

Singapore Management University  
Institutional Knowledge at Singapore Management University

---

Research Collection School Of Economics

School of Economics

---

7-2009

# Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method

Denis H. Y. LEUNG

Singapore Management University, [denisleung@smu.edu.sg](mailto:denisleung@smu.edu.sg)

You Gan WANG


CSIRO

Min ZHU

University of Sydney

**DOI:** <https://doi.org/10.1093/biostatistics/kxp002>

Follow this and additional works at: [https://ink.library.smu.edu.sg/soe\\_research](https://ink.library.smu.edu.sg/soe_research)

 Part of the [Econometrics Commons](#), and the [Medicine and Health Sciences Commons](#)

---

## Citation

LEUNG, Denis H. Y.; WANG, You Gan; and ZHU, Min. Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. (2009). *Biostatistics*. 10, (3), 436-445. Research Collection School Of Economics.

**Available at:** [https://ink.library.smu.edu.sg/soe\\_research/514](https://ink.library.smu.edu.sg/soe_research/514)

This Journal Article is brought to you for free and open access by the School of Economics at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Economics by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [libIR@smu.edu.sg](mailto:libIR@smu.edu.sg).

# Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method

DENIS H. Y. LEUNG\*

*School of Economics, Singapore Management University, 90 Stamford Road, Singapore*  
denisleung@smu.edu.sg

YOU-GAN WANG

*Commonwealth Scientific and Industrial Research Organization, Mathematical and Information Sciences, CSIRO Long Pocket Laboratories, 120 Meiers Road, Indooroopilly, Queensland 4068, Australia*

MIN ZHU

*Finance Discipline, School of Business and Economics, University of Sydney, NSW 2006, and Division of Mathematical and Information Sciences, Commonwealth Scientific and Industrial Research Organisation, PO Box 120, Cleveland, QLD 4163, Australia*

## SUMMARY

The method of generalized estimating equations (GEEs) provides consistent estimates of the regression parameters in a marginal regression model for longitudinal data, even when the working correlation model is misspecified (Liang and Zeger, 1986). However, the efficiency of a GEE estimate can be seriously affected by the choice of the working correlation model. This study addresses this problem by proposing a hybrid method that combines multiple GEEs based on different working correlation models, using the empirical likelihood method (Qin and Lawless, 1994). Analyses show that this hybrid method is more efficient than a GEE using a misspecified working correlation model. Furthermore, if one of the working correlation structures correctly models the within-subject correlations, then this hybrid method provides the most efficient parameter estimates. In simulations, the hybrid method's finite-sample performance is superior to a GEE under any of the commonly used working correlation models and is almost fully efficient in all scenarios studied. The hybrid method is illustrated using data from a longitudinal study of the respiratory infection rates in 275 Indonesian children.

*Keywords:* Empirical likelihood; Generalized estimating equations; Longitudinal data.

## 1. INTRODUCTION

Generalized estimating equations (GEEs) have been found to be very useful in analysis of correlated and longitudinal outcomes using marginal regression models. Following Liang and Zeger (1986), many

\*To whom correspondence should be addressed.

aspects of GEE have been explored. Reviews of GEE include Pendergast *and others* (1996) and Desmond (1997).

In a marginal regression model, the primary interest is in the regression parameters, which characterize the expectations of the subject's response over time. However, in order to make proper inference about the regression parameters, the within-subject covariance (correlation) structures must be taken into consideration. The GEE approach has been popular because estimates of mean parameters remain consistent even if the correlation or the covariance structure is misspecified. On the other hand, accurate modeling of the correlation structure generally improves statistical inference on means (Albert and McShane, 1995; Fitzmaurice, 1995; Hall and Severini, 1998). Wang and Carey (2003) analyzed how efficiency can be affected by (i) the choice of the working correlation structure, (ii) the method by which the working correlation parameters are estimated, and (iii) the layout of the design matrix. Higher moments can be incorporated into estimation using a generalized version of GEE called GEE2 (Liang *and others*, 1992). However, bias or efficiency losses may be introduced if higher moment assumptions of GEE2 are incorrectly specified. For this reason, GEE2 has yet to receive wide application.

