Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

The perils of quasi-likelihood information criteria

Yishu Wang[a], Orla Murphy[b], Maxime Turgeon[a,c], ZhuoYu Wang[a], Sahir R. Bhatnagar[a,c], Juliana Schulz[b] and Erica E. M. Moodie[a,*]

Received 15 September 2015; Accepted 20 September 2015

In this paper, we consider some potential pitfalls of the growing use of quasi-likelihood-based information criteria for longitudinal data to select a working correlation structure in a generalized estimating equation framework. In particular, we examine settings where the fully conditional mean does not equal the marginal mean as well as hypothesis testing following selection of the working correlation matrix. Our results suggest that the use of any information criterion for selection of the working correlation matrix is inappropriate when the conditional mean model assumption is violated. We also find that type I error differs from the nominal level in moderate sample sizes following selection of the form of the working correlation but improves as sample size is increased as the selection is then concentrated on a single correlation structure. Our results serve to underline the potential dangers that can arise when using information criteria to select correlation structure in routine data analysis. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: generalized estimating equation; information criterion; model selection; quasi-likelihood

1 Introduction

We consider the analysis of longitudinal data, where – as in many settings – analysts are frequently called upon to make modelling assumptions and statisticians often perform model selection in the course of their work. For the parameterized component of the model, the task has been made simpler through the introduction of a variety of procedures and criteria (e.g. stepwise selection and the Bayesian information criterion, respectively) that aim to automate model selection. Dependence on distributional assumptions can be relaxed through semi-parametric choices, for example, by the use of generalized estimating equations (GEEs) in lieu of a mixed effects model; GEEs are consistent under quite general settings, even when the true correlation structure is not correctly specified by the so-called "working correlation" structure (Liang & Zeger, 1986). In this way, model choice is guided by the data and can therefore be seen as an objective procedure. While alternatives to GEEs for estimation marginal models exist (Qu et al., 2000) with associated selection criteria (Wang & Qu, 2009; Zhou & Qu, 2012; Westgate, 2014), they have not, to date, become common tools of analysis.

The quasi-likelihood information criterion (QIC) (Pan, 2001) and several variants have been proposed for the purpose of model selection in a GEE setting. The use of these criteria has been facilitated by their implementation in commonly used software such as SAS (SAS Institute, Cary, NC, USA) (macros QIC and QICu in PROC GENMOD) and Stata (Cui,

[a]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, 1020 Pine Ave W., Montreal, QC H3A 1A2, Canada
[b]Department of Mathematics and Statistics, McGill University, 805 Sherbrooke St W., Montreal, QC H3A 2K6, Canada
[c]Lady Davis Research Institute, Jewish General Hospital, Montreal, QC H3T 1E2, Canada
*Email: erica.moodie@mcgill.ca

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

2007). The QIC is now routinely taught in some graduate courses in epidemiology, and consequently, it is being frequently adopted for selecting the correlation structure to be used in a GEE. In 2014, a Web of Science analysis indicates that more than 80% of the 111 citations to Pan (2001) are in non-statistical journals. Similarly, in the first 8 months of 2015, the article has been cited more than 100 times, primarily outside the statistical literature in articles that employ the selection approach in applications in oncology (Martin et al., 2015), HIV (Stein et al., 2015) and ecology (Montoya et al., 2015). The theoretical justification for the criterion is not in dispute. However, its routine use is not without peril, as we shall demonstrate.

In this paper, we consider two dangers of the routine use of the QIC and its variants in data analysis. We begin in Section 2 with a review of the QIC and its variants, before turning to a particular form of marginal model in which only the independence structure should be used for estimation in Section 3. We then turn our attention to the type I error that results in inference concerning mean model parameters following selection of the correlation structure in GEE in Section 4. Finally, we consider the use of the quasi-likelihood information criteria in the analysis of the relationship between current vitamin A deficiency and the presence of a respiratory infection in Section .

