures causes variance inflation.

Simpler models capture the underlying structure better hence have superior predictive performance. We therefore propose a modification to the bias correction term of QIC so as mitigate its appetite for complex correlation structures and improve its selection rates of the true parsimonious working correlation structure.

3.2 The proposed penalty adjusted quasi-likelihood information criteria

We denote the penalty adjusted QIC as $QIC_{m2}(R)$ and write the new method as:

$$QIC_{m2}(R) = -2Q(\hat{\beta}^R|y; I, (\mathbf{Y}, \mathbf{X}) + 2\lambda tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\}, \tag{3.2}$$

where $\lambda = f(p, q)$ is the tuning parameter based on the weighted Euclidean distance from the origin using p and q as cost components. p is the number of regression parameters while q is the number of correlation parameters.

Definition 3.1. Let $R = \{R_i\}$ $i = 1 \ldots S$ be the set of S working correlation structures whose set of correlation parameters $q_i = \{0, 1, ..., 0.5m(m-1)\}$. Further, let $M = \{M_j\}$, $j = 1...p$ be the set of regression models that are a subset of the full model with p regressors. The weighted Euclidean distance from the origin of (p, q) is:

$$d\,(p, q) = \sqrt{\{wp^2 + (1-w)q^2\}}, w = [0, 1]$$

Definition 3.2. Considering the transformations:

$$\begin{cases} p^* = \frac{k}{p}, & M_k \subset M_p, k = 1..p-1 : 0 \le p^* \le 1 \\ q^* = \frac{q_i}{q_{\max}}, & q_{\max} = 0.5m(m-1) : 0 \le q^* \le 1 \end{cases}$$

then;

$$d\left(p^*, q^*\right) = \sqrt{\{wp^{*2} + (1-w)q^{*2}\}}, \quad w = [0, 1] \tag{3.3}$$

To obtain λ, we first consider the penalization adopted by [19] in which the penalty term of the original QIC is multiplied by $2p$ and then set $w = 0$ for Equation (3.3) so that

$$\begin{aligned} 2\lambda tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\} &= 4ptr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\} + \tfrac{2q}{m(m-1}tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\} \\ &= 2\left\{2p + \tfrac{q}{m(m-1}\right\} tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\} \end{aligned} \tag{3.4}$$

Hence,

$$\lambda = 2p + \frac{q}{m(m-1)}, \quad p > 0, q \ge 0, m > 1 \tag{3.5}$$

The proposed method is given by

$$QIC_{m2}(R) = -2Q(\hat{\beta}^R|y; I, (\mathbf{Y}, \mathbf{X})) + 2\left\{2p + \frac{q}{m(m-1)}\right\} tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\}, \tag{3.6}$$

where $I(\hat{\beta}^I|y) = \sum_{i=1}^{n} D_i^T A_i^{-1} D_i$ such that

$$\Sigma(\hat{\beta}^R) = \{I(\hat{\beta}^I|y)\}^{-1} \left\{\frac{1}{n}\sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i\right\} \{I(\hat{\beta}^I|y)\}^{-1}$$

Remark 3.4. If the working correlation matrix is independence, then the $QIC_{m2}(R)$ above reduces to

$$QIC_{m2}(R) = -2Q(\hat{\beta}^R|y; I, (\mathbf{Y}, \mathbf{X})) + 4p\, tr\{I(\hat{\beta}^I|y)\Sigma(\hat{\beta}^R)\} \tag{3.7}$$

Remark 3.5. For a correctly specified model with p covariates, $QIC_{m2}(R)$ has a stronger penalty term than $QIC^{(R)}$ since $\lambda > 1$; hence, a smaller probability of selecting over-parameterized correlation structures. As m increases, $\frac{q}{m(m-1)} \to 0$. Therefore, $QIC_{m2}(R)$ performs better in selecting the correct parsimonious correlation matrix.

