

Comparison of Criteria to Select Working Correlation Matrix in Generalized Estimating Equations *

ZHU XIAOLU ZHU ZHONGYI

(*Department of Statistics, School of Management, Fudan University, Shanghai, 200433*)

Abstract

In this paper, we compare two modified Gaussian pseudolikelihood criteria (GPCs) with existing Gaussian pseudolikelihood criterion and empirical likelihood based criteria to choose the working correlation matrix in generalized estimating equations approach. Rich simulation studies are conducted to investigate the performance of these criteria under a range of model settings. The results show that the modified criteria outperform the original GPC and empirical likelihood based criteria in most cases in terms of selection accuracy. Empirical likelihood based criteria perform better to identify exchangeable structure in data with binary response. In the end, these criteria are applied to epilepsy seizure and Madras longitudinal schizophrenia study clinical data sets analysis.

Keywords: Longitudinal data, model selection, pseudolikelihood, empirical likelihood.

AMS Subject Classification: 62F03.

§1. Introduction

Longitudinal data analysis is very popular in practice since it focuses on common data which has repeated observations over time related to a class of individuals. One of the most important characteristics of longitudinal data is its repeated measurements of dynamic relations within individuals over time, so the analysis of the within-subject relationship has been paid much attention in the literature. Liang and Zeger (1986) proposed an approach of generalized estimating equations (GEE) which introduces the working correlation structure to characterize this within-subject relationship and this approach has been widely used due to many attractive features. Rather than making assumptions on likelihood functions, GEE approach only requires partial specifications of marginal means and working covariance structures about the joint distribution of repeated measurements. The estimates of regression coefficients yielded by the GEE approach are always consistent even when the working correlation matrix is misspecified. Another attractive feature of GEE is that it relies on a sequential of estimating equations and the computational

*The research was supported by National Natural Science Foundation of China (11271080).

Received February 27, 2013. Revised March 14, 2013.

implementation can be easily achieved. Because of these semiparametric advantages, the GEE method is quite simple and useful to deal with frequently occurred non-Gaussian longitudinal data.

Although GEE approach yields consistent estimates for regression parameters, the problem of how to partially specify the joint distribution of repeated observations is still open to discussion, which may influence the estimation efficiency. For example, the efficiency of estimates is limited by the misspecification of working correlation matrix. Bai et al. (2010) showed that the efficiency of statistical inference for GEE can be improved with a proper chosen working correlation matrix. Then how to choose a proper working correlation matrix becomes an important problem. There has been recent work in the literature to achieve this purpose.

Qu et al. (2000) proposed quadratic inference functions (QIF) to improve the estimation efficiency. By representing the inverse of working correlation matrix with the linear combination of basis matrices, QIF can yield more efficient parameter estimates than GEE estimates when working correlation matrix is misspecified. Based on this methodology, other approaches (eg. Wang, 2011) have been derived to select the working correlation matrix.

Based on quasi-likelihood procedures, Pan (2001) proposed a selection criterion called quasi-likelihood under the independence model criterion (QIC) by modifying the AIC (Akaike, 1973) criterion to improve the efficiency of GEE estimators, and this approach is also used to select the working correlation matrix. However, as mentioned by Hin and Wang (2009), due to its assumption of independent correlation structure which actually cannot portray the real within-subject relationship, QIC approach is not that powerful when used to choose a proper working correlation matrix. By modifying QIC, Hin and Wang (2009) constructed another selection criterion called correlation information criterion (CIC) and showed that CIC is more powerful than QIC.

Hin et al. (2007) compared QIC with Rotnitzky-Jewell (RJ) criterion to identify the true correlation structure via rich simulations with Gaussian and binomial responses, and exchangeable or AR(1) working correlation matrix. They found that RJ performs better when the true structure is exchangeable while QIC performs better for an AR(1) structure.

QIC uses independence structure to define the quasi-likelihood and CIC is not likelihood-like, so they may ignore some underlying information about the within-subject correlation among repeated observations. The empirical likelihood approach proposed by Chen and Lazar (2012) takes advantage of desirable likelihood properties from empirical likelihood without specification of parametric distributions. They conducted two criteria which replace the parametric likelihood components with derived empirical likelihood

terms in AIC and BIC to choose appropriate working correlation structure. Simulation study demonstrated that EL-based criteria are more powerful compared with QIC and CIC methods under certain settings.

Another method Carey and Wang (2011) used is Gaussian pseudolikelihood criterion (GPC) which is the extended QL function in Hall and Severini (1998). GPC is mainly compared with another form of Rotnitzky-Jewell criterion by Carey and Wang (2011). As shown by the simulation results, the correct identification rates of the GPC are relatively higher in most cases when data sets are generated from two specific continuous distribution. And they only considered two kinds of working correlation matrix. However, the question whether it is more effective compared to EL-based criteria when used in discrete data sets or to select from a broader sets of working correlation matrices is still open to doubt.

Taking the GPC as the likelihood-like term in AIC and BIC, we generalize the Gaussian pseudolikelihood criterion raised by Carey and Wang (2011) and modify Gaussian pseudolikelihood criteria as AGPC and BGPC by adding penalty terms related to number of free parameters. Compared with QIC, we do not assume the independent correlation structure which is not reasonable in practice. Since the number of free parameters are considered in our modified criteria, the modified criteria are more rigorous and reliable than GPC method. This paper, in Section 2, will compare the modified criteria with some competitive criteria including GPC and EL-based criteria EAIC and EBIC, which show superiority in recent work. Sections 3 through 5 present simulation studies on criteria of GPC, AGPC, BGPC, EAIC and EBIC to select working correlation matrix, two applications and summary of the paper respectively. Simulation studies show that the modified criteria tend to increase the selection accuracy and they are particularly good at identifying the underlying independent structure and other structures with strong correlation.