2 Information criteria based on the quasi-likelihood

Akaike's information criterion (Akaike, 1973) was derived based on the idea of minimizing the Kullback–Leibler distance of the assumed model from the true, data-generating model. Differences in the Akaike's information criterion are informative. In the absence of parametric modelling, a quasi-likelihood could be used in place of a true likelihood in order to compute a similar "score" that can be used for model comparison and selection.

We begin with a brief description of the derivation of information criteria for GEEs. Suppose that data arise from a longitudinal study and are composed of n clusters, with the ith cluster consisting of m_i observations; in what follows, $t\,(t = 1, \ldots, m_i)$ will index observation times within the ith cluster $(i = 1, \ldots, n)$. For cluster i, let the m_i-vector $\boldsymbol{y}_i$ denote the response vector, with corresponding $m_i \times p$ covariate matrix $\boldsymbol{x}_i$, of which the tth row $\mathbf{x}_i^t$ is the value of the p covariates at time t. Let $\boldsymbol{\beta}$ denote a p-vector representing the parameters of interest. We let $\boldsymbol{\mu_i}$ be the m_i-vector of mean response values for the ith cluster, and we fix a link function $g(\cdot)$ that relates the parameters $\boldsymbol{\beta}$ of the linear specification to the mean so that $\mu_{it} = g^{-1}(\mathbf{x}_i^t\boldsymbol{\beta})$ for any $t = 1, \ldots, m_i$. Finally, let $v(\cdot)$ denote a variance function that defines the mean–variance relation and A_i denote the $m_i \times m_i$ diagonal matrix of marginal variances for the ith cluster, with the tth diagonal element given by $\phi v(\mu_{it})$, where ϕ is a dispersion parameter.

In a GEE framework, estimates $\hat{\boldsymbol{\beta}}$ are obtained by solving the following equation:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ is an $m_i \times p$ matrix and $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$ is known as the *model-based* variance–covariance matrix. In this formulation, $\mathbf{R}_i(\alpha)$ is an $m_i \times m_i$ matrix representing the *working* correlation matrix for cluster i. The robust estimated variance of the GEE estimates is obtained through a sandwich estimator as

$$\mathbf{V}_{GEE} = \left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \operatorname{cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} D_i\right) \left(\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1}.$$

The empirical estimate of $\hat{\mathbf{V}}_{GEE}$ may be obtained by replacing $\operatorname{cov}(\mathbf{Y}_i)$ by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T$, and $\boldsymbol{\beta}$, ϕ and α by $\hat{\boldsymbol{\beta}}$, $\hat{\phi}$ and $\hat{\alpha}$, respectively. The GEE mean and variance parameter estimators, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}_{GEE}$, are consistent for the true values provided that the mean model is correctly specified, even when the working correlation structure is misspecified, although poor choice of the working correlation structure can lead to efficiency losses.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

The ISI's Journal for the Rapid Dissemination of Statistics Research

The quasi-likelihood under the independence model assumption, for example, assuming a diagonal working correlation structure, for GEEs is an information criterion derived from the quasi-likelihood for the GEE. The quasi-likelihood (Wedderburn, 1974; McCullagh & Nelder, 1989) is

$$Q(\mu, \phi; y) = \int_y^{\mu} \frac{y - t}{\phi v(t)} dt$$

where $\mu = \mathrm{E}(y)$ and $\mathrm{var}(y) = \phi v(\mu)$. In the longitudinal data setting, we have $\mathrm{E}(\mathbf{y}_{it}) = \mu_{it}$ and $\mathrm{var}(y_{it}) = \phi v(\mu_{it})$. Assuming that the within-cluster observations are independent, the quasi-log-likelihood can be written as

$$Q(\boldsymbol{\beta}, \phi; I, (\mathbf{Y}, \mathbf{X})) = \sum_{i=1}^{n} \sum_{t=1}^{m_i} Q(\boldsymbol{\beta}, \phi; (y_{it}, x_{it})).$$

Under the independence assumption, we have $\mathbf{V}_i = \mathbf{A}_i$, and it is easy to verify that

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \partial Q(\boldsymbol{\beta}, \phi; I, (\mathbf{Y}, \mathbf{X}))/\partial \boldsymbol{\beta}.$$