4 Simulation Study

4.1 Simulation design

We considered the same model used by [7] and [3] in which the binary response vector $y_i = (y_{i1}, \ldots, y_{it})^T$, $i = 1, \ldots, n$ and $t = 1, \ldots, m$ was assumed to be Bernoulli while the covariates $x_{1it} \sim Bernoulli(0.5)$ and $x_{2it} = t - 1$ hence,

$$\begin{aligned} logit\{E(y_i|x_{1it}, x_{2it})\} &= \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} \\ & \quad for \quad i = 1...n, \quad t = 1...m \\ & \quad \beta_0 = 0.25 = -\beta_1 = -\beta_2 \end{aligned} \tag{4.1}$$

The simulation sample sizes and measurements per subject considered were $n = \{50, 100, 200, 500\}$ and $m = \{3, 5\}$ respectively. Low and medium correlation levels of 0.2 and 0.5, respectively, were considered for the AR(1) and compound symmetry structures [20]. This was useful in determining the effect of the level of correlation on the selection probability of a true correlation structure. We considered the correlation structures below:

i. Independent: $R_0 = Diag(1, \ldots, 1)$.

ii. Compound Symmetry: $R_0 = (1-\alpha)I_m + \alpha\Lambda_m$ for $m = \{3, 5\}$ and $\alpha = \{0.2, 0.5\}$; where I_m is a $m \times m$ identity matrix, and Λ_m is a $m \times m$ matrix of 1's.

iii. AR(1): $R_0 = \{(\alpha)^{|j-k|}\}_{1\le j,k\le m}$ for $m = \{3, 5\}$ and $\alpha = \{0.2, 0.5\}$.

iv. Toeplitz: For $m = 3$ and $m = 5$, we respectively use

$$R_0 = \begin{pmatrix} 1 & 0.50 & 0.35 \\ 0.50 & 1 & 0.50 \\ 0.35 & 0.50 & 1 \end{pmatrix} \quad \text{and} \quad R_0 = \begin{pmatrix} 1 & 0.50 & 0.35 & 0.30 & 0.25 \\ 0.50 & 1 & 0.50 & 0.35 & 0.30 \\ 0.35 & 0.50 & 1 & 0.50 & 0.35 \\ 0.30 & 0.35 & 0.50 & 1 & 0.50 \\ 0.25 & 0.30 & 0.35 & 0.50 & 1 \end{pmatrix}$$

[3] and [7] considered AR(1), compound symmetry and independent correlation matrices in their simulation. The inclusion of Toeplitz matrix allows for the assessment of the ability of our criteria to impose penalty on the number of correlation parameters estimated. Simulations based on 1,000 replications were used to establish the performance of $QIC_{m2}(R)$ compared to EAIC and EBIC [13], CIC [12] and the original QIC [7].

Motivated by [21], we considered the Selection probability of the true structure $\frac{f^{R_{0i}}}{1000}$ such that the Probability of selecting a mispecified structure is $1 - \{\frac{f^{R_{0i}}}{1000}\}$ and mean squared error of prediction (MSEP)

$$MSEP = \frac{1}{1,000} \sum_{K=1}^{1,000} \sum_{i=1}^{n} (\hat{\mu}_{i,R_*}^{(K)} - \mu_i) V_i^{-1} (\hat{\mu}_{i,R_*}^{(K)} - \mu_i)^T, \tag{4.2}$$

in evaluating the performance of the selection criteria. $f^{R_{0i}}$ is the frequency of selection of the i^{th} true correlation structure R_0 and $\hat{\mu}_{i,R_*}^{(K)}$ is the estimator of $\mu_i = g^{-1}(X_i^T\beta)$ under the correlation structure selected by each criteria.

4.2 Simulation results

For independent correlation structure the proposed criteria $QIC_{m2}(R)$ selects the true independent correlation structure with probabilities of 0.921, 0.991, 1.000 and 1.000 for samples of 50, 100, 200 and 500, respectively (Table 1). This implies that $QIC_{m2}(R)$ selected the true independent structure with a probability of 1 for larger sample sizes. The selection probabilities of EAIC and EBIC were above 0.700 and increased with the the increase in sample size to 0.936 and 0.999, respectively, for sample sizes of 500. The other criteria QIC and CIC instead choose the Toeplitz structure which estimates $'m-1'$ correlation parameters. Their selection probabilities were however less than 50%. For the sample sizes considered, $QIC_{m2}(R)$ outperformed both EAIC and EBIC in choosing the true independent structure. We further observed that the estimation accuracy of the GEE model with the independent structure which is the true structure is better than that of the GEE model under the Toeplitz structure preferred by QIC and CIC.