§2. Methods for Model Selection

Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$, $i = 1, 2, \dots, K$, denote the K response vectors where Y_{it} is the t -th, $t = 1, \dots, n_i$, observation of the i -th subject. The design matrix for subject i is denoted as $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{in_i})^T$, where $\mathbf{X}_{it} = (x_{it1}, x_{it2}, \dots, x_{itp})^T$ denote the p dimensional vector of covariates for Y_{it} . For each subject i , the n_i observations are often correlated with each other and the working correlation matrix $R(\boldsymbol{\alpha})$ in GEE method proposed by Liang and Zeger (1986) is assumed to describe this kind of correlation, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_s)^T$ is an s dimensional vector to characterize $R(\boldsymbol{\alpha})$. Assuming $n_i = n$, four common used structures in this paper are as follows

(A) Independence (IN), when there is no within subject correlation, $R(\boldsymbol{\alpha}) = I_n$, where

I_n denotes the identity matrix;

- (B) Exchangeable (EX), in which α involves one unknown parameter;
- (C) One-order autoregressive (AR(1)), in which α involves one unknown parameter;
- (D) Stationary structure (ST), which involves $n - 1$ unknown parameters in α .

$$\begin{array}{ccc}
 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, & \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}, & \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n-1} & \cdots & \alpha^2 & \alpha & 1 \end{pmatrix}, \\
 \text{IN} & \text{EX} & \text{AR(1)} \\
 \\
 & \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} \\ \alpha_1 & 1 & \alpha_1 & \cdots & \alpha_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{n-2} & \alpha_{n-3} & \cdots & \ddots & \alpha_1 \\ \alpha_{n-1} & \alpha_{n-2} & \cdots & \alpha_1 & 1 \end{pmatrix}. & \\
 & \text{ST} &
 \end{array}$$

In the GEE approach, the conditional expectation of Y_{it} is denoted as $E(Y_{it}|X_{it}) = \mu_{it}(\beta) = g^{-1}(X_{it}^T \beta)$ where $g(\cdot)$ is the link function of GLM model and β is an unknown p dimensional regression parameter vector of interest. It is also assumed that $\text{Var}(Y_{it}) = v(\mu_{it}(\beta))\phi = \sigma_{it}^2\phi$. In addition, $V_i = A_i^{1/2}R(\alpha)A_i^{1/2}\phi$ is defined as working covariance structure of \mathbf{Y}_i , where A_i is a diagonal matrix with σ_{it}^2 , $t = 1, \dots, n_i$, along the diagonal and ϕ is the over-dispersion parameter.

2.1 Modified Gaussian Pseudolikelihood Criteria

First, we briefly introduce the Gaussian pseudolikelihood criterion (GPC) used by Carey and Wang (2011) since our criteria are its extensions. GPC is based on the extended QL function in Hall and Severini's work in 1998. Hall and Severini (1998) constructed an extended quasi-likelihood function and its first derivatives provide valid estimating function for β and α . They also established the consistency and asymptotic multivariate normality properties of associated parameters estimated by extend QL functions. The extended quasi-likelihood function shares some features with multivariate Gaussian likelihood function. Denote V_i as $W_i(\alpha) = A_i^{1/2}R(\alpha)A_i^{1/2}\phi$ and $E(\mathbf{Y}_i|\mathbf{X}_i) = \boldsymbol{\mu}_i$, where $W_i(\alpha)$

is a given or estimated working covariance structure in estimating equations. Then the extended quasi-likelihood function is

$$LG = -\frac{1}{2} \sum_i \{(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T W_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) + \log(|W_i|)\}. \quad (2.1)$$

Then the Gaussian pseudolikelihood criterion defined by Carey and Wang (2011) is $GPC = -2LG$. By using the estimations of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from GEE, it is easy to fit $\boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$ and $R_i(\hat{\boldsymbol{\alpha}})$. Model selection is accomplished by selecting from candidate models the one minimizing GPC.

In Carey and Wang (2011), GPC is used to distinguish only two kinds of working correlation matrices which are EX and AR(1). It is obvious that both structures contain one unknown parameter in $R(\boldsymbol{\alpha})$. When selecting from various structures whose number of free parameters vary from each other, GPC is not necessarily efficient without this consideration. As mentioned by Hall and Severini (1998), GPC is an extended quasi-likelihood function and it has likelihood-like properties. It also performs well in selecting working correlation matrix in Gaussian and Lognormal responses. Since GEE has no likelihood function, Pan (2001) modified the AIC by replacing the likelihood component with a quasi-likelihood one under working independence model. Pan (2001) also redefined the discrepancy and conducted the QIC through some approximation and adjustment for the penalty term. Just as QIC proposed by Pan (2001), modified criteria are improved by replacing the likelihood component in AIC with this extended quasi-likelihood component. The first modified criterion is

$$AGPC = -2LG + 2 \dim(\boldsymbol{\theta}) = GPC + 2 \dim(\boldsymbol{\theta}), \quad (2.2)$$

Schwarz (1978) conducted the Bayes based procedure BIC and it differs from AIC only in the penalty term. The modified BGPC criterion with the same penalty term as BIC's is

$$BGPC = -2LG + \log(K) \dim(\boldsymbol{\theta}) = GPC + \log(K) \dim(\boldsymbol{\theta}), \quad (2.3)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ is vector of free parameters to be estimated.

As discussed by Schwarz (1978), AIC and BIC procedures differ significantly from each other for large numbers of observations. So do AGPC and BGPC, which will be seen especially in simulation studies with four candidates.

Different from QIC in Pan (2001), we replace the parametric likelihood term with an extended QL function directly and skip further approximation and derivations. In addition, when conducting the criterion, we do not assume independence working correlation matrix which is essential in QIC.