Thus, the GEE with independent working correlation can be regarded as an empirical quasi-log-likelihood score function. If we let $\boldsymbol{\Omega}_I = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1} \mathbf{D}_i$, then, in a similar manner, it can be verified that

$$\boldsymbol{\Omega}_I = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1} \mathbf{D}_i = -\partial^2 Q(\boldsymbol{\beta}, \phi; I, (\mathbf{Y}, \mathbf{X}))/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T.$$

Thus, in the independence model, $\boldsymbol{\Omega}_I$ can be referred to as an empirical quasi-log-likelihood information matrix.

Based on the derivation of the Akaike's information criterion and the resemblance between the quasi-likelihood and the likelihood, Pan (2001) defined a criterion, termed the quasi-likelihood under the independence model criterion, for GEEs, as

$$QIC_P(R) = -2Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, (\mathbf{Y}, \mathbf{X})) + 2\mathrm{tr}(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{GEE}) \tag{1}$$

where all three entities $Q(\hat{\boldsymbol{\beta}}, \hat{\phi}; I, (\mathbf{Y}, \mathbf{X}))$, $\hat{\boldsymbol{\Omega}}_I$ and $\hat{\mathbf{V}}_{GEE}$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(R)$. Similar to Akaike's information criterion, the QIC is a trade-off between model fit, as measured by the quasi-likelihood, and a penalty for over-complexity, as measured by the trace. The optimal variance–covariance matrix is that which gives the smallest QIC. Variations on Pan's development of $QIC_P(R)$ have also been proposed: Hardin & Hilbe (2007) proposed an alternative QIC, $QIC_{HH}(R)$, which contains the same components as $QIC_P(R)$ but where $\hat{\boldsymbol{\Omega}}_I$ in the $QIC_{HH}(R)$ is evaluated at $\hat{\boldsymbol{\beta}}(I)$ rather than $\hat{\boldsymbol{\beta}}(R)$. Thus, $QIC_P(R)$ and $QIC_{HH}(R)$ are identical when the working correlation is independent. Hin & Wang (2008) noted that, assuming that the GEE estimator $\hat{\boldsymbol{\beta}}(R)$ is consistent, the difference between $QIC_P(R)$ and $QIC_{HH}(R)$ is only $O_p(n^{-1/2})$, which is small for datasets with even only a moderate number of clusters ($n \geq 50$). Hin & Wang (2008) thus proposed another information criterion called the correlation information criterion, which is defined as

$$CIC(R) = \mathrm{tr}(\hat{\boldsymbol{\Omega}}_I \hat{\mathbf{V}}_{GEE}) \tag{2}$$

where $\hat{\boldsymbol{\Omega}}_I$ and $\hat{\mathbf{V}}_{GEE}$ are evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(R)$. From the definition, it can be seen that $CIC(R)$ is equal to half of the penalty component of the criterion $QIC_P(R)$ given by the second term in Equation (1). Similar to the correlation information criterion, the penalty in $QIC_{HH}(R)$ can be used as yet another criterion, denoted T_{2HH}.

Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

3 Mean parameter estimate bias following selection of the correlation structure

3.1. The conditional mean model assumption

Pepe & Anderson (1994) raise an important reminder that the parameters of the (conditional) data-generating model may not always be of primary interest. For example, in a setting where prior covariates affect current outcome but are unlikely to be available to the researcher, the cross-sectional model parameters are of greater scientific relevance.