In GEE modeling, the most commonly used working correlation models are the exchangeable, AR(1) and MA(1). Wang and Carey (2003) found that among the 3, AR(1) is the most robust. However, they also demonstrated scenarios where the exchangeable and MA(1) models give better results than the AR(1) model. Therefore, it remains an issue of how to choose a working correlation model in a particular GEE analysis. The AR(1) and MA(1) working correlation models appear to be favored by users of GEE because (i) in most situations, they are sufficient as an approximation to the true correlation structure and (ii) they represent sensible compromises between the independence model (which ignores within-subject correlations) and the completely unstructured model (which requires the estimation of large number of nuisance parameters). These considerations lead us to propose a method that incorporates all 3 working correlation models in a single framework, yielding a method that is efficient if one of these 3 working correlation models correctly captures the true correlation structure, and robust even if none of the working correlation models is correct. The proposed method can be generalized to combining any number of GEEs with working correlation models other than the exchangeable, AR(1) and MA(1) models.

Each GEE with a particular working correlation model is a mean-zero estimating equation under the true parameters. When we are interested in combining multiple GEEs, then there are more estimating equations than the number of parameters. In situations involving independent data where the number of estimating equations may be larger than the number of parameters, Qin and Lawless (1994) showed how to combine efficiently the estimation equations using an empirical likelihood (EL) (Owen, 1988). We exploit this attribute of the EL technique to combine GEEs. The individual GEEs using different working correlations are used as constraints in an EL for the parameters of interest. The parameter estimates are then obtained by maximizing the EL. Other than its role as a tool for combining estimating equations, EL also inherits a number of desirable properties from its parametric counterparts, as described in Owen (2001).

The rest of this paper is organized as follows. Section 2 presents the basic problem, the modeling framework of the proposed method, and its large-sample properties. The results of a simulation study are summarized in Section 3. In Section 4, the method is illustrated using a real data set. Section 5 concludes the paper with a discussion. Detailed simulation results and proofs are given as supplementary material available at *Biostatistics* online, <http://biostatistics.oxfordjournals.org>.

## 2. COMBINING GEEs

Consider a longitudinal study in which there are  $n$  subjects, each of whom is measured at  $K$  time points. Let  $y_i = (y_{i1}, \dots, y_{iK})^T$  denote the underlying outcome for the  $i$ th subject,  $x_i$  an associated vector of  $r \times 1$  covariates, and  $x_{ik}$  the value of the covariate at time  $k$ . Denote the marginal mean outcome at the

$k$ th measurement for the  $i$ th subject by  $\mu_{ik}(\beta) = g(x_{ik}^T \beta)$ , for a vector of unknown parameters,  $\beta$ . For conciseness, we suppress the explicit association of  $\mu_{ik}$  with  $\beta$  if there is no confusion.

Following Liang and Zeger (1986), a GEE can be used to estimate the regression parameters,  $\beta$ ,

$$\sum_{i=1}^n D_i^T V_i^{-1} \{y_i - \mu_i\} = 0, \quad (2.1)$$

where  $\mu_i = (\mu_{i1}, \dots, \mu_{ik})^T$ ,  $D_i = \partial \mu_i / \partial \beta^T$ , and  $V_i$  is the covariance matrix of  $y_i$ . The matrix  $V_i$  is often modeled as  $\phi A_i^{1/2} R(\alpha) A_i^{1/2}$ , where  $A_i$  is a diagonal matrix representing the variances of  $y_{ik}$ ,  $R(\alpha)$  is a “working correlation” depending on a set of unknown parameters  $\alpha$ , and  $\phi$  is a scale parameter used to model over-dispersion or under-dispersion. Liang and Zeger (1986) showed that, whether or not  $V_i$  is correctly specified, the estimators of  $\beta$  obtained from (2.1) remain consistent. In addition, if  $V_i = \text{Cov}(y_i)$  can be consistently estimated up to  $n^{-1/2}$ , then the estimator of  $\beta$  is fully efficient. On the other hand, an incorrectly specified  $V_i$  will lead to a loss of efficiency (Wang and Carey, 2003).