Table 1.True independent structure selection frequency out of 1,000 replications and corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	**726**	112	108	54	1.116	**727**	124	106	13	1.114
	QIC	**196**	180	205	419	1.127	**166**	337	299	198	1.127
	$QIC_{m2}(R)$	**921**	30	42	7	1.109	**952**	12	27	9	1.011
	CIC	**248**	194	231	327	1.123	**197**	197	126	380	1.126
	EBIC	**893**	50	49	8	1.108	**959**	9	32	0	1.102
100	EAIC	**712**	117	115	60	1.100	**882**	30	75	13	1.100
	QIC	**187**	204	186	423	1.123	**175**	329	306	190	1.126
	$QIC_{m2}(R)$	**991**	3	5	1	1.004	**987**	2	11	0	1.001
	CIC	**240**	184	194	382	1.121	**152**	225	230	393	1.122
	EBIC	**940**	30	25	5	1.009	**980**	3	17	0	1.003
200	EAIC	**711**	121	114	54	.994	**877**	38	80	5	.990
	QIC	**172**	204	198	426	1.123	**170**	321	305	355	1.125
	$QIC_{m2}(R)$	**1000**	0	0	0	.903	**1000**	0	0	0	.900
	CIC	**252**	194	200	354	1.117	**200**	240	189	371	1.121
	EBIC	**951**	26	19	0	.984	**992**	9	32	0	.935
500	EAIC	**936**	33	31	0	.983	**892**	43	56	9	.980
	QIC	**236**	201	224	339	1.122	**177**	298	199	326	1.125
	$QIC_{m2}(R)$	**1000**	0	0	0	.823	**1000**	0	0	0	.808
	CIC	**253**	183	226	338	1.119	**208**	208	171	413	1.120
	EBIC	**999**	0	1	0	.823	**1000**	0	0	0	.808

When the true working correlation structure was compound symmetry with a weak correlation of $\alpha = 0.2$ which is in the lower third, [20] and fewer number of measurements per subjects $m = 3$, simulation results in Table 2 show that $QIC_{m2}(R)$ failed to select the correct structure and instead prefered the independent structure. The results buttress findings by [22] that when the correlation between responses is weak, $\hat{\beta}$ obtained using the independence structure is more efficient. Likewise, as observed by [23], under this circumstance, the robust estimates for working independence and compound symmetry correlation will both be correct. This is true since the difference in MSEP of models under the two structures was marginal. Both EAIC and EBIC outperformed $QIC_{m2}(R)$ in this setting and selected the true compound symmetry structure with higher probabilities.

However, increasing the number of measurements per subject to five, resulted to $QIC_{m2}(R)$ outperforming all the other criteria in selecting the true structure. This can be attributed to assertions by [23] that increasing the number of measurements per subject makes the true correlation structure distinct hence easily identifiable by a selection criteria.

When the degree of correlation is increased to 0.5 (in the upper third, [20]), $QIC_{m2}(R)$ selects the compound symmetry structure with probabilities of 0.732, 0.871, 0.965 and 0.987 for the respective sample sizes of 50, 100, 200 and 500 (Table 3). These are comparable to EAIC and EBIC whose respective probabilities are 0.750, 0.818, 0.851 and 0.861 for EAIC and 0.835, 0.923, 0.977 and 0.989 for EBIC. QIC and CIC still selected the Toeplitz structure with higher probabilities rather than the true structure regardless of the increase in α. Just like EAIC and EBIC, the probability of $QIC_{m2}(R)$ selecting the true compound symmetry structure asymptotically approaches one as $n \to \infty$ and it was more than twice that of the original QIC. The results indicate that by increasing α or m, $QIC_{m2}(R)$ overcomes its poor performance in selecting the compound symmetry structure when α is weak and measurements per subject (m) are small. Selection of the true compound symmetry structure equally minimized the MSEP values.