GEEs provide a modelling framework that allows for marginal analysis of such cross-sectional associations. The unbiasedness of the GEE estimators is a direct consequence of finding the root of an unbiased estimating equation $U(\boldsymbol{\beta})$, which relies on the correct specification of the mean model. Recall that this estimating equation can be written as a sum over subjects, or clusters, as

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = 0.$$

This results in a system of p equations. We can write the kth component of $U_i(\boldsymbol{\beta})$ as follows:

$$U_i^k(\boldsymbol{\beta}) = \left(D_i^k\right)^T V_i^k (Y_{it} - \mu_{it}),$$

where D_i^h (resp. V_i^h) is the hth column of $\mathbf{D}_i$ (resp. $\mathbf{V}_i$). Taking the expectation of U_i^k, we have

$$\mathrm{E}\left[U_i^k(\boldsymbol{\beta})\right] = \mathrm{E}\left\{\mathrm{E}\left[U_i^k(\boldsymbol{\beta}) \mid X_{i1}, \ldots, X_{im_i}\right]\right\} = \mathrm{E}\left\{\sum_{t=1}^{m_i} \left(D_i^k\right)^T V_i^k \mathrm{E}[\,(Y_{it} - \mu_{it}) \mid X_{i1}, \ldots, X_{im_i}]\right\}.$$

That is, we must condition on covariates at all time points, $X_{ij}, j = 1, \ldots, m_i$, in order to move $\left(D_i^k\right)^T V_i^k$ outside of the inner, conditional expectation. It thus follows that the expectation $\mathrm{E}\left[U_i^k(\boldsymbol{\beta})\right]$ will be zero if the following condition holds:

$$\mathrm{E}[\,Y_{it} \mid X_{i1}, \ldots, X_{im_i}] = \mathrm{E}[\,Y_{it} \mid X_{it}]\,. \tag{3}$$

Consistency of the estimator $\hat{\boldsymbol{\beta}}(R)$ is guaranteed if the marginal mean is equal to the fully conditional mean. The only exception to this restriction is when a diagonal correlation structure is used, in which case the GEE approach will yield unbiased estimator of $\boldsymbol{\beta}$ even if condition (3) is not satisfied.

The focus of this section is to assess the performance of the information criteria based on the quasi-likelihood in selecting a correlation structure that will yield unbiased mean parameter estimators. Following Pepe & Anderson (1994), three data-generating processes are considered for Y_{it}, which yield the same marginal model:

$$\mu_{it} = \mathrm{E}[\,Y_{it} \mid X_{it}] = \beta X_{it}. \tag{4}$$

These three data-generating processes are as follows:

Model A

$$Y_{it} = \alpha Y_{i(t-1)} + \beta X_{it} + \epsilon_{it}, \quad t = 1, \ldots, n_i,$$

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

The ISI's Journal for the Rapid Dissemination of Statistics Research

where $Y_{i0} = 0$, (X_{it}, ϵ_{it}) have mean zero and are independent of one another and $Y_{i(t-1)}$. The conditional mean of this model is $\mathrm{E}[Y_{it} \mid X_{i1}, \ldots, X_{im_i}] = \beta \sum_{j \le t} \alpha^{t-j} X_{ij}$.

Model B

$$Y_{it} = \beta X_{it}(\alpha Y_{i(t-1)}) + \epsilon_{it}, \quad t = 1, \ldots, n_i,$$

where $\alpha = 1/\beta, Y_{i0} = 1/\alpha$ and, similarly to the previous model, (X_{it}, ϵ_{it}) are independent of each other and $Y_{i(t-1)}$. In this process, X_{it} has mean one, and ϵ_{it} has mean zero. The conditional mean for this model is $\mathrm{E}[Y_{it} \mid X_{i1}, \ldots, X_{im_i}] = \beta \prod_{j \le t} X_{ij}$.

Model C

$$Y_{it} = \eta_i + \beta X_{it} + \epsilon_{it}, \quad t = 1, \ldots, n_i,$$

where the random intercept η_i has mean zero and is independent of (X_{it}, ϵ_{it}), which are again independent with mean zero. In this case, the conditional mean $\mathrm{E}[Y_{it} \mid X_{i1}, \ldots, X_{im_i}] = \beta X_{it}$.