As Liang and Zeger (1986) pointed out, (2.1) can be re-expressed as a function of  $\beta$  by writing  $\alpha \equiv \alpha(\beta, \phi)$  and  $\phi \equiv \phi(\beta)$ . An iterative algorithm can then be used to estimate  $\beta$ ,  $\alpha$ , and  $\phi$ , starting with initial estimates of  $\alpha$  and  $\phi$ . For suggested methods for estimating  $\alpha$  and  $\phi$ , see Liang and Zeger (1986), Chaganty (1997), and Chaganty and Shults (1999). For ease of exposition, we assume  $\phi = 1$ . Liang and Zeger (1986, pp 17–18) discussed choices for  $R(\alpha)$ , while Wang and Carey (2003) studied their relative efficiencies. For any chosen working correlation matrix  $R \equiv R(\alpha)$ , write  $S_i(\beta) \equiv D_i^T A_i^{-1/2} R^{-1} A_i^{-1/2} (y_i - \mu_i)$ . Then, the GEE (2.1) estimates are solutions of

$$S(\beta) \equiv \sum_{i=1}^n S_i(\beta) = 0. \quad (2.2)$$

Now, consider different, linearly independent choices of  $R(\alpha)$ , say  $R^j(\alpha)$ ,  $j = 1, \dots, J$ , and write

$$S^j(\beta) \equiv \sum_{i=1}^n S_i^j(\beta) = 0, \quad (2.3)$$

for the estimating equation (2.2) but using working correlation matrix  $R^j(\alpha)$ . Let  $h_i(\beta) \equiv (S_i^1(\beta)^T, \dots, S_i^j(\beta)^T, \dots, S_i^J(\beta)^T)^T$ , and note that  $h_i \equiv h_i(\beta)$  is a function of  $\beta$  only. In general, the dimension of  $h_i$  is higher than the dimension of  $\beta$ . Our propose is to use EL to combine the estimating equations  $S_i^1, \dots, S_i^J$ . If one of the  $S_i^1, \dots, S_i^J$  is the optimal estimating equation, in the sense that it solves (2.2) with  $A_i^{-1/2} \{R(\alpha)\}^{-1} A_i^{-1/2} = V_i^{-1}$ , then the EL estimate will be optimal. If none of them is optimal, then the EL estimate is still consistent and combines optimally the information in  $S_i^1, \dots, S_i^J$ . In practice, a few popular choices of  $R(\alpha)$  may be used; for example, exchangeable, AR(1) and MA(1).

We now describe how to use an EL framework to combine the GEEs in  $h_i$ . Let  $F$  be the distribution function associated with the observations  $\{(y_i, x_i)\}_{i=1}^n$ . Denote  $p_i = dF(y_i | x_i)$  as the jump size of  $F$  at  $(y_i, x_i)$ . Then, the nonparametric likelihood of the data can be written as  $\prod_{i=1}^n dF(y_i | x_i) \equiv \prod_{i=1}^n p_i$ , subject to the constraints  $0 \leq p_i \leq 1, i = 1, \dots, n$ , and  $\sum_{i=1}^n p_i = 1$ . Without any other information, the maximum nonparametric likelihood estimate of  $F$  is the empirical distribution function  $F_n(y_i | x_i) = \sum_{i=1}^n I(y_i \leq y | x_i)$ , which corresponds to  $p_i = 1/n$ . However, suppose we know that  $E(h_i(\beta)) = 0$  under  $F$ . Then, the empirical distribution function is no longer desirable because  $E(h_i(\beta)) \neq 0$  under  $dF_n \equiv p_i = 1/n$ . Instead, we can use the (empirical) likelihood

$$L(\beta) = \prod_{i=1}^n p_i \quad (2.4)$$

subject to the constraints

$$0 \leq p_i \leq 1, \quad i = 1, \dots, n; \quad \sum_{i=1}^n p_i = 1,$$

$$\sum_{i=1}^n p_i \{S_i^1(\beta)^T, \dots, S_i^j(\beta)^T, \dots, S_i^J(\beta)^T\}^T \equiv \sum_{i=1}^n p_i h_i(\beta) = 0.$$

In this formulation, maximizing the EL gives a set of  $p_i$ s such that  $E(h_i(\beta)) = 0$  under  $\{p_i\}_{i=1}^n$ . Since the resulting values of  $\{p_i\}_{i=1}^n$  depend on the extra conditions  $E(h_i(\beta)) = 0$ , which in turn depend on the value of  $\beta$ , the EL estimate of  $F$  is sensitive to the value of  $\beta$ .