Table 2. Compound symmetry structure ($\alpha = 0.2$) selection Frequency from 1,000 independent replications and corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	156	**538**	209	97	1.118	189	**536**	180	95	1.114
	QIC	165	**235**	189	391	1.122	186	**292**	180	397	1.121
	$QIC_{m2}(R)$	609	**235**	120	32	1.120	212	**666**	113	9	1.074
	CIC	274	**189**	239	307	1.127	223	**398**	158	256	1.126
	EBIC	332	**469**	169	30	1.120	277	**512**	165	146	1.120
100	EAIC	37	**675**	188	100	1.062	42	**613**	179	166	1.063
	QIC	158	**302**	156	389	1.121	153	**351**	84	412	1.121
	$QIC_{m2}(R)$	712	**219**	65	4	1.105	116	**807**	74	3	1.001
	CIC	264	**190**	233	313	1.126	223	**398**	156	233	1.139
	EBIC	131	**676**	168	25	1.062	137	**660**	171	22	1.066
200	EAIC	1	**784**	130	85	1.070	0	**702**	94	204	1.055
	QIC	116	**341**	102	441	1.122	162	**447**	66	325	1.122
	$QIC_{m2}(R)$	758	**202**	40	0	1.101	52	**920**	28	0	1.000
	CIC	210	**246**	241	303	1.122	233	**415**	155	197	1.122
	EBIC	11	**844**	131	14	1.049	14	**857**	112	17	1.051
500	EAIC	0	**852**	11	137	1.027	0	**704**	4	292	1.098
	QIC	121	**329**	80	470	1.107	129	**489**	64	318	1.108
	$QIC_{m2}(R)$	865	**129**	6	0	1.108	8	**988**	4	0	1.000
	CIC	131	**283**	214	372	1.107	158	**549**	147	146	1.108
	EBIC	0	**970**	27	3	1.011	0	**948**	39	13	1.081

Table 3. Frequency of selection of true compound symmetry structure ($\alpha = 0.5$) from 1,000 independent replications and corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	0	**750**	103	147	1.018	9	**828**	111	52	1.018
	QIC	157	**732**	188	51	1.047	0	**861**	139	0	1.003
	CIC	185	**186**	327	302	1.215	137	**343**	135	385	1.166
	EBIC	0	**835**	104	51	1.051	0	**859**	141	0	1.010
100	EAIC	0	**818**	36	146	1.018	0	**890**	85	25	1.000
	QIC	131	**367**	92	410	1.231	134	**444**	195	227	1.179
	$QIC_{m2}(R)$	12	**871**	102	15	1.000	0	**951**	49	0	.900
	CIC	128	**231**	295	346	1.271	195	**491**	157	157	1.155
	EBIC	0	**923**	50	27	1.006	0	**914**	86	0	1.000
200	EAIC	0	**851**	1	148	1.009	0	**971**	0	29	.998
	QIC	124	**379**	78	419	1.231	160	**506**	123	211	1.201
	$QIC_{m2}(R)$	4	**965**	29	2	.990	0	**987**	13	0	.988
	CIC	72	**325**	211	392	1.219	98	**580**	186	136	1.150
	EBIC	0	**977**	6	17	1.000	0	**1000**	0	0	.918
500	EAIC	0	**816**	0	184	1.011	0	**990**	0	10	.975
	QIC	147	**373**	73	407	1.218	123	**576**	10	299	1.153
	$QIC_{m2}(R)$	4	**987**	9	0	.977	0	**1000**	0	0	.891
	CIC	10	**410**	154	426	1.216	29	**655**	150	166	1.110
	EBIC	0	**989**	0	11	.978	0	**1000**	0	0	.890

When the true correlation structure was $AR(1)$, simulation results in Tables 4 and 5 show that for a weak degree of correlation of $\alpha = 0.2$ with a smaller number of measurements per subject $m = 3$, $QIC_{m2}(R)$ failed to select the true structure and instead preferred the independent structure for the data and its selection probabilities for the true AR(1) structure decreased with increase in the sample size. This selection, resulted to a model with lower predictive performance compared to when the true AR(1) structure was used. On the other hand, EAIC and EBIC performed better than the other criteria and their selection probabilities increased with the sample size. QIC and CIC selected the Toeplitz structure rather than the true structure and the resultant model had the lowest predictive performance. The results further showed that, for the same weak correlation, the performance of our proposed criteria in selecting the true AR(1) structure was superior to that of EAIC and EBIC when m is increased to 5. Also, by increasing α to 0.5, the performance of $QIC_{m2}(R)$ was comparable to or exceeded that of EAIC and EBIC. For sufficiently large n and m, $QIC_{m2}(R)$ selected the true AR(1) structure with a probability of one.