2.2 Empirical Likelihood Versions of AIC and BIC: EAIC and EBIC

Chen and Lazar (2012) constructed another two criteria to select the working correlation matrix by substituting empirical likelihood for parametric likelihood in AIC and BIC. These two criteria are proved to be more effective than QIC and CIC in Chen and Lazar (2012). So we also investigate and compare their performance with GPC and modified ones in this paper. In the framework of EAIC and EBIC, they mainly focused on the derivation of empirical likelihood ratio (ELR) under a full model which assumes the stationary (ST) working correlation matrix denoted as $R_F(\boldsymbol{\alpha})$, where $p+n-1$ free parameters are recorded in $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \alpha_1, \dots, \alpha_{n-1})$.

To obtain the full model ELR, new estimating function ($g^F(\cdot)$) is defined first under the assumed full model as

$$g^F((\mathbf{Y}_i, \mathbf{X}_i), \boldsymbol{\beta}, \alpha_1, \dots, \alpha_{n-1}; R_F(\boldsymbol{\alpha})) = \begin{pmatrix} (\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^T)^T A_i^{-1/2} R_F^{-1}(\alpha_1, \dots, \alpha_{n-1}) A_i^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \\ \sum_{t=1}^{n-1} e_{it}(\boldsymbol{\beta}) e_{i,t+1}(\boldsymbol{\beta}) - \alpha_1 \hat{\phi}(\boldsymbol{\beta}) (n-1-p/K) \\ \vdots \\ \sum_{t=1}^1 e_{it}(\boldsymbol{\beta}) e_{i,t+n-1}(\boldsymbol{\beta}) - \alpha_{n-1} \hat{\phi}(\boldsymbol{\beta}) (1-p/K) \end{pmatrix}_{(p+n-1) \times 1}, \quad (2.4)$$

where the Pearson residuals are

$$e_{it}(\boldsymbol{\beta}) = (Y_{it} - \mu_{it}(\boldsymbol{\beta})) / \sqrt{v(\mu_{it}(\boldsymbol{\beta}))} \quad \text{and} \quad \hat{\phi}(\boldsymbol{\beta}) = \sum_{i=1}^K \sum_{t=1}^n e_{it}^2 / (Kn - p).$$

Then ELR function can be expressed by replacing the new estimating equation for the old GEE term:

$$\mathcal{R}^F(\boldsymbol{\beta}, \boldsymbol{\alpha}^T) = \sup \left\{ \prod_{i=1}^K K \omega_i : \omega_i \geq 0, \sum_{i=1}^K \omega_i = 1, \sum_{i=1}^K \omega_i g^F((\mathbf{Y}_i, \mathbf{X}_i), \boldsymbol{\beta}, \boldsymbol{\alpha}^T; R_F(\boldsymbol{\alpha})) = 0 \right\}. \quad (2.5)$$

Here, estimating equations for each candidates share the same full working correlation matrix which is a more general stationary structure. Based on this new $g^F(\cdot)$, ELRs can be signed with different and comparable values for various working correlation matrices. This is because maximum empirical likelihood estimators (MELE) are the same with GEE estimators without this modification and ELRs will all equal to 1.

Corresponding to each one of four candidate working correlation structures, Chen and Lazar (2012) calculated GEE estimates $\hat{\boldsymbol{\theta}}_G$ as $(\hat{\boldsymbol{\beta}}_{\text{IN}}, \mathbf{0}^T)^T$ assuming IN structure, $(\hat{\boldsymbol{\beta}}_{\text{EX}}^T, \hat{\boldsymbol{\alpha}}_{\text{EX}}^T)^T$ assuming EX structure, $(\hat{\boldsymbol{\beta}}_{\text{AR}(1)}^T, \hat{\boldsymbol{\alpha}}_{\text{AR}(1)}^T)^T$ assuming AR(1) structure and

$(\hat{\beta}_{ST}^T, \hat{\alpha}_{ST}^T)^T$ assuming ST structure, then obtained ELR values by plugging GEE estimates into (2.5). They constructed two criteria to select the working correlation matrix which lead to larger ELR.

$$EAIC = -2 \log \mathcal{R}^F(\hat{\theta}_G) + 2 \dim(\theta), \quad (2.6)$$

$$EBIC = -2 \log \mathcal{R}^F(\hat{\theta}_G) + \log(K) \dim(\theta), \quad (2.7)$$

where $\dim(\theta)$ is the dimension of free parameters $(\beta^T, \alpha^T)^T$ to be estimated. It is obvious that minimum EAIC and EBIC suggest the most likely model.

2.3 Comparison of These Criteria

All these criteria are likelihood-like, and the modified methods and EL-based criteria can be viewed as extensions of AIC and BIC. GPC use the extended quasi-likelihood function directly as the selection criterion, which performs well in Gaussian and Lognormal data. Compared with GPC, our modified criteria take number of free parameters into account and make it the penalty term. This modification is more reasonable to improve the selection accuracy when working correlation matrix has various numbers of parameters to be estimated. These methods are proved to be superior in recent literature works. Here, we compare their performance in a broader way with more candidate working correlation matrices and responses from both continuous and discrete distributions.

We will see in the simulation studies for example that GPC will lose efficiency when ST structure is in the selecting candidates, since it has more free parameters than other candidates. Without the penalty of free parameters, GPC tend to choose more complicated ST structure. The EL-based approach uses the empirical likelihood term which differs from our extended QL term. And both likelihood-like terms enjoy good properties. In general, these criteria are all in the low computational complexity. They have their own motivations and merits, and their performance under certain circumstances are investigated through the following simulation studies.

§3. Simulation Studies

The main purpose of simulation studies is to compare the modified method with the original criterion and EL-based criteria, and to find their specific applicability and dominant positions under various settings when choosing the underlying working correlation matrix. For every model specification, we record the selection percentage of each four candidates (IN, EX, AR(1) and ST) over 1000 simulation runs. The number of subjects are set to be 50, 75 and 150. Both discrete and continuous responses are generated to

compare the selection accuracy (selection percentage of underlying working correlation matrix) of different criteria.