The EL (2.4) can be maximized as a constrained maximization problem. By introducing Lagrange multipliers  $\eta, \lambda = (\lambda_1^T, \dots, \lambda_j^T, \dots, \lambda_J^T)^T$ , where each  $\lambda_j$  is  $r \times 1$ , the log-EL can be written as

$$\log L(\beta) = \sum_{i=1}^n \log p_i + \eta \left( 1 - \sum_{i=1}^n p_i \right) - n \lambda^T \sum_{i=1}^n p_i h_i(\beta). \quad (2.5)$$

The values of  $\{p_i\}_{i=1}^n$  can be profiled out by differentiating (2.5) with respect to  $p_i$  to give

$$\frac{1}{p_i} - \eta - n \lambda^T h_i(\beta) = 0 \Rightarrow n - \eta = 0 \Rightarrow \eta = n. \quad (2.6)$$

Equation (2.6) implies that the optimal values of  $\{p_i\}_{i=1}^n$  are

$$p_i = \frac{1}{n\{1 + \lambda^T h_i(\beta)\}}. \quad (2.7)$$

Furthermore, the constraint  $\sum_{i=1}^n p_i h_i(\beta) = 0$  implies that  $\lambda$  satisfies the equation

$$\sum_{i=1}^n \frac{h_i(\beta)}{1 + \lambda^T h_i(\beta)} = 0. \quad (2.8)$$

Using (2.7) and (2.8),  $\eta$  and  $\{p_i\}_{i=1}^n$  can be profiled out in the negative log-EL to give

$$\ell(\beta) \equiv -\log L(\beta) = \sum_{i=1}^n \log\{1 + \lambda^T h_i(\beta)\} - n \log(n). \quad (2.9)$$

Let  $h_i^\beta(\beta) = \partial h_i(\beta) / \partial \beta^T$ . Differentiating (2.9) with respect to  $\beta$  leads to

$$\sum_{i=1}^n \frac{\lambda^T h_i^\beta(\beta)}{1 + \lambda^T h_i(\beta)} = 0. \quad (2.10)$$

The maximum EL estimates  $(\hat{\beta}, \hat{\lambda})$  are the solutions to (2.8) and (2.10). Note that (2.8) consists of  $J$  equations for each parameter and (2.10) consists of  $r$  equations, so in total there are  $(J+1)r$  simultaneous equations to solve. We now give the results for the large-sample behavior of the parameter estimates using the proposed method.

THEOREM 2.1 Under the conditions given in the supplementary material available at *Biostatistics* online, as  $n \rightarrow \infty$ ,

$$n^{1/2}(\hat{\beta} - \beta_*) \xrightarrow{d} \text{MVN}(0, (\Sigma_{12}^T \Sigma_{22}^{-1} \Sigma_{12})^{-1}), \quad (2.11)$$

where  $\Sigma_{12}$  and  $\Sigma_{22}$  are defined in the supplementary material available at *Biostatistics* online.

THEOREM 2.2 If one of the  $S_i^1, \dots, S_i^J$  is the optimal estimating equation, then the EL estimate will be optimal in the sense that it will be equivalent to the GEE estimate with the correct specification of  $R(\alpha)$ . In that case, as  $n \rightarrow \infty$ ,

$$n^{1/2}(\hat{\beta} - \beta_*) \xrightarrow{d} \text{MVN} \left( 0, \left( \lim_{n \rightarrow \infty} \sum_i D_i^{-1} V_i^{-1} D_i^T \right)^{-1} \right). \quad (2.12)$$

In Theorem 2.2,  $S_i^1, \dots, S_i^J$  refer to the researcher's "guesses" of the optimal estimating equation. In practice, it is possible that none of the guesses correspond to the optimal estimating equation. We demonstrate that even in that case, the EL method is still optimal in the sense that it optimally combines the guesses  $S_i^1, \dots, S_i^J$ . This fact can be established by considering the following. Expand the left-hand side of (2.8) in a Taylor expansion around  $\lambda = 0$  to give

$$\begin{aligned} 0 &= \sum_{i=1}^n \frac{h_i(\beta)}{1 + \lambda^T h_i(\beta)} = \sum_{i=1}^n h_i(\beta) - \sum_{i=1}^n h_i(\beta) h_i(\beta)^T \lambda + o_p(1) \\ \Rightarrow \lambda &= \frac{\sum_{i=1}^n h_i(\beta)}{\sum_{i=1}^n h_i(\beta) h_i(\beta)^T} + o_p(1). \end{aligned} \quad (2.13)$$

