eAppendix for "A call caution in using information criteria to select the working correlation structure in generalized estimating equations"

Wilhemina Adoma Pels*, Shomoita Alam*, Lindsay N Carpp,
and Erica E M Moodie (*joint first authors)

In this supplement, we briefly review the most common information criteria for selecting a working correlation structure in the context of generalized estimating equations, before turning to a particular form of marginal model in which only the independence structure should be used for estimation to demonstrate the bias that can occur when an incorrect structure is chosen – a bias first raised in a cautionary note over two decades ago [7].

Quasi-likelihood and related criteria

Suppose data are collected on n individuals over time, with the i^{th} individual being observed m_i times; in what follows t ($t = 1, \ldots, m_i$) will index observation times within the i^{th} person ($i = 1, \ldots, n$). 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 t^{th} 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. Let $\boldsymbol{\mu_i}$ be the m_i-vector of mean response values for the i^{th} individual, and fix a link function $g(\cdot)$ that relates the parameters $\boldsymbol{\beta}$ to the mean (e.g. a log function for count data). In a GEE, estimates $\hat{\boldsymbol{\beta}}$ are obtained by solving the equation

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

where $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ and $\mathbf{V}_i$ is the (analyst-specified) working correlation structure. Empir-

ical, or "robust", standard errors are then used to ensure correct estimation of the standard errors even when the working correlation structure is not of the correct form (though efficiency is compromised by a badly mis-specified working correlation). GEE mean and variance parameter estimators are consistent for the true values provided that the mean model is correctly specified [4].

The GEE framework is semi-parametric, in that a likelihood or error distribution need not be specified for estimation or inference. The *quasi-likelihood* was proposed [5] to provide many of the convenient features of a likelihood. Inspired by Akaike's information criterion, the quasi-likelihood information criterion, QIC_P, for generalized estimating equations [6] was next proposed. The QIC_P uses the observed quasi-likelihood, penalized for complexity, as a method for choosing between competing models, thus balancing model fit and parsimony. The optimal variance-covariance matrix is that which gives the smallest quasi-likelihood information criterion. Several variations on QIC_P have been proposed, including:

- QIC_{HH}: an alternative formulation that evaluates the quasi-likelihood at mean model parameter estimates generated under the independence working correlation [1];
- CIC: the correlation information criterion [2], which is simply one half of the penalty of QIC_P with no adjustment for the value of the quasi-likelihood;
- T_{2HH}: like the CIC, it is one half of the penalty of QIC_{HH} [1];
- DBC: the determinant-based criterion [3], which is based on a bias-corrected robust covariance estimator.

Simulation study

It is sometimes the case that the parameters of the data-generating model are not of primary

interest. For example, in a setting where prior history may affect a patient's outcome, but a physician only has access to current health, the *cross-sectional* model parameters are of greater scientific relevance. It has long been known in the statistical literature that GEEs will yield biased estimates of cross-sectional model parameters when the true data-generating mechanism relies on additional covariate history – including previous outcome measures – unless an independence correlation structure is assumed [7].

We assess the performance of the five information criteria to yield unbiased mean parameter estimators following selection of a working correlation structure. We consider three data-generating processes:

Model A

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

where $Y_{i0} = 0$, (X_{it}, ϵ_{it}) have mean zero and are mutually independent of each other and of $Y_{i(t-1)}$.

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 (X_{it}, ϵ_{it}) are independent of each other and $Y_{i(t-1)}$. X_{it} has mean one and ϵ_{it} has mean zero.

Model C

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

where η_i and (X_{it}, ϵ_{it}) are mutually independent, each with mean zero.