For discrete response simulation, each of the subject is observed 3 times and α is set to be 0.25 or 0.5. Binary data is drawn from the model $\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 x_{ij2} + \beta_3(j - 1)$, $j = 1, 2, 3$, which was also adopted by Chen and Lazar (2012) and the data generation process is also the same. And Poisson data is generated from the model $\text{log}(\mu_{ij}) = \beta_1 + \beta_2 x_{ij2} + \beta_3(j - 1)$, $j = 1, 2, 3$, and the data is generated with Matlab function Poisson (Dalthorp and Madsen). The parameters $\beta_1 = -\beta_2 = -\beta_3 = 0.25$ and x_{ij2} independently and identically follow Bernoulli distribution with $P(x_{ij2} = 1) = 1/2$.

3.1 Selection Accuracy when ST is Excluded from Candidates

All these criteria perform better with a larger sample size and stronger within subject correlation in all settings. For Binary data shown in Table 1, GPC is no longer in force if $R(\alpha)$ is independence. It tends to choose EX or AR(1) structure which have one more free parameter in the working correlation matrix. After adding a penalty term, AGPC and BGPC have significant improvement in terms of selection accuracy. EBIC also performs similar to BGPC, but along with the increase number of individuals, the benefit of BGPC leads to a more favorable comparison.

Table 1 Binary response, selection from candidates: EX, AR(1) and IN

Binary	$R(\alpha)$	$R_0(\alpha) : K = 50$					$R_0(\alpha) : K = 75$					$R_0(\alpha) : K = 150$				
		EX		AR(1)		IN	EX		AR(1)		IN	EX		AR(1)		IN
		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5	
EX	GPC	77.0	87.1	27.1	12.4	49.2	79.8	89.2	22.3	8.4	50.9	90.8	96.6	12.2	1.7	50.6
	AGPC	71.8	87.1	23.2	12.4	12.5	78.6	89.2	21.2	8.4	10.7	90.8	96.6	12.2	1.7	11.8
	BGPC	65.0	87.1	19.4	12.4	4.6	74.4	89.2	18.0	8.4	2.8	90.6	96.6	12.2	1.7	2.0
	EAIC	73.1	89.7	24.0	15.4	13.8	79.4	92.0	22.1	10.5	10.7	91.3	97.4	12.7	3.0	11.8
	EBIC	66.2	89.7	20.0	15.4	4.5	75.3	92.0	19.2	10.5	3.0	91.1	97.4	12.7	3.0	2.2
AR(1)	GPC	23.0	12.9	72.9	87.6	50.7	20.2	10.8	77.7	91.6	49.1	9.2	3.4	87.8	98.3	49.4
	AGPC	21.3	12.9	64.7	87.5	11.5	19.6	10.8	74.6	91.6	11.8	9.2	3.4	87.7	98.3	8.9
	BGPC	19.1	12.9	53.8	87.5	4.1	18.3	10.8	67.4	91.6	2.6	9.1	3.4	86.6	98.3	1.6
	EAIC	20.4	10.2	64.4	84.5	11.5	18.9	8.0	73.8	89.5	11.7	8.7	2.6	87.2	97.0	9.1
	EBIC	17.7	10.2	53.3	84.2	4.0	17.6	8.0	67.0	89.5	2.7	8.6	2.6	86.1	97.0	1.6
IN	GPC	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	AGPC	6.9	0.0	12.1	0.1	76.0	1.8	0.0	4.2	0.0	77.5	0.0	0.0	0.1	0.0	79.3
	BGPC	15.9	0.0	26.8	0.1	91.3	7.3	0.0	14.6	0.0	94.6	0.3	0.0	1.2	0.0	96.4
	EAIC	6.5	0.1	11.6	0.1	74.7	1.7	0.0	4.1	0.0	77.6	0.0	0.0	0.1	0.0	79.1
	EBIC	16.1	0.1	26.7	0.4	91.5	7.1	0.0	13.8	0.0	94.3	0.3	0.0	1.2	0.0	96.2

Note: $R_0(\alpha)$ corresponds to the underlying working correlation matrix. The highlighted numbers stand for the selection accuracy for each method under various model settings.

When the underlying working correlation matrix is AR(1), GPC criterion has the highest selection accuracy. AGPC is more effective than EAIC and EBIC, and BGPC performs better than EBIC. When the initial correlation is stronger with $\alpha = 0.5$ or when the sample size is large, Gaussian pseudolikelihood criterion and modified criteria show similar correct selection rates.

If $R(\alpha)$ is exchangeable, the EL-based methods tend to select the correct working correlation matrix more often when $\alpha = 0.5$. When $\alpha = 0.25$ the performance of BGPC and EBIC seem to be similar, the same happens between AGPC and EAIC.

In general, although the modified criteria do not have the absolute highest selection accuracy, they perform similar to the best one along with the increase of sample size. Moreover, performance of GPC and modified ones tends to be identical when the correlation gets stronger.