Substitute (2.13) back into the left-hand side of (2.10) to give

$$\begin{aligned} \sum_{i=1}^n \left\{ \sum_{i=1}^n p_i h_i^\beta(\beta) \right\} \left\{ \sum_{i=1}^n h_i(\beta) h_i(\beta)^T \right\}^{-1} h_i(\beta) &= o_p(1) \\ \Rightarrow n^{-1} \sum_{i=1}^n \left\{ n^{-1} \sum_{i=1}^n h_i^\beta(\beta) \right\} \left\{ n^{-1} \sum_{i=1}^n h_i(\beta) h_i(\beta)^T \right\}^{-1} h_i(\beta) &= 0 \end{aligned} \quad (2.14)$$

asymptotically. Expression (2.14) is in the form of the optimal combination of  $S_i^1, \dots, S_i^J$  (Small and McLeish, 1994, p 94).

In practice, finding the solution to the maximum EL via (2.8) and (2.10) may encounter numerical problems. Furthermore, solving (2.10) requires finding  $h_i^\beta(\beta)$ , which is not straightforward analytically. Therefore, we follow the method of Mittelhammer *and others* (2003) by profiling out the Lagrange multipliers as well, so that for fixed  $\beta$ , the Lagrange multipliers are  $\lambda(\beta) = (\lambda_1^T(\beta), \dots, \lambda_J^T(\beta))^T$ . Given  $\beta$ , the first and second derivatives of (2.9) with respect to  $\lambda$  are

$$\sum_{i=1}^n \frac{h_i(\beta)}{1 + \lambda^T h_i(\beta)} \quad \text{and} \quad \sum_{i=1}^n \frac{-h_i(\beta) h_i(\beta)^T}{\{1 + \lambda^T h_i(\beta)\}^2}. \quad (2.15)$$

Therefore, with some abuse of notation, for given  $\beta$  and a starting value  $\lambda^0$ , the following Newton–Raphson procedure can be used:

$$\lambda^k = \lambda^{k-1} + \sum_{i=1}^n \left\{ \frac{h_i(\beta)}{1 + (\lambda^{k-1})^T h_i(\beta)} \right\}^{-1} \left\{ \frac{h_i(\beta) h_i(\beta)^T}{(1 + (\lambda^{k-1})^T h_i(\beta))^2} \right\}, \quad (2.16)$$

and the solution used as  $\lambda(\beta)$ . Substituting  $\lambda(\beta)$  back into (2.9) then gives

$$\ell(\beta) = \sum_{i=1}^n \log\{1 + \lambda^T(\beta) h_i(\beta)\} - n \log(n) \quad (2.17)$$

which can be maximized with respect to  $\beta$ . Hence, the algorithm can be seen as a nested algorithm with an outside loop that involves maximizing (2.17) with respect to  $\beta$ , while for each  $\beta$ , the inside loop evaluates  $\lambda(\beta)$  using (2.16). The overall maximum gives the maximum EL estimate  $\hat{\beta}$ . Therefore, instead of solving  $(J + 1)r$  simultaneous equations, only a function of  $r$  parameters needs to be maximized. This method becomes especially useful when the number of estimating equations,  $J$ , is large. In our simulations, we used a simple modification of Owen’s S program for the inside loop (<http://www-stat.stanford.edu/~owen/empirical/>). The outside loop was performed using the `optim` function in R. Details of the program can be found in the supplementary material available at *Biostatistics* online.

As an example, let  $\mu_{ik} = \beta_0 + x_{ik}\beta_1$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, K$ , and  $\beta = (\beta_0, \beta_1)$ . Furthermore, suppose 2 different choices of  $R(\alpha)$  are used, namely, the AR(1) with  $\alpha_{ij} = \alpha^{|i-j|}$  and the exchangeable with  $\alpha_{ij} = \alpha$ , for all  $i \neq j$ . Then,

$$h_i(\beta) = (S_i^1(\beta)^T, S_i^2(\beta)^T)^T = \begin{pmatrix} \underline{1}^T A_i^{-1/2} \{R_i^1(\alpha)\}^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\} \\ x_i^T A_i^{-1/2} \{R_i^1(\alpha)\}^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\} \\ \underline{1}^T A_i^{-1/2} \{R_i^2(\alpha)\}^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\} \\ x_i^T A_i^{-1/2} \{R_i^2(\alpha)\}^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\} \end{pmatrix},$$

where  $\underline{1} = (1, \dots, 1)^T$ . Furthermore,  $\lambda = (\lambda_1^T, \lambda_2^T)^T \equiv (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})^T$  and