When the true correlation structure was Toeplitz, simulation results in Table 6, show that $QIC_{m2}(R)$ fails to select the Toeplitz structure even with the increase in n or m. Instead it preferred a parsimonious AR(1) structure. Likewise, for sample sizes of 50 and 100 both EAIC and EBIC selected the AR(1) structure instead of the Toeplitz structure. However, for sample sizes greater than 200 the consistency of EAIC and EBIC in selecting an over-parameterized structure starts to set in. The results justifies the study objective in which we sought to improve on the penalty term of QIC to aid in the selection of parsimonious correlation structures in GEE.

Table 4. AR(1) ($\alpha = 0.2$) structure Selection Frequency from 1,000 independent replications and corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	237	210	**465**	88	1.109	282	70	**464**	184	1.109
	QIC	152	195	**247**	406	1.264	156	176	**334**	334	1.205
	$QIC_{m2}(R)$	669	69	**244**	18	1.118	237	104	**642**	17	1.103
	CIC	271	192	**214**	323	1.222	142	107	**329**	422	1.200
	EBIC	455	147	**373**	25	1.199	399	34	**518**	49	1.117
100	EAIC	68	1186	**627**	119	1.067	0	11	**705**	284	1.029
	QIC	140	171	**297**	392	1.219	61	154	**470**	315	1.215
	$QIC_{m2}(R)$	726	39	**234**	1	1.109	2	21	**975**	3	1.002
	CIC	247	186	**220**	347	1.220	48	102	**584**	266	1.166
	EBIC	236	152	**589**	23	1.167	0	21	**914**	65	1.009
200	EAIC	1	136	**854**	9	1.046	26	63	**687**	204	1.068
	QIC	106	188	**311**	395	1.201	91	140	**486**	283	1.200
	$QIC_{m2}(R)$	808	13	**179**	0	1.109	99	46	**845**	0	1.000
	CIC	190	214	**279**	317	1.210	92	105	**532**	261	1.161
	EBIC	32	130	**838**	0	1.016	62	60	**831**	37	1.018
500	EAIC	0	38	**957**	5	1.001	0	16	**852**	32	1.011
	QIC	92	153	**338**	417	1.204	71	127	**516**	286	1.195
	$QIC_{m2}(R)$	907	0	**93**	0	1.103	13	3	**984**	1	1.001
	CIC	127	176	**296**	401	1.114	53	62	**552**	333	1.106
	EBIC	0	38	**962**	0	1.001	0	26	**984**	0	1.001

5 Application: Mother's Stress and Children's Morbidity Study

We apply our proposed criteria to the Mother's Stress and Children's Morbidity (MSCM) data. The data was considered by [24] and contains 2,004 observations on 13 variables for 167 mothers and children who enrolled in

Table 5. fAR(1) ($\alpha = 0.5$) structure selection frequency from 1,000 independent replications and corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	0	140	**722**	138	1.031	0	42	**700**	258	1.020
	QIC	99	211	**324**	366	1.124	99	130	**351**	420	1.121
	$QIC_{m2}(R)$	29	181	**768**	22	1.031	16	74	**873**	37	1.020
	CIC	190	199	**275**	336	1.124	220	133	**338**	309	1.121
	EBIC	0	141	**789**	70	1.019	0	69	**830**	101	1.010
100	EAIC	0	43	**828**	129	1.017	0	10	**736**	254	1.010
	QIC	109	205	**293**	393	1.120	52	127	**443**	380	1.114
	$QIC_{m2}(R)$	4	119	**876**	1	1.017	0	30	**970**	0	1.001
	CIC	122	243	**304**	331	1.119	150	121	**429**	300	1.113
	EBIC	0	53	**910**	37	1.007	0	20	**940**	40	1.000
200	EAIC	0	0	**835**	165	1.017	0	0	**770**	230	1.010
	QIC	93	177	**317**	413	1.120	70	92	**513**	325	1.111
	$QIC_{m2}(R)$	4	50	**946**	0	1.009	0	0	**1000**	0	.891
	CIC	66	211	**339**	384	1.122	78	99	**513**	310	1.107
	EBIC	0	6	979	15	1.003	0	9	983	8	.992
500	EAIC	0	0	**846**	154	1.000	0	0	**790**	210	1.006
	QIC	105	189	**333**	373	1.195	60	196	**474**	270	1.194
	$QIC_{m2}(R)$	0	20	**980**	0	.910	0	0	**1000**	0	.871
	CIC	13	110	**418**	459	1.181	66	123	**503**	308	1.177
	EBIC	0	0	**985**	15	.990	0	0	**988**	12	.958