For models A and B, $\mathrm{E}[Y_{it} \mid X_{i1}, \ldots, X_{im_i}] \neq \mathrm{E}[Y_{it} \mid X_{it}]$, that is, condition (3) is not satisfied; the condition is, however, satisfied in model C. In our simulation study, data are generated 1000 times for $n = 50$ and 100 subjects each with $m_i = 5$ observations. The covariates and errors are generated independently using a normal distribution with their specified means and unit variance. The parameter of interest β is set to 0.5 for all three cases, and for model A, we set $\alpha = 0.5$. In model C, a linear mixed effects model, the random intercepts η_i are generated independently of ϵ_{it} from a standard normal distribution. Models A and C correspond closely to models (7) and (9) in Pepe & Anderson (1994); because of inconsistencies in reporting, we were unable to reproduce their model (8).

For each of the generated datasets $\{(Y_{it}, X_{it}) : i = 1, \ldots, n, t = 1, \ldots, m_i = 5\}$, parameter estimates $\hat{\beta}$ are obtained from the marginal model (4). Five correlation structures are considered: an independence structure, empirical correlations for models A and B, respectively, a first-order autoregressive correlation structure and an exchangeable correlation structure (which is the true correlation structure in model C). The empirical correlations were defined such that the (t,j)th entry of the covariance matrix was $\mathrm{cov}(Y_{it} - X_{it}\hat{\beta}, Y_{ij} - X_{ij}\hat{\beta})$ for model A and $\mathrm{cov}(Y_{it} - X_{it}Y_{i1}\hat{\beta}, Y_{ij} - X_{ij}Y_{i1}\hat{\beta})$ for model B, where $\hat{\beta}$ was estimated using an independence working correlation structure.

3.2. Results

As anticipated, the results in the top panel of Table I demonstrate that an independence working correlation structure yields unbiased estimates for the marginal mean μ_{it} in all three data-generating models, while violation of the key assumption (3) leads to bias in the estimate $\hat{\beta}$ (as in the case for models A and B). For model C, where condition (3) is satisfied, estimates are unbiased for all working correlation matrices, but efficiency is gained when the true (i.e. exchangeable) correlation structure is used.

We next turn our attention to selection of the working correlation structure; see Figure 1. Results for model C are ideal: the exchangeable structure is selected most often, which leads to unbiased and efficient estimation. For models A and B, results are less encouraging: the information criterion of Hardin and Hilbe favours the independence working correlation structure, whereas the other three criteria do not consistently select this structure, instead favouring working correlation structures that lead to biased estimates of $\boldsymbol{\beta}$, as observed in the bottom panel of Table I. That is, if the fully conditional mean model does not equal the marginal model that is being estimated, the use of any of the quasi-likelihood-based selection criteria can lead to biased estimation of the mean model parameters. The $QIC_{HH}(R)$ criterion provided the least biased estimators, while the performance of the mean estimators following selection using $CIC(R)$ or $T_{2HH}(R)$ was substantial. The bias is more acute in the small sample setting – precisely when one might be most tempted to select a working correlation model in the hopes of improving efficiency.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

Table I. Mean and standard deviation of the estimates of β in the marginal mean $\mu_{it} = X\beta$ with $\beta = 0.5$, using generalized estimating equations with various fixed working correlation matrices (top panel) and when the correlation matrix is chosen via different selection criteria (bottom panel).

	$n = 50$			$n = 100$		
	Model A	Model B	Model C	Model A	Model B	Model C
Fixed correlation structure						
I	0.499 (0.073)	0.502 (0.401)	0.499 (0.091)	0.500 (0.049)	0.507 (0.286)	0.501 (0.062)
Emp(A)	0.418 (0.063)	0.317 (0.411)	0.515 (0.120)	0.415 (0.042)	0.297 (0.504)	0.508 (0.063)
Emp(B)	0.428 (0.063)	0.295 (0.460)	0.501 (0.098)	0.425 (0.043)	0.300 (0.199)	0.500 (0.055)
Exch	0.452 (0.067)	0.331 (0.306)	0.501 (0.073)	0.452 (0.045)	0.339 (0.226)	0.500 (0.049)
AR(1)	0.415 (0.060)	0.319 (0.252)	0.500 (0.079)	0.414 (0.041)	0.323 (0.189)	0.499 (0.054)
Criterion to choose correlation structure						
$QIC_P(R)$	0.470 (0.080)	0.312 (0.266)	0.500 (0.079)	0.485 (0.059)	0.298 (0.194)	0.501 (0.054)
$QIC_{HH}(R)$	0.483 (0.079)	0.497 (0.400)	0.500 (0.088)	0.492 (0.055)	0.481 (0.285)	0.501 (0.060)
$CIC(R)$	0.420 (0.063)	0.307 (0.264)	0.501 (0.073)	0.415 (0.044)	0.293 (0.194)	0.500 (0.050)
$T_{2HH}(R)$	0.419 (0.063)	0.306 (0.262)	0.501 (0.073)	0.414 (0.044)	0.292 (0.193)	0.500 (0.050)