$$\begin{aligned} \lambda^T h_i(\beta) &= \lambda_{11} \underline{1}^T A_i^{-1/2} \{(R_i^1(\alpha))^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\}\} \\ &\quad + \lambda_{12} x_i^T A_i^{-1/2} \{(R_i^1(\alpha))^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\}\} \\ &\quad + \lambda_{21} \underline{1}^T A_i^{-1/2} \{(R_i^2(\alpha))^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\}\} \\ &\quad + \lambda_{22} x_i^T A_i^{-1/2} \{(R_i^2(\alpha))^{-1} A_i^{-1/2} \{y_i - (\beta_0 \underline{1} + \beta_1 x_i)\}\}. \end{aligned}$$

### 3. SIMULATIONS

We carried out a simulation study to evaluate the moderate sample properties of the proposed method. Two sets of simulations were used:

Set A:  $x_{ik}$ ,  $i = 1, \dots, n$ ,  $k = 1, \dots, 10$ , are independent and identically distributed as  $N(1, \sigma_x^2)$  with  $\sigma_x = 1$ . In this setup,  $x_{ik}$ s are subject-specific covariates that may change over time and are different between subjects but there is no time trend.

Set B:  $x_i = (x_{i1}, \dots, x_{ik})$  followed  $MVN(0, \sigma_x^2 R)$  with  $R$  a  $10 \times 10$  matrix with unit diagonal and off-diagonal elements equal to 0.2 and  $\sigma_x = 0.5$ . In this setup, the intrasubject covariates are correlated over time.

Each set of simulations was based on 1000 runs. Samples sizes were  $n = 100$  and 200. The following model was used for the mean response at time  $k$  for the  $i$ th subject,  $E(y_{ik}) \equiv \mu_{ik} = \beta_0 - \beta_1 x_{ik}$ ,  $k = 1, \dots, 10$ ,  $i = 1, \dots, n$ . The true values of  $(\beta_0, \beta_1)$  were  $(1, -1)$ . The simulation study shows that the proposed method is nearly as efficient as the standard GEE using the correct working correlation model and is superior to the standard GEE using an incorrect working correlation model. We also evaluated the empirical coverage probability of 95% confidence intervals of  $(\beta_0, \beta_1)$  using Theorem 2 and found that they are close to the nominal level. Details of the results are given as supplementary material available at *Biostatistics* online.

#### 4. APPLICATION TO INDONESIAN CHILDREN'S INFECTION DATA

In this section, we apply the proposed method to data from a longitudinal study of the respiratory infection rate in a group of Indonesian children (Diggle *and others*, 2002). The sample consists of 275 preschool children examined at 3-month intervals for 18 months. The maximum number of visits is therefore  $K = 6$ . In total, the 275 children generated 1200 repeated measures of the response (infection versus no infection). The primary interest in this study is to determine the relationship between respiratory infection and Vitamin A deficiency while adjusting for a number of confounders, as listed in Table 1.

We fitted the data by a GEE using an exchangeable (CS), AR(1), and MA(1) working correlation. We then used the proposed method using  $R^1(\alpha) = \text{CS}$ ,  $R^2(\alpha) = \text{AR}(1)$ , and  $R^3(\alpha) = \text{MA}(1)$ . The results are given in Table 1. Standard errors of the estimates for  $\text{GEE}_{\text{MA}(1)}$ ,  $\text{GEE}_{\text{CS}}$ , and  $\text{GEE}_{\text{AR}(1)}$  were obtained from the R routine `geese` in the `geepack` package. Those for the proposed method were estimated using Theorem 1. The results using the 4 methods are quite similar. The conclusions from all methods are the same, that there is no evidence of increased risk for infection due to xerophthalmia. These conclusions are similar to those in earlier studies (e.g. Zeger and Karim, 1991; Lin and Carroll, 2001).