Table 6.Toeplitz structure selection frquency and the corresponding MSEP

		m=3					m=5				
n	Citeria	IN	CS	AR(1)	TOEP	MSEP	IN	CS	AR(1)	TOEP	MSEP
50	EAIC	0	435	529	**36**	1.109	0	70	381	**612**	1.109
	QIC	120	289	254	**337**	1.264	143	196	233	**428**	1.205
	$QIC_{m2}(R)$	41	404	549	**6**	1.118	5	190	805	**0**	1.103
	CIC	175	260	337	**228**	1.222	190	143	286	**381**	1.200
	EBIC	1	441	539	**19**	1.199	0	381	619	**0**	1.117
100	EAIC	0	352	549	**99**	1.067	0	288	591	**121**	1.029
	QIC	120	277	211	**392**	1.219	140	254	205	**401**	1.215
	$QIC_{m2}(R)$	11	426	563	**0**	1.109	0	125	875	**0**	1.002
	CIC	160	288	346	**206**	1.220	63	292	312	**333**	1.166
	EBIC	0	393	606	**1**	1.167	0	287	701	**12**	1.009
200	EAIC	0	160	328	**512**	1.046	0	78	570	**352**	1.068
	QIC	127	271	130	**472**	1.201	112	288	109	**491**	1.200
	$QIC_{m2}(R)$	6	461	533	**0**	1.109	0	90	910	**0**	1.000
	CIC	72	302	389	**237**	1.210	90	287	173	**450**	1.161
	EBIC	0	406	586	**0**	1.016	0	278	722	**0**	1.018
500	EAIC	0	1	70	**929**	1.001	0	29	200	**771**	1.011
	QIC	133	260	138	**469**	1.204	90	230	169	**511**	1.195
	$QIC_{m2}(R)$	7	355	638	**0**	1.103	0	120	880	**0**	1.001
	CIC	10	254	319	**417**	1.114	30	231	149	**590**	1.106
	EBIC	0	63	339	**598**	1.001	0	180	790	**30**	1.001

the study. The study variables were mother's stress, children's illness status, mother's marriage status, mother's highest education level, mother's employment status, health status of children at baseline(chlth), health status of mother at baeline(mhlth), children's race, household size(housize), average mother's stress of the 1-16 days (bstress), average children's illness of the 1-16 days(billness) and study time(week) [*See [24] for variable details*]. Data covering the period of day 17 to 28 were considered, as the data for day 1-16 exhibited weak correlation therefore $m = 12$. The data had 0.97% missing values [24] hence the MCAR assumption was applied. We adopted the logit link function and fitted the following model:

$$\begin{aligned}\log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) &= \beta_0 + \beta_1 illness_{it} + \beta_2 married_{it} + \beta_5 chlth_{it} + \beta_3 education_{it} \\ &\quad + \beta_4 employed_{it} + \beta_6 mhlth_{it} + \beta_7 race_{it} + \beta_8 csex_{it} + \beta_9 housize_{it} \\ &\quad + \beta_{10} bstress_{it} + \beta_{11} billness_{it} + \beta_{12} week_{it} \qquad (5.1)\end{aligned}$$

where $\mu_{it} = E(Y_{it}|X_{itp})$, Y_{it} is the binary indicator of the presence or absence of mother's stress during the t^{th} visit, $t = 1, \ldots, 12$. and X_{itp} is the p^{th} covariate.

We compared the EAIC, QIC, $QIC_{m2}(R)$, CIC and EBIC values for the four models each under the correlation structures: independence, compound symmetry,$AR(1)$ and Toeplitz. EAIC, $QIC_{m2}(R)$ and EBIC chose the $AR(1)$ working correlation structure for the MSCM data while QIC and CIC chose the Toeplitz structure for the data Table 7. The selection of $AR(1)$ by $QIC_{2m}(R)$ is supported by views by [25] that when the intra-subject measurements are equispaced in time, the correlation between consecutive measurements on a subject will decrease with increase in the distance between measurement times.