For Models A and B, the true data-generating mechanism relies on covariate history and so

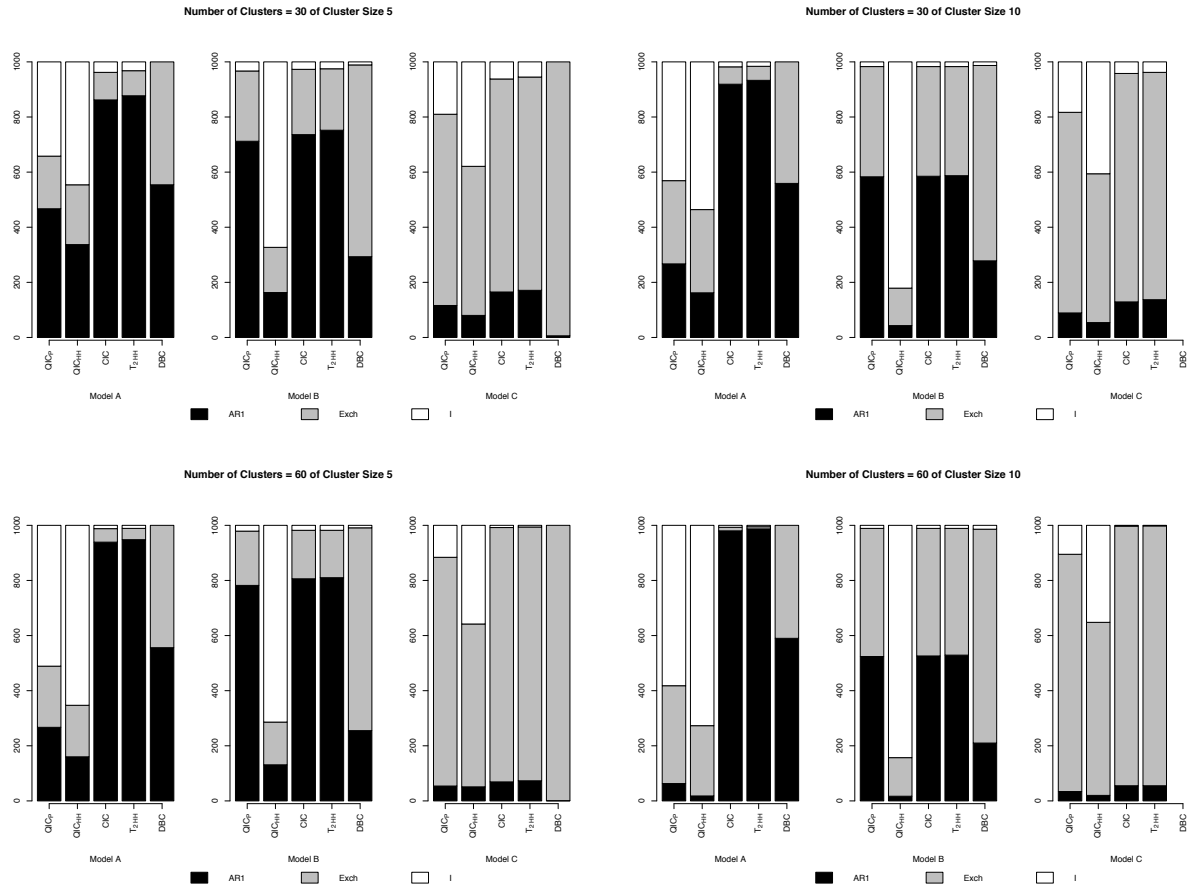
unbiased estimation is only assured through the use of an independence working correlation structure, whereas in Model C the true outcome model is cross-sectional. Models A and B do not yield named or recognizable correlation structures, whereas the true correlation structure from Model C is exchangeable (also called compound symmetric, i.e. all off-diagonal elements share the same correlation).

Data are generated 1,000 times for $n = 30$ and 60 individuals, each with $m_i = 5$ and 10 observations. The selection criteria are used to choose between three different structures: an independence structure, an exchangeable structure, and a first-order autoregressive structure.