As a means to compare the merits of the different models, we used Akaike's information criterion (AIC) for GEE as developed by Pan (2001). Let  $Q(\beta) \equiv -\sum_{i=1}^n \{y_i - \mu_i(\beta)\}^T V_i^{-1} (y_i - \mu_i(\beta))$ . Then, to assess the merits of a model with parameter estimates  $\hat{\beta}$  obtained using a working correlation matrix  $R$ , the AIC is defined as  $-2Q(\hat{\beta}) + 2 \text{trace}(\hat{\Sigma}_I^{-1} \hat{\Sigma}_R)$ , where  $\hat{\Sigma}_I^{-1}$  is the inverse of the variance of the model coefficients under an independence working correlation and  $\hat{\Sigma}_R$  is that under working correlation  $R$ . The AIC values for the 4 methods are given in Table 2. In Table 2, we also give the second term of the AIC, that is,  $2 \text{trace}(\hat{\Sigma}_I^{-1} \hat{\Sigma}_R)$ , which has been shown in Hin and Wang (2009) to be more accurate in capturing the true correlation structure. The method proposed in this paper has the lowest AIC value and

Table 1. *Parameter estimates (SE) using 4 methods to analyze the Indonesian children's infection data*

Parameter	Method			
	$\text{GEE}_{\text{MA}(1)}$	$\text{GEE}_{\text{CS}}$	$\text{GEE}_{\text{AR}(1)}$	EL
Intercept	-2.371 (0.162)	-2.367 (0.162)	-2.377 (0.162)	-2.370 (0.146)
Age	-0.0317 (0.00628)	-0.0316 (0.00628)	-0.0315 (0.00627)	-0.0317 (0.00578)
Xerophthalmia	0.680 (0.431)	0.651 (0.438)	0.717 (0.419)	0.763 (0.372)
Cos (season)	-0.543 (0.161)	-0.538 (0.160)	-0.550 (0.161)	-0.537 (0.153)
Sex	-0.398 (0.237)	-0.396 (0.237)	-0.394 (0.237)	-0.408 (0.227)
Height for age	-0.0488 (0.0244)	-0.0493 (0.0243)	-0.0478 (0.0244)	-0.0498 (0.0224)



Table 2. Goodness of fit for the 4 methods in the Indonesian children's infection data analysis

Method	AIC	$2 \text{ trace}(\hat{\Sigma}_I^{-1} \hat{\Sigma}_R)$
GEE <sub>MA(1)</sub>	3312.993	12.099
GEE <sub>CS</sub>	3313.351	12.118
GEE <sub>AR(1)</sub>	3313.575	12.137
EL	3310.236	10.199

the lowest value in  $2 \text{ trace}(\hat{\Sigma}_I^{-1} \hat{\Sigma}_R)$  and therefore, by these measures, is the most preferred method for this data set.

## 5. CONCLUSION AND FUTURE RESEARCH

We have introduced a method for combining GEEs in analyzing longitudinal data so as to improve the efficiency of the GEE method when, as is typically the case in practice, correct specification of the correlation structure is problematic.

Validity of the GEE approach requires correct specification of the mean function. If some observations are missing completely at random (Little and Rubin, 1987), the mean function is not affected, and in those cases, the method proposed here remains valid. However, if the missingness probability depends on the observed responses (missing at random) or on the missing responses conditional on the observed responses (nonignorable missingness), then correct modeling of the missingness probability is required for the GEE approach, and therefore our method, to be valid. However, correct specification of the missingness probability is a nontestable condition (Gill *and others*, 1997; Manski, 2003), therefore, if missing at random or nonignorable missingness are suspected, some sort of sensitivity analysis is necessary.

A related method to the one proposed in this paper is the quadratic inference function (QIF) by Qu *et al.* (2000). In their work, the inverse of the working correlation matrix is approximated by a linear combination of basis matrices,  $M_i, i = 1, \dots, m$ , such as

$$R(\alpha)^{-1} \approx a_0 M_0 + a_1 M_1 + \dots + a_m M_m, \quad (5.1)$$

where  $M_1 = I_{K \times K}$  is an identity matrix and  $M_i, i = 2, \dots, m$ , are known symmetric matrices. Instead of estimating  $a_0, \dots, a_m$  directly, they recognized that a GEE based on (5.1) is equivalent to solving the linear combination of a vector of estimating equations:

$$g_n(\beta) = \frac{1}{n} \sum_{i=1}^n g_i(\beta) = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} D_i^T A_i^{-1/2} M_1 A_i^{-1/2} \{y_i - \mu_i\} \\ D_i^T A_i^{-1/2} M_2 A_i^{-1/2} \{y_i - \mu_i\} \\ \vdots \\ D_i^T A_i^{-1/2} M_m A_i^{-1/2} \{y_i - \mu_i\} \end{pmatrix},$$