Table 2 Poisson response, selection from candidates: EX, AR(1) and IN

Poisson	$R(\alpha)$	$R_0(\alpha) : K = 50$					$R_0(\alpha) : K = 75$					$R_0(\alpha) : K = 150$				
		EX		AR(1)		IN	EX		AR(1)		IN	EX		AR(1)		IN
		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5	
EX	GPC	77.6	89.8	31.4	17.4	48.4	78.7	91.4	25.1	10.8	47.7	88.2	97.9	13.6	3.5	50.6
	AGPC	71.0	89.8	26.4	17.4	11.9	76.1	91.4	23.3	10.8	11.2	88.1	97.9	13.4	3.5	10.9
	BGPC	60.6	89.8	21.3	17.4	3.8	70.5	91.4	20.6	10.8	3.9	87.5	97.9	13.0	3.5	1.7
	EAIC	69.8	87.9	26.4	17.9	16.5	75.1	90.8	23.0	13.2	13.6	87.0	97.2	13.9	5.0	12.3
	EBIC	60.1	87.8	22.1	17.9	7.5	69.0	90.8	21.3	13.2	5.8	86.1	97.2	13.5	5.0	2.2
AR(1)	GPC	22.4	10.2	68.4	82.6	51.2	21.3	8.6	74.9	89.2	52.0	11.8	2.1	86.4	96.5	49.2
	AGPC	18.0	10.2	59.9	82.6	12.5	19.8	8.6	69.9	89.2	12.9	11.8	2.1	86.1	96.5	11.4
	BGPC	14.6	10.2	48.5	82.2	3.8	17.4	8.6	61.9	89.2	2.7	11.6	2.1	85.1	96.5	2.1
	EAIC	19.8	12.1	61.8	82.0	18.5	21.5	9.2	70.6	86.8	16.0	13.0	2.8	85.6	95.0	14.0
	EBIC	16.9	12.1	52.4	81.8	7.3	18.4	9.2	63.4	86.8	5.2	12.8	2.8	84.6	95.0	3.2
IN	GPC	0.0	0.0	0.2	0.0	0.4	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2
	AGPC	11.0	0.0	13.7	0.0	75.6	4.1	0.0	6.8	0.0	75.9	0.1	0.0	0.5	0.0	77.7
	BGPC	24.8	0.0	30.2	0.4	92.4	12.1	0.0	17.5	0.0	93.4	0.9	0.0	1.9	0.0	96.2
	EAIC	10.4	0.0	11.8	0.1	65.0	3.4	0.0	6.4	0.0	70.4	0.0	0.0	0.5	0.0	73.7
	EBIC	23.0	0.1	25.5	0.3	85.2	12.6	0.0	15.3	0.0	89.0	1.1	0.0	1.9	0.0	94.6

Note: $R_0(\alpha)$ corresponds to the underlying working correlation matrix. The highlighted numbers stand for the selection accuracy for each method under various model settings.

Table 2 summarizes the selection results of working correlation matrices under the Poisson models. The same with the Binary models, GPC lose efficiency to choose underlying independent structure. BGPC enjoys absolute advantage over other criteria in this case. If the true structure is EX, AGPC is more effective than EAIC, and BGPC is more effective than EBIC. If the true structure is AR(1), EAIC is more effective than AGPC

and EBIC is more effective than BGPC when $\alpha = 0.25$ and $K = 50$ or 100 . But when sample size goes larger or within subject correlation goes stronger, the situation becomes inverse.

Overall, the Gaussian pseudolikelihood based criteria perform better than EL-based criteria. GPC and the modified criteria perform identically best to choose the right structure both for AR(1) and EX structure with $\alpha = 0.5$. Under weaker working correlation structure, even if GPC is better than AGPC and BGPC, the gaps narrow as the sample size goes larger. EL-based criteria still enjoy small advantages when correlation in AR(1) patten is not strong.

3.2 Selection Accuracy when ST is Included in Candidates

Table 3 Binary response, selection from candidates: EX, AR(1), IN and ST

Binary	$R_0(\alpha) : K = 50$					$R_0(\alpha) : K = 75$					$R_0(\alpha) : K = 150$				
	EX		AR(1)		IN	EX		AR(1)		IN	EX		AR(1)		IN
	$R(\alpha)$	α													
EX	GPC	3.2 7.4	1.5	1.2	1.1	2.7 4.3	1.2	0.0	0.6	2.4 5.2	0.4	0.0	0.6		
	AGPC	63.6 74.6	22.7	12.4	10.7	68.0 75.7	21.0	7.1	9.8	80.1 75.4	12.1	0.3	10.6		
	BGPC	62.2 81.8	19.3	12.4	4.3	71.3 85.0	17.9	8.4	2.6	89.1 92.7	12.2	1.7	2.0		
	EAIC	64.7 77.6	23.4	15.4	11.8	69.4 81.1	21.8	10.1	9.7	81.4 83.4	12.6	0.4	10.8		
	EBIC	63.3 84.6	19.9	15.4	4.2	72.5 88.6	19.1	10.5	2.8	89.7 95.6	12.7	3.0	2.2		
AR(1)	GPC	0.2 0.5	0.8 5.0	0.8		0.2 0.0	1.0 3.2	0.1		0.1 0.0	0.8 3.5	0.1			
	AGPC	21.0 12.7	57.8 79.5	10.7		19.5 8.8	65.0 81.9	10.5		8.7 0.8	79.7 85.2	8.3			
	BGPC	19.0 12.9	51.7 85.3	4.1		18.3 10.8	65.5 90.1	2.6		9.1 3.1	85.3 97.7	1.6			
	EAIC	20.1 10.0	56.9 73.4	10.8		18.8 7.7	64.5 78.5	10.5		8.4 0.7	79.3 84.1	8.5			
	EBIC	17.6 10.2	50.8 77.5	4.0		17.6 7.8	64.9 85.9	2.7		8.6 2.6	85.0 96.0	1.6			
IN	GPC	0.0 0.0	0.0 0.0	0.1		0.0 0.0	0.0 0.0	0.0		0.0 0.0	0.0 0.0	0.0			
	AGPC	5.7 0.0	11.6 0.1	72.9		1.6 0.0	3.9 0.0	73.9		0.0 0.0	0.0 0.0	76.9			
	BGPC	15.7 0.0	26.2 0.1	90.9		7.0 0.0	14.4 0.0	94.1		0.3 0.0	1.2 0.0	96.3			
	EAIC	5.3 0.0	11.0 0.1	71.5		1.5 0.0	3.7 0.0	73.9		0.0 0.0	0.0 0.0	76.6			
	EBIC	16.0 0.0	26.2 0.1	91.0		6.8 0.0	13.7 0.0	93.8		0.3 0.0	1.2 0.0	96.1			
ST	GPC	96.6 92.1	97.7 93.8	98.0		97.1 95.7	97.8 96.8	99.3		97.5 94.8	98.8 96.5	99.3			
	AGPC	9.7 12.7	7.9 8.0	5.7		10.9 15.5	10.1 11.0	5.8		11.2 23.8	8.2 14.5	4.2			
	BGPC	3.1 5.3	2.8 2.2	0.7		3.4 4.2	2.2 1.5	0.7		1.5 4.2	1.3 0.6	0.1			
	EAIC	9.9 12.4	8.7 11.1	5.9		10.3 11.2	10.0 11.4	5.9		10.2 15.9	8.1 15.5	4.1			
	EBIC	3.1 5.2	3.1 7.0	0.8		3.1 3.6	2.3 3.6	0.7		1.4 1.8	1.1 1.0	0.1			