I, independence; Emp(A), empirical correlation based on model A; Emp(B), empirical correlation based on model B; Exch, exchangeable correlation; AR(1), first-order autoregressive correlation. Shown are average estimates from 1000 simulated datasets.

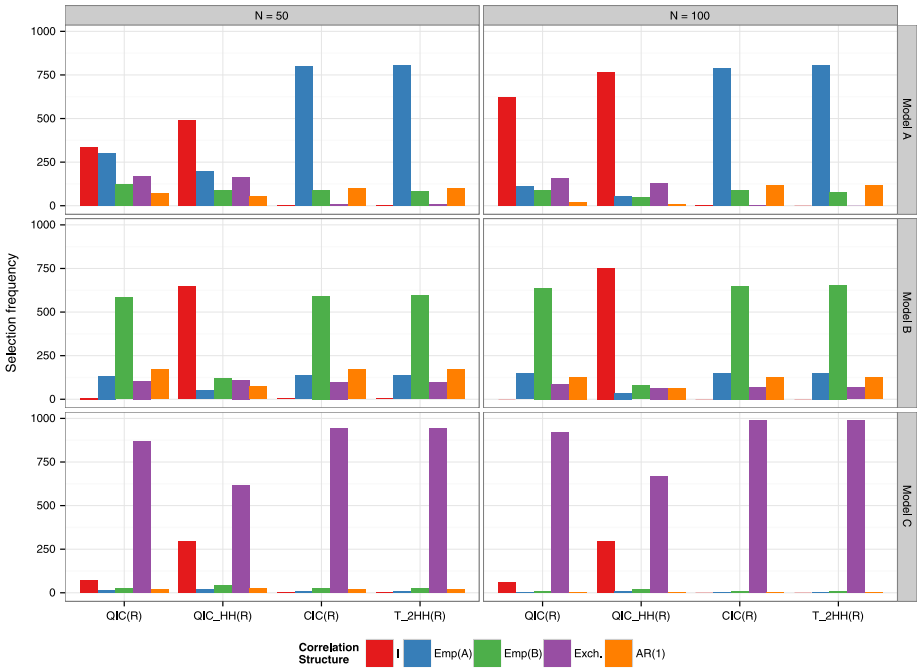

Figure 1. Frequency of working correlation matrix selected for the marginal model by $QIC_P(R)$, QIC_{HH}, $CIC(R)$ and $T_{2HH}(R)$ from 1000 simulated datasets.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

The ISI's Journal for the Rapid Dissemination of Statistics Research

Table II. Type I error using either a fixed working correlation matrix or following selection of the working correlation using QIC_P, QIC_{HH}, $CIC(R)$ and T_{2HH} at a 0.05 significance level from 1000 simulated datasets.

	$n = 25$	$n = 50$	$n = 100$	$n = 200$
Independent	0.089	0.069	0.057	0.043
Exchangeable	0.087	0.074	0.060	0.041
$QIC_P(R)$	0.130	0.102	0.077	0.055
$QIC_{HH}(R)$	0.135	0.111	0.088	0.064
$CIC(R)$	0.131	0.094	0.076	0.046
$T_{2HH}(R)$	0.130	0.093	0.076	0.046

4 Type I error following selection of the correlation structure

It is well known (Hurvich & Tsai, 1990) that type I error rates will not be at the nominal level following variable selection. We now consider the impact of selecting the working correlation structure in a GEE on the type I error rate in testing the null hypothesis of no effect of a covariate on the outcome at a level of significance of 0.05. We consider only model C, where the conditional mean model assumption is satisfied, and both the independent and exchangeable working correlation structures yield unbiased estimators. Unlike in the previous section, we now take $\beta = 0$ and consider the type I error obtained when assuming a fixed working correlation structure (independent or exchangeable) and the error obtained following selection of the working correlation structure using each of the four information criteria described in the previous sections.