Table 7. Working correlation structure for MSCM data

	Working Correlation Structure			
	IN	CS	AR(1)	TOEP
EAIC	266.899	68.619	26.006	46.000
QIC	2047.577	2047.358	2047.336	2047.335
$QIC_{2M}(R)$	2082.434	2081.037	2080.182	2081.240
CIC	1.597×10^{-5}	1.471×10^{-5}	1.437×10^{-5}	1.434×10^{-5}
EBIC	317.474	123.409	80.796	142.936

The estimates of regression parameters, SE of the estimates and p-values under the $AR(1)$ and Toeplitz correlation structures are presented in Table 8.

Table 8. Parameter estimates, standard errors, P-values for the AR(1) and Toeplitz structures.

	AR(1)			TOEPLITZ		
COVARIATE	ESTIMATE	SE	P-value	ESTIMATE	SE	P-vale
Illness	.731	.179	$<$ .0001	.696	.186	$<$ .0001
Married	-.032	.233	.889	-.055	.2326	.925
Education	-.419	.221	.055	.423	.223	.032
Employment	.618	.240	$<$ 0.01	-.617	.244	$<$ .01
Child health status	-.230	.122	.058	-.228	.212	.160
Mother health status	-.200	.118	.087	-.204	.117	.053
Race	.067	.237	.761	.064	.238	.692
Sex	-.022	.211	.911	-.0215	.212	.899
House size	.064	.237	.793	.0793	.241	.736
Stress	3.897	.693	$<$ .0001	3.9375	.6980	$<$ .0001
Billness	.426	.692	.544	.443	.708	.560
Week	-.399	.163	.015	-.406	.164	.013

The results show that estimates of regression coefficients are not similar under the two structures hence will lead to different estimates of $E(Y_{it})$ and 10 out of the 12 covariates have lower Standard Errors under the $AR(1)$ structure compared to only 2 which have lower standard errors under Toeplitz structure. As suggested by [26], the standard errors of GEE estimators are smaller when the appropriate working correlation structure is used. Hence in the spirit of [27], we therefore conclude that the model under the $AR(1)$ correlation structure can predict the mother's stress status more accurately than the model under the Toeplitz structure. Lower SE imply that the estimates under $AR(1)$ will have shorter confidence intervals hence more precise. To further ascertain the predictive performance of the GEE model with the $AR(1)$ working correlation structure chosen by EAIC, $QIC_{m2}(R)$ and EBIC for the MSCM data compared to the GEE model with the Toeplitz structure chosen by CIC and QIC, K-fold cross-validation method($K = \{6, 10, 12\}$) was used to generate the MSE error of prediction for the two models.

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} PE_{-k}(\lambda) \tag{5.2}$$

$$PE_{-k}(\lambda) = \frac{1}{\mid N_{-k} \mid}\sum_{i\in N_{-k}} \frac{1}{m}\sum_{i=1}^{m}(Y_{it} - g^{-1}(X_i^T\beta))^2$$

is the prediction error based on the N_{-k} set of subjects in the training dataset. $|N_{-k}|$ is the cardinality of N_{-k}. Relative Efficiency of the model under the $AR(1)$ correlation matrix relative to the model under the Toeplitz structure was also established.

$$RE = \frac{MSE_{\hat{\beta}}^{TOEP}}{MSE_{\hat{\beta}}^{AR(1)}} \tag{5.3}$$

For all K, the RE values were more than 1. Hence, based on [27], the model under the $AR(1)$ working correlation structure had a higher estimation efficiency of the mother's stress status than the model with the Toeplitz structure. The gain in efficiency established for the AR(1) structure was 2% more than the one for the Toeplitz structure when $K = 6$, 6% more than the one for the Toeplitz structure when $K = 10$ and 26% more than the one for the Toeplitz structure when $K = 12$. [13] also held the view that the use of the correct correlation matrix enhances efficiency of GEE estimators.