Note: $R_0(\alpha)$ corresponds to the underlying working correlation matrix. The highlighted numbers stand for the selection accuracy for each method under various model settings.

To have a better understanding of selection behavior of these criteria, we include the stationary structure into candidate working correlation matrices. In application, it

is better to include a more complicated model if we have no enough information about data. The result differs from simulation with three candidate matrices because GPC becomes invalid not only in independent structure setting but also in EX and AR(1) structures as seen in Table 3 and Table 4. GPC tends to select more complicated stationary structure under all model settings. When taking the number of free parameters into account, modified criteria AGPC and BGPC do improve dramatically in choosing the right structures.

Table 4 Poisson response, selection from candidates: EX, AR(1), IN and ST

Poisson	$R(\alpha)$ α	$R_0(\alpha) : K = 50$					$R_0(\alpha) : K = 75$					$R_0(\alpha) : K = 150$				
		EX		AR(1)		IN	EX		AR(1)		IN	EX		AR(1)		IN
		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5		0.25	0.5	0.25	0.5	
EX	GPC	24.4	42.0	11.2	8.3	8.7	22.7	43.2	6.9	2.2	7.4	24.3	39.1	2.6	0.1	4.8
	AGPC	65.2	82.1	25.9	17.4	11.5	68.9	82.5	23.1	10.5	10.5	79.7	83.5	13.3	1.2	9.6
	BGPC	59.4	87.3	21.3	17.4	3.8	68.8	87.7	20.6	10.8	3.6	86.0	96.0	13.0	3.5	1.6
	EAIC	60.5	72.0	25.1	17.7	14.9	65.8	75.0	22.3	12.9	11.7	76.5	81.0	13.8	3.3	10.9
	EBIC	57.0	81.3	21.8	17.8	7.2	67.0	85.6	21.3	13.2	5.3	84.0	94.8	13.5	5.0	2.0
AR(1)	GPC	2.7	2.3	7.8	21.6	5.5	2.6	1.0	7.5	19.9	3.2	0.3	0.0	7.2	22.9	2.4
	AGPC	17.4	10.2	53.7	73.7	10.9	19.7	7.4	62.4	80.7	11.2	11.7	0.8	76.0	83.9	9.9
	BGPC	14.5	10.2	47.1	79.1	3.8	17.4	8.6	60.4	87.1	2.4	11.6	2.1	83.7	95.6	2.1
	EAIC	18.9	11.7	52.3	68.9	15.2	21.4	8.7	59.7	74.1	13.6	12.4	1.5	74.8	82.7	12.4
	EBIC	16.4	12.0	48.9	75.6	6.6	18.4	9.2	60.0	83.3	4.6	12.8	2.8	82.1	93.2	3.2
IN	GPC	0.0	0.0	0.1	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.2
	AGPC	10.6	0.0	12.4	0.0	72.5	4.0	0.0	6.2	0.0	73.1	0.1	0.0	0.5	0.0	75.5
	BGPC	24.3	0.0	29.7	0.4	91.6	12.0	0.0	17.4	0.0	93.0	0.9	0.0	1.9	0.0	96.2
	EAIC	9.5	0.0	9.9	0.1	60.0	3.1	0.0	5.4	0.0	66.0	0.0	0.0	0.5	0.0	69.9
	EBIC	22.4	0.1	24.2	0.3	83.6	12.2	0.0	14.8	0.0	87.7	1.1	0.0	1.8	0.0	94.5
ST	GPC	72.9	55.7	80.9	70.1	85.5	74.7	55.8	85.6	77.9	89.1	75.4	60.9	90.2	77.0	92.6
	AGPC	6.8	7.7	8.0	8.9	5.1	7.4	10.1	8.3	8.8	5.2	8.5	15.7	10.2	14.9	5.0
	BGPC	1.8	2.5	1.9	3.1	0.8	1.8	3.7	1.6	2.1	1.0	1.5	1.9	1.4	0.9	0.1
	EAIC	11.1	16.3	12.7	13.3	9.9	9.7	16.3	12.6	13.0	8.7	11.1	17.5	10.9	14.0	6.8
	EBIC	4.2	6.6	5.1	6.3	2.6	2.4	5.2	3.9	3.5	2.4	2.1	2.4	2.6	1.8	0.3

Note: $R_0(\alpha)$ corresponds to the underlying working correlation matrix. The highlighted numbers stand for the selection accuracy for each method under various model settings.

In general, for model settings with AR(1) and independence structure, BGPC almost performs best in choosing the correct working correlation matrix and this pattern tends to be stable when there are more individuals. For model settings with EX structure, the EL-based criteria still have moderate high selection accuracy especially for EBIC, but BGPC is still competitive and promising under these circumstances since its selection accuracy is slightly inferior to that of EBIC.

For Poisson data, modified criteria perform well under each model setting; AGPC is better than EAIC and BGPC behaves better than EBIC except that EBIC achieves 48.9% selection accuracy while BGPC has lower selection accuracy of 47.1% in the case $K = 50$ and $R(0.25)$ has AR(1) patten. When the correlation is weaker both for EX and AR(1) structure in data with fewer subjects, AGPC is more powerful than BGPC. However, BGPC tends to transcend the correct selection performance of AGPC when sample size is large.