As anticipated [7], only the independence working correlation structure yields unbiased estimates for Models A and B. For model C, estimates are unbiased for all working correlation matrices, but efficiency is gained when the true (exchangeable) correlation structure is used (eTable 1). If, rather than fixing the working correlation structure, model selection is undertaken using one of the criteria described above, bias is substantial for all but QIC_{HH} (eTable 2) due to the frequent selection of a working correlation structure other than independence (eFigure 1).

eTable 1. Mean and and standard deviation of $\widehat{\beta}$ for fixed working correlation

Number of Clusters, n	Cluster Size	Fixed Correlation Structure	Model A	Model B	Model C
30	5	I	0.500 (0.095)	0.495 (0.535)	0.503 (0.118)
		Exch	0.457 (0.086)	0.324 (0.409)	0.504 (0.094)
		AR(1)	0.420 (0.079)	0.314 (0.339)	0.504 (0.105)
30	10	I	0.500 (0.068)	0.516 (1.483)	0.500 (0.083)
		Exch	0.473 (0.065)	0.324 (1.005)	0.501 (0.062)
		AR(1)	0.408 (0.054)	0.314 (0.807)	0.500 (0.072)
60	5	I	0.498 (0.068)	0.487 (0.366)	0.499 (0.085)
		Exch	0.451 (0.060)	0.319 (0.286)	0.499 (0.063)
		AR(1)	0.414 (0.055)	0.309 (0.237)	0.499 (0.071)
60	10	I	0.500 (0.048)	0.496 (1.115)	0.502 (0.058)
		Exch	0.472 (0.044)	0.304 (0.778)	0.501 (0.042)
		AR(1)	0.406 (0.038)	0.292 (0.662)	0.501 (0.050)



eFigure 1. Frequency of working correlation matrix selected for the marginal model by QIC_P, QIC_{HH}, CIC, T_{2HH} and DBC from 1000 simulated datasets for 30 (top row) and 60 (bottom row) clusters.

eTable 2. Mean and standard deviation of $\widehat{\beta}$ when selection criteria are used to choose between Independence, Exchangeable, and AR(1) working correlation structures. The true value of β is 0.5.

Number of clusters	Cluster size	Selection criteria	Model A	Model B	Model C
30	5	QIC_P	0.472 (0.100)	0.329 (0.361)	0.504 (0.107)
		QIC_{HH}	0.483 (0.099)	0.490 (0.534)	0.503 (0.116)
		CIC	0.429 (0.085)	0.324 (0.357)	0.505 (0.098)
		T_{2HH}	0.428 (0.084)	0.320 (0.352)	0.505 (0.097)
		DBC	0.436 (0.083)	0.327 (0.374)	0.504 (0.094)
30	10	QIC_P	0.482 (0.076)	0.312 (0.940)	0.501 (0.073)
		QIC_{HH}	0.491 (0.073)	0.497 (1.557)	0.500 (0.079)
		CIC	0.415 (0.061)	0.311 (0.940)	0.501 (0.065)
		T_{2HH}	0.413 (0.060)	0.310 (0.937)	0.501 (0.065)
		DBC	0.436 (0.065)	0.339 (0.922)	0.501 (0.062)
60	5	QIC_P	0.480 (0.076)	0.319 (0.249)	0.499 (0.073)
		QIC_{HH}	0.490 (0.073)	0.484 (0.365)	0.498 (0.082)
		CIC	0.417 (0.057)	0.315 (0.247)	0.499 (0.064)
		T_{2HH}	0.416 (0.057)	0.314 (0.248)	0.499 (0.064)
		DBC	0.430 (0.060)	0.322 (0.267)	0.499 (0.063)
60	10	QIC_P	0.491 (0.052)	0.288 (0.709)	0.501 (0.048)
		QIC_{HH}	0.496 (0.049)	0.495 (1.115)	0.501 (0.055)
		CIC	0.408 (0.040)	0.287 (0.709)	0.501 (0.043)
		T_{2HH}	0.408 (0.039)	0.285 (0.708)	0.501 (0.043)
		DBC	0.432 (0.050)	0.310 (0.740)	0.501 (0.042)

References

[1] JW Hardin and JM Hilbe. *Generalized Linear Models and Extensions*. Stata press, 2007.

[2] Lin-Yee Hin and You-Gan Wang. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine*, 28(4):642–658, 2008.

[3] A Jaman, AHM Mahbub Latif, W Bari, and AS Wahed. A determinant-based criterion for working correlation structure selection in generalized estimating equations. *Statistics in Medicine*, 35(11):1819–1833, 2015.

[4] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[5] Peter McCullagh and John A Nelder. *Generalized linear models*. CRC press, 1989.

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

[7] MS Pepe and GL Anderson. 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, 1994.