We observe in Table II that the type I error differs significantly from the nominal level for $n = 25$ when the working correlation matrix is set to be fixed. When selecting from either the independent, exchangeable, empirical correlation for model A or a first-order autoregressive structure using the four information criteria, the type I error is more than twice the nominal level. When the sample size increases to $n = 200$, all methods yield the nominal error rate, as the selection of the working correlation matrix concentrates on a single structure (Figure 1) so that results are very similar to those obtained assuming a fixed, exchangeable, correlation structure. However, when sample sizes are moderate ($n = 50, 100$), selection of the working correlation structure leads to inflation of the type I error rate relative to estimators based on fixed working correlation matrices.

5 Example analysis: vitamin A and respiratory infection

Diggle et al. (2002) describe a dataset in which the response is the presence or absence of a respiratory illness in 275 children, with up to six measurements collected every 3 months. The children's age and the presence of an ocular condition, xerophthalmia, are recorded. Xerophthalmia serves as an indicator of chronic vitamin A deficiency; it is of interest to determine whether vitamin A deficiency is associated with the presence of a respiratory infection. A regression of respiratory infection on child's age, current xerophthalmia status and lagged xerophthalmia status suggests that there is a strong dependence on lagged xerophthalmia. Therefore, we have evidence to suggest that condition (3) is not satisfied.

We apply each of the selection criteria to choose between the independence, exchangeable and autoregressive correlation structures in two random sub-samples of the data of sizes $n = 50$ and $n = 100$ children as well as in the

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

Table III. Estimate (standard error) of the log odds ratio of the presence of a respiratory illness as a function of xerophthalmia status.

	$\hat{\beta}(I)$	$\hat{\beta}$(Exch)	$\hat{\beta}$(AR(1))
$n = 50$	2.26 (0.86)	2.36 (0.79)	2.25 (0.85)
$n = 100$	0.76 (0.65)	0.64 (0.69)	0.78 (0.63)
$n = 275$	0.72 (0.42)	0.59 (0.45)	0.64 (0.44)

full dataset. All models depend on age (centred) and current xerophthalmia status (Table III). In the full dataset, all information criteria select the independence working correlation structure as optimal. However, at the smaller sample sizes, results are less consistent. At $n = 50$, all criteria suggest that the exchangeable correlation should be retained, whereas when $n = 100$, the autoregressive structure is deemed optimal by all methods. We note, however, that the log odds ratio varies much more dramatically across sample sizes than across working correlation structures: this is explained by the infrequent occurrence of xerophthalmia, which is present at fewer than 5% of all visits.

To assess the stability of the selection at different sample sizes, we repeated the sub-sampling procedure 500 times and computed the information criteria (generating a new sub-sample if the selected data contained any empty cells in the 2×2 table of respiratory infection by xerophthalmia). For samples of size $n = 50$, the most commonly selected (%) correlation structures by each of the criteria were as follows: $QIC_P(R)$: independence (45.2%); $QIC_{HH}(R)$: independence (42.8%); $CIC(R)$: autoregressive (41.8%); and $T_{2HH}(R)$: autoregressive (45.2%). For samples of size $n = 100$, the independence structure was most commonly selected by all of the criteria, with the following probabilities: $QIC_P(R)$: 66.6%; $QIC_{HH}(R)$: 65.4%; $CIC(R)$: 55.8%; and $T_{2HH}(R)$: 49.8%. In all cases (both sample sizes and all four criteria), the *least* chosen correlation structure was selected in at least 15% of the sub-samples.