In general, under the most of model settings for Poisson data, BGPC enjoys the highest rate of recognition of right working correlation matrix and this advantage becomes more stable as sample size goes larger and correlation goes stronger.

In simulation of continuous response, each of the subject is observed 4 times and α is set to be 0.25 or 0.5. The marginal mean model for Gaussian distribution is $\mu_{ij} = \beta_0 + \beta_1 x_{ij}$, which is also adopted by Carey and Wang (2011). The Lognormal model is set to be $E(\log(Y_{ij})) = \beta_0 + \beta_1 x_{ij}$. In the true model, $\beta_0 = 0$, $\beta_1 = 1$, x_{ij} independent and identically follow distribution of $U(j, j + 1)$, and $V(\mu) = \sigma^2 = 1$, $\phi = 1$.

The main findings are similar to those of discrete data, the detailed tables are omitted here.

3.3 Conclusions from the Simulation Studies

(1) Under all model settings, GPC becomes invalid in identifying independent structures. When the stationary structure is in the candidates, GPC lose efficiency to select the any correct underlying working correlation matrix. GPC is proved to select a more complicated structure since it treats all models equally. The modified criteria are significantly improved in these cases. BGPC can identify the true independent working correlation matrix with superior rates.

(2) The modified criteria outperform other criteria in most cases when sample size is large and within subject correlation is strong. In general, at least one of the modified criteria outperforms EL-based criteria in most cases and is competitive in other cases. When the stationary structure is excluded from the candidate pool, AGPC is more effective than BGPC if within-subject correlation is weak, and both the modified criteria can obtain the same efficiency when correlation becomes stronger. When the stationary structure is in the candidates, BGPC performs better than AGPC in Binary data with stronger correlation or larger sample size. While for continuous data and Poisson data, BGPC is consistently more powerful than AGPC.

(3) For Binary data, EL-based criteria have higher selection accuracy than the modified ones on rare situations. However the differences are minor.

§4. Two Examples

4.1 Epilepsy Seizure Data

We analyze the data of 59 patients with epilepsy who were randomly treated with progabide or placebo which is also analyzed in Carey and Wang (2011). As mentioned by Carey and Wang (2011), this is discrete data with high degree of extra-Poisson variation and within subject correlation. Hence we use Poisson distribution in model fitting. The response variable is the number of epileptic seizures documented in continuous every 2-week after ingestion of drug or placebo. Five covariates related to this data are Progabide Treatment (Whether treated with progabide), Base Count (Cumulative 8-week seizure counts before trial), Age (Age of patients), Visit 4 (Indicator of the fourth period) and interaction Progabide Treatment * Base Count between Progabide Treatment and Base Count. Log transformations have been applied to covariates Base Count, Age and Progabide Treatment * Base Count. Five new variables are represented as Protr, Lbase, Lage, V4 and Protr * Lbase. Variance function is set to be $v(\mu) = \mu^2$ just as the selection result in Carey and Wang (2011). And we choose the proper working correlation matrix from four candidate matrices including stationary structure. The fitted results are summarized in Table 5.

Table 5 Model fitting results for epilepsy seizure data

$R(\alpha)$	EX	AR(1)	IN	ST
intercept	-0.607(0.826)	-0.799(0.838)	-0.607(0.826)	-0.786(0.840)
Protr	-0.794(0.401)	-0.856(0.415)	-0.794(0.401)	-0.854(0.415)
Lbase	0.872(0.123)	0.860(0.114)	0.872(0.123)	0.860(0.114)
Lage	0.302(0.239)	0.367(0.245)	0.302(0.239)	0.364(0.245)
V4	-0.127(0.096)	-0.086(0.101)	-0.127(0.096)	-0.105(0.099)
Protr * Lbase	0.286(0.207)	0.305(0.215)	0.286(0.207)	0.304(0.215)
GPC	1371.517	1374.020	1403.950	1371.341*
AGPC	1385.517*	1388.020	1415.950	1389.341
BGPC	1400.060*	1402.563	1428.415	1408.039
EAIC	19.159	16.719*	67.592	18.013
EBIC	33.702	31.261*	80.057	36.711

Note: Standard errors are given in parentheses, * indicates the minimum criteria values.

GPC tends to choose the more complicated stationary structure which is also the case in simulation study. The choice of the modified criteria is exchangeable structure

which is the same with the result in Carey and Wang (2011), while choice of EL-based criteria is AR(1) structure. Although selection result for the modified criteria is consistent with that in Carey and Wang (2011), Carey and Wang (2011) only choose from EX and AR(1) structures and they share the same number of free parameters. When adding the independence and stationary working correlation matrices into the candidate pool, GPC used in Carey and Wang (2011) lose efficiency. The modified criteria are more robust to identify the underlying structure. We also see that the standard errors of the model with EX structure are the smallest except for Base Count effect.

4.2 Madras Longitudinal Schizophrenia Study

The data set is available in Diggle et al. (2002) investigating the course of positive and negative psychiatric symptoms over the first year after initial hospitalization for schizophrenia. The Madras longitudinal study collected data on 6 symptoms which are classified into positive and negative symptoms. So the response variable Y is the binary variable indicating the presence of “thoughts disorder”. The logistic regression is fitted with main effects “MONTH” ($0, \dots, 11$), duration in the first year following hospitalization for schizophrenia; “AGE” (0 stands for age-at-onset less than 20, 1 otherwise), onset ages of patients; “GENDER” (0 = male, 1 otherwise); and also two interaction terms between “MONTH” and the other two main effects. In this example, there are 87 patients and 17 of them have incomplete data with duration less than 12 months. We choose the working correlation matrix from EX, AR(1) and IN structures.