6 Conclusions

In the present paper, we proposed a modification to the penalty term of QIC model selection criterion in GEE and came up with a new criteria $QIC_{m2}(R)$. The Performance evaluation and comparison with other criteria was done using simulation studies through which we established that$QIC_{m2}(R)$ often selected the true parsimonious correlation structure and its performance became better when the degree of correlation was strong in which case the performance was regardless of the number of observations taken per subject. In the case of a weak correlation, increasing the number of measurements per subject significantly improved its proportion of selecting the true $AR(1)$ and compound symmetry structures. Furthermore, fitting the GEE model with the correlation structure selected by $QIC_{m2}(R)$ improved the relative efficiency of the GEE estimators which is one of the primary interest in GEE modeling. We therefore recommend for the routine use of $QIC_{m2}(R)$ to select a working correlation structure rather than the original QIC.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Liang K, Zeger S. Longitudinal data analysis using generalized linear models, Biometrika. 1986 emph73,12-22.

[2] Wedderburn RWM. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometri KA. 1974;61:439-447.

[3] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. NJ.John Wiley and Sons; 2004.

[4] Kaurmann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. Journal of America Statistics Association. 2008;96(456):1387-1396.

[5] Sutradhar B, Das K. The accuracy of efficiency of estimating equation approach. Biometrics. 2000;90:29-41.

[6] Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: Implications for GEE performance. Biometrika. 2003;90:29-41.

[7] Pan W. Akaike information criteria in generalized estimating equations. Biometrics. 2001;57:120-125.

[8] Barnett G, Koper N, Annette JD, Schmiegelow V, Manseau M. Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. Methods in Ecology and Evolution. 2010;1:15-24.

[9] Nyabwanga RN, Onyango F, Otumba EO. Consistyency inference property of QIC in selecting the true working correlation structure for generalized estimating equations. American Journal of Theoretical and Applied Statistics. 2019;8;74-84.

[10] Hyu-Joo K, Cavanaugh JE, Tad AD, Fore SA. Model selection for overdispersed data and their application to the characterization of a host-parasite relationship. Environ. Ecol. Stat; 2014.

[11] Gul I, Mahbub AL, Presser J. A prediction criteria for working correlation structure selection in GEE. 2018;Sta bfseries 57ME arXiv:1803.06383v1.

[12] Hin L, Wang. Working Correlation Structure identification in generalized estimating equations. Statistic in Medicine. 2001;28:642–658.

[13] Chen J, Nicole L. Selection of Working correlation structure in Generalized Estimating Equations via Empirical Likelihood. Journal of computational and graphical Statistics. 2012;53:98-109.

[14] McCullagh P, Nelder JA. Generalized linear models. Chapman and Hall London, Second Edition; 1989

[15] Wang Y, Orla M, Maxime TC, Wang Z, Sahir RB, Juliana S, Erica EMM. The perils of quasi-likelihood information criteria. 2015;4:246–254.

[16] Hardin JW, Hilbe JM. Generalized estimating equations. Chapman and Hall New York; 2003.

[17] Wentao G. Bootstrap-adjusted quasi-likelihood information criteria for mixed model selection. PHD Dissertation, Bowling Green State University; 2019.

[18] Westgate PM. Improving the correlation structure selection approach for generalized estimating equations and balanced longitudinal data. Statistics in Medicine. 2014;33:2222–2237.

[19] Deroche CB. Diagnostics and model selection for generalized linear models and generalized estimating equations. Applied Mixed Models in Medicine. Doctoral Dissertation; 2015.
Available:http://scholarcommons.sc.edu/etd/3059. 410

[20] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum; 1988.

[21] Shinpei I. Consistent selection of working correlation structure in GEE analysis based on Stein's loss function. Hiroshima Math. J. 2015;45:91-107.

[22] Zeger SL. analysis of discrete longitudinal data: Commentary. Statistics in Medicine. 1988;7:161-168.

[23] Hin L, Carey J, Wang Y. Criteria for working–correlation–structure selection in GEE: assessment via simulation. The American Statistician. 2007;61(4)
DOI: 10.1198/000313007X245122

[24] Asar J, Ilk YG. mmm: An R package for analyzing multivariate longitudinal data with multivariatemarginal models. Computer Methods and Programs in Biomedicine. 2013;112:649–654.

[25] Shults J, Sun W, Xin T, Hanjoo K, Jay A, Hilbe J, 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.

[26] Kwang MJ. Note on working correlation in the GEE of longitudinal counts data. Communications of the Korean Statistical Society.2011;18(6):751-759.

[27] Qu A, Lindsay BG, Li B. Improving generalized estimating equations using QIF. Biometrika. 2000;70(87)4:823–836.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
https://www.sdiarticle5.com/review-history/120449