which can be performed using the generalized method of moments (Hansen, 1982). Their method gives  $\hat{\beta}_{\text{QIF}} = \arg \min_{\beta} Q_n(\beta) \equiv g_n^T(\beta) C_n^{-1}(\beta) g_n(\beta)$ , where  $C_n(\beta) = 1/n^2 \sum_{i=1}^n g_i(\beta) g_i^T(\beta)$  is an estimate of the variance of  $g_n(\beta)$ . We used a modest simulation study to compare our method to QIF and found that the 2 methods give very similar results throughout when the true correlation structure is AR(1), whereas for the CS structure there does seem to be a substantial difference in favor of EL when  $\alpha = 0.7$ . Detailed results of the study are given as supplementary material available at *Biostatistics* online. However, we view these results as preliminary. More work needs to be done to compare these 2 related methods.

Finally, our proposed method is motivated on combining GEEs to find an optimal combination of working correlations for a single data set. There are also situations where multiple longitudinal studies are to be combined in a single analysis, for example, in a meta-analysis or multicenter study (e.g. Inoue *and others*, 2004). In that case,  $S^1(\beta)$ ,  $\dots$ ,  $S^J(\beta)$  may be viewed as GEEs from the different studies that share a common parameter  $\beta$  of interest. The difference between that situation and the one considered here is the multiple samples in the former. The method proposed here can be modified using a multiple sample EL.

## 6. ACKNOWLEDGMENTS

We thank Dr Annie Qu and Ms Guei-feng (Cindy) Tsai for their valuable comments. We also thank the referees and the coeditor for their valuable comments.

## FUNDING

Research Center at Singapore Management University to D.L. (05-C208-SMU-003).

## REFERENCES

- ALBERT, P. S. AND MCSHANE, L. M. (1995). A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. *Biometrics* **51**, 627–638.
- CHAGANTY, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* **63**, 39–54.
- CHAGANTY, N. R. AND SHULTS, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference* **76**, 145–161.
- DESMOND, A. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *Journal of Statistical Planning and Inference* **60**, 77–121.
- DIGGLE, P. J., HEAGERTY, P., LIANG, K. L. AND ZEGER, S. L. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
- FITZMAURICE, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309–317.
- GILL, R. D., VAN DER LAAN, M. J. AND ROBINS, J. M. (1997). Coarsening at random: characterizations, conjectures, counter-examples. In: Lin, D. Y. and Fleming, T. R. (editors), *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. New York: Springer, pp. 255–294.
- HALL, D. AND SEVERINI, T. A. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association* **93**, 1365–1375.
- HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.
- HIN, L.-Y., WANG, Y.-G. (2009) Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* **28**, 642–658.
- INOUE, L. Y., ETZIONI, R., SLATE, E., MORRELL, C. AND PENSON, D. F. (2004). Combining longitudinal studies of PSA. *Biostatistics* **5**, 483–500.
- LIANG, K. Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

- LIANG, K. Y., ZEGER, S. L. AND QAQISH, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society, Series B* **54**, 3–24.
- LIN, X. AND CARROLL, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045–1056.
- LITTLE, R. AND RUBIN, D. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- MANSKI, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer.
- MITTELHAMMER, R., JUDGE, G. AND SCHOENBERG, R. (2003). Empirical evidence concerning the finite sample performance of EL-type structural equation estimation and inference methods. *CUDARE Working Paper Series, Paper 945*. Berkeley, CA: University of California.
- OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.
- OWEN, A. (2001). *Empirical Likelihood*. Boca Raton, FL: Chapman and Hall.
- PAN, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–125.
- PENDERGAST, J. F., GANGE, S. J., NEWTON, M. A., LINDSTROM, M. J., PALTA, M. AND FISHER, M. R. (1996). A survey of methods for analysing clustered binary response data. *International Statistical Review* **64**, 89–118.
- QIN, J. AND LAWLESS, J. (1994). Empirical likelihood and general estimating functions. *Annals of Statistics* **22**, 300–325.
- QU, A., LINDSAY, B. AND LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.
- SMALL, C. G. AND MCLEISH, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. New York: Wiley.
- WANG, Y. G. AND CAREY, V. (2003). Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. *Biometrika* **90**, 29–41.
- ZEGER, S. L. AND KARIM, M. R. (1991). Generalized linear models with random effects. *Journal of the American Statistical Association* **86**, 79–86.

[Received November 1, 2007; revised May 12, 2008; second revision October 18, 2008;  
accepted for publication January 20, 2009]