Table 6 Model fitting results for Madras longitudinal schizophrenia study

$R(\alpha)$	EX	AR(1)	IN
intercept	0.621(0.315)	0.543(0.291)	0.643(0.305)
MONTH	-0.272(0.064)	-0.233(0.055)	-0.254(0.059)
AGE	1.057(0.546)	0.621(0.459)	0.811(0.493)
GENDER	-0.591(0.525)	-0.132(0.419)	-0.388(0.449)
MONTHAGE	-0.087(0.093)	-0.097(0.084)	-0.137(0.094)
MONTHGENDER	-0.140(0.096)	-0.156(0.088)	-0.113(0.096)
GPC	752.6658	441.1584*	441.1584*
AGPC	766.6658	455.1584	453.1584*
BGPC	783.8462	472.3388	467.8845*
EAIC	2950.754	2974.191	2906.411*
EBIC	2967.935	2991.371	2921.137*

Note: Standard errors are given in parentheses, * indicates the minimum criteria values.

As illustrated in Table 6, all the criteria choose independence working correlation matrix. However, it is worth noting that differences between independence and AR(1) structures are minor for GPC and the modified criteria. What's more, standard errors of model with AR(1) structure are much smaller than others. So this case indicates that AR(1) structure is more proper to present the within subject correlation. And this result is consistent with Hin and Wang (2009).

§5. Summary

In this paper, we compare two modified criteria based on the Gaussian pseudolikelihood criterion raised by Carey and Wang (2011) with the original criterion and empirical likelihood based criteria to choose the working correlation matrix in GEE approach. The new criteria are modified by adding a penalty term describing number of free parameters.

Rich simulation studies have been conducted both in discrete and continuous data to investigate the performance of these criteria. The results show that in general Gaussian pseudolikelihood criterion performs well in selection from candidate matrices with stationary working correlation matrix excluded, but it is invalid to identify independent structure. When choosing from four candidate structures, Gaussian pseudolikelihood criterion loses efficiency. The modified criteria exhibit remarkable and robust improvement in these cases. Compared with other criteria, the modified criteria is especially skilled in recognizing independent structure and underlying working correlation matrix with strong correlation.

In our simulation study, the EL-based criteria have similar correct recognition rates when real data is generated from Gaussian distribution. EAIC and EBIC also show superiority in choosing EX working correlation matrix in binary data. It's worth noting that the modified criteria perform more stable in various settings, and their advantages become obvious when choosing structures with stronger with-subject correlation and larger sample sizes. Even when EL-based criteria dominate in certain cases, the differences between the modified criteria and EL-based criteria are minor.

From our simulation, it appears that modified criteria, especially BGPC, dominate in most cases and are stable to choose the underlying working correlation matrix with a broad range of candidates. But for binary data, EL-based criteria are more reliable in terms of identifying exchangeable correlation structure. Our next step is to examine their simultaneous selection efficiency in multiple panels with different settings.

References

- [1] Akaike, H., Information theory and an extension of the maximum likelihood principle, In Petrov, B.N. and Csaki, F. (Eds.), *Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, 1973, 267–281.
- [2] Bai, Y., Fung, W.K. and Zhu, Z.Y., Weighted empirical likelihood for generalized linear models with longitudinal data, *Journal of Statistical Planning and Inference*, **140**(11)(2010), 3446–3456.
- [3] Carey, V.J. and Wang, Y.-G., Working covariance model selection for generalized estimating equations, *Statistics in Medicine*, **30**(26)(2011), 3117–3124.
- [4] Chen, J. and Lazar, N.A., Selection of working correlation structure in generalized estimating equations via empirical likelihood, *Journal of Computational and Graphical Statistics*, **21**(1)(2012), 18–41.
- [5] Dalthorp, D. and Madsen, L., Simulation of correlated count-valued random variables: Brief description, Available at <http://www.stat.oregonstate.edu/node/3709>.
- [6] Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L., *Analysis of Longitudinal Data*, Oxford University Press: Oxford, 2002.
- [7] Hall, D.B. and Severini, T.A., Extended generalized estimating equations for clustered data, *Journal of the American Statistical Association*, **93**(444)(1998), 1365–1375.
- [8] Hin, L.Y., Carey, V.J. and Wang, Y.-G., Criteria for working-correlation-structure selection in GEE: Assessment via simulation, *The American Statistician*, **61**(4)(2007), 360–364.
- [9] Hin, L.Y. and Wang, Y.-G., Working-correlation-structure identification in generalized estimating equations, *Statistics in Medicine*, **28**(4)(2009), 642–658.
- [10] Liang, K.Y. and Zeger, S.L., Longitudinal data analysis using generalized linear models, *Biometrika*, **73**(1)(1986), 13–22.
- [11] Pan, W., Akaike's information criterion in generalized estimating equations, *Biometrics*, **57**(1)(2001), 120–125.
- [12] Qu, A., Lindsay, B.G. and Li, B., Improving generalised estimating equations using quadratic inference functions, *Biometrika*, **87**(4)(2000), 823–836.
- [13] Schwarz, G., Estimating the dimension of a model, *The Annals of Statistics*, **6**(2)(1978), 461–464.
- [14] Wang, P., Zhou, J.H. and Qu, A., Model selection for correlation structure for diverging clustered data, Manuscript, 2011.

广义估计方程中工作相关阵选择准则的比较

朱晓露 朱仲义

(复旦大学管理学院统计学系)

在本文中, 我们比较了广义估计方程中相关阵基于高斯伪似然、修正的高斯伪似然和经验似然的选择方法. 通过大量的模拟研究, 我们发现修正的高斯伪似然方法优于其他两种方法. 对二项离散模型, 经验似然方法在选择可交换相关结构时有更好的表现. 最后, 通过两个实例分析, 进一步分析了各个选择方法之间的优劣性.

关键词: 纵向数据, 模型选择, 伪似然, 经验似然.

学科分类号: O212.1.