6 Discussion

GEEs are popular for a variety of reasons, including the population average or marginal interpretation, the reduced dependence on distributional assumptions and the consistency of the estimators under fairly general assumptions, even when the working correlation matrix is misspecified. This latter property does not hold uniformly, as was first demonstrated by Pepe & Anderson (1994): in some settings, only a diagonal working correlation will yield unbiased estimators. Further, even when the mean model assumption is met, the type I error rate may not be at the nominal level following selection of the working correlation structure. These results suggest that caution is warranted in the use of any information criterion to decide among candidate working correlation structures in GEEs and that they should not be routinely adopted as part of a standard analytic strategy without careful consideration of whether the conditional mean model assumption is met and whether the sample size is sufficient so that type I error rates will be near the nominal level.

Acknowledgements

Dr Moodie is supported by a "Chercheurs boursiers" career award from the Fonds de recherche de Québec en santé. Author Murphy is supported by a doctoral award from the Natural Sciences in Engineering Research Council of Canada. Authors Turgeon and Schulz are supported by doctoral awards from the Fonds de recherche de Québec en nature et technologies.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.95

The ISI's Journal for the Rapid Dissemination of Statistics Research

References

Akaike, H (1973), 'Information theory and an extension of the maximum likelihood principle', Second International Symposium on Information Theory, Akademinai Kiado, 267–281.

Cui, J (2007), 'QIC program and model selection in GEE analyses', *Stata Journal*, **7**(2), 209–220.

Diggle, P, Heagerty, P, Liang, KY & Zeger, S (2002), *Analysis of Longitudinal Data*, (2nd edition), *Oxford University Press*, Oxford, UK.

Hardin, JW & Hilbe, J (2007), *Generalized Linear Models and Extensions*, 3rd edn., *Stata Press*, College Station TX.

Hin, LY & Wang, YG (2008), 'Working-correlation-structure identification in generalized estimating equations', *Statistics in Medicine*, **28**(4), 642–658.

Hurvich, C & Tsai, CL (1990), 'The impact of model selection on inference in linear regression', *American Statistician*, **44**(3), 214–217.

Liang, KY & Zeger, SL (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**(1), 13–22.

Martin, LJ, Melnichouk, O, Huszti, E, Connelly, PW, Greenberg, CV, Minkin, S & Boyd, NF (2015), 'Serum lipids, lipoproteins, and risk of breast cancer: a nested case-control study using multiple time points', *Journal of the National Cancer Institute*, **107**(5), djv032.

McCullagh, P & Nelder, JA (1989), *Generalized Linear Models*, 2nd edn., Vol. 37, *CRC Press*, Boca Raton, FL.

Montoya, D, Yallop, ML & Memmott, J (2015), 'Functional group diversity increases with modularity in complex food webs', *Nature Communications*. DOI: 10.1038/ncomms8379.

Pan, W (2001), 'Akaike's information criterion in generalized estimating equations', *Biometrics*, **57**(1), 120–125.

Pepe, MS & Anderson, GL (1994), 'A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data', *Communications in Statistics – Simulation and Computation*, **23**(4), 939–951.

Qu, A, Lindsay, BG & Li, B (2000), 'Improving generalised estimating equations using quadratic inference functions', *Biometrika*, **87**, 823–836.

Stein, R, Shapatava, E, Williams, W, Griffin, T, Bell, K, Lyons, B & Uhl, G (2015), 'Reduced sexual risk behaviors among young men of color who have sex with men: findings from the community-based organization behavioral outcomes of Many Men, Many Voices (CBOP-3MV) Project', *Prevention Science*. DOI: 10.1007/s11121–015–0565–8.

Wang, L & Qu, A (2009), 'Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(1), 177–190.

Wedderburn, RWM (1974), 'Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method', *Biometrika*, **61**(3), 439–447.

Westgate, PM (2014), 'Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach', *Biometrical Journal*, **56**(3), 461–476.

Zhou, J & Qu, A (2012), 'Informative estimation and selection of correlation structure for longitudinal data', *Journal of the American Statistical Association*, **107**, 701–710.