Statistica Sinica **22** (2012), 755-776
doi:http://dx.doi.org/10.5705/ss.2010.090

A FINITE MIXTURE MODEL FOR WORKING CORRELATION MATRICES IN GENERALIZED ESTIMATING EQUATIONS

Lili Xu[1], Nan Lin[2], Baoxue Zhang[1] and Ning-Zhong Shi[1]

[1]*Northeast Normal University and* [2]*Washington University in St. Louis*

Abstract: The generalized estimating equations (GEE) method has been widely used to analyze longitudinal data since it was proposed by Liang and Zeger (1986). It is well known that the efficiency of the GEE estimator can be seriously affected by the choice of the working correlation matrix. To address the associated misspecification issue, we propose an estimator called mix-GEE based on a finite mixture model for the working correlation. Under mild regularity conditions, the mix-GEE estimator is consistent, asymptotically normal, and asymptotically efficient if data are from a Gaussian mixture model. An important feature of the mix-GEE method is that it guarantees the positive definiteness of the estimated working correlation matrix if either the AR(1) or exchangeable structure is included. It is numerically more stable and displays a better finite sample efficiency than the hybrid GEE method (Leung, Wang, and Zhu (2009)). The value of our method is further demonstrated by simulation studies and data examples.

Key words and phrases: GEE, longitudinal data, working correlation, misspecification, finite mixture, positive definite, pseudo-likelihood, PL-EM algorithm.

1. Introduction

Consider a longitudinal study, in which y_{ij} is the response measured at the jth time point on the ith subject, and $\boldsymbol{x}_{ij}$ is the corresponding p-dimensional covariate, $i = 1, \ldots, N$, $j = 1, \ldots, T$. Observations from different subjects are assumed to be independent whereas those from the same subject are correlated. Suppose that $E(y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij} = \mu(\boldsymbol{x}_{ij}^{\tau}\boldsymbol{\beta})$ and $Var(y_{ij}|\boldsymbol{x}_{ij}) = v_i(\mu_{ij})$, where $\mu(\cdot)$ is a link function and $\boldsymbol{\beta}$ are regression parameters of interest. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^{\tau}$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})^{\tau}$ and $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})^{\tau}$. The generalized estimating equations (GEE) estimator (Liang and Zeger (1986)) of $\boldsymbol{\beta}$ is the solution to the estimating equation

$$U(\boldsymbol{\beta}, R) = \sum_{i=1}^{N} U_i(\boldsymbol{\beta}, R) = \sum_{i=1}^{N} D_i^{\tau} A_i^{-1/2} R^{-1}(\alpha) A_i^{-1/2}(\mathbf{y_i} - \boldsymbol{\mu}_i) = 0, \qquad (1.1)$$

where $A_i = \mathrm{diag}(v_i(\mu_{i1}), \ldots, v_i(\mu_{iT}))$, $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, and $R(\alpha)$ is a working correlation matrix with nuisance parameter α. In the following, we denote by $\hat{\boldsymbol{\beta}}^R$ the solution to (1.1) for a given working correlation matrix R.

Under mild regularity conditions the GEE estimator is consistent, but its efficiency depends on the choice of the working correlation matrix. Extensive studies have been done to provide good choices of the working correlation matrix. Often, one is obtained by simultaneously estimating the nuisance parameter α in (1.1) for a given correlation structure R, such as exchangeable (CS), AR(1), or MA(1). The exchangeable structure $R_{CS}(\alpha)$ has 1 on the diagonal and α elsewhere; the AR(1) structure has $R_{ij} = \alpha^{|i-j|}$; the MA(1) structure $R_{MA}(\alpha)$ has the diagonal and the main off-diagonal components equal to 1 and α, respectively, with the rest of the components equal to 0.

Liang and Zeger (1986) developed several moment methods to estimate these nuisance parameters but these estimates may not exist even in some simple cases (Crowder (1995)). Many other methods have been developed since then, such as extended quasi-likelihood (Hall and Severini (1998)), the quasi-least squares method (Chaganty (1997); Chaganty and Shults (1999)), unbiased estimating equations (Wang and Carey (2004)), decoupled pseudo-likelihood (Wang and Carey (2003); Liu, Lin, and Zhang (2008)) and quadratic inference functions (Qu, Lindsay, and Li (2000)). GEE estimators based on these methods are efficient when the working correlation structure is correctly specified, but correct specification of the working correlation structure is not an easy task. In practice, the working correlation structure is often selected using some ad hoc method from a small list of given structures, and efficiency of the method can be seriously affected when the structure is misspecified (Wang and Carey (2003)), though consistency remains still holds. Recently, from an empirical likelihood perspective, Leung, Wang, and Zhu (2009) proposed a hybrid GEE method to address this misspecification problem by combining multiple GEEs with different linearly independent working correlation matrices. If $R_j(\alpha)$, $j = 1, \ldots, J$, is a list of working correlation structures, then the hybrid GEE estimator of $\boldsymbol{\beta}$ is given by maximizing the empirical likelihood function $L(\boldsymbol{\beta}) = \Pi_{i=1}^N p_i$ subject to $\sum_{i=1}^N p_i = 1$ $(0 \le p_i \le 1)$ and

$$\sum_{i=1}^N p_i h_i(\boldsymbol{\beta}) = \begin{pmatrix} \sum_{i=1}^N p_i S_i^1(\boldsymbol{\beta}) \\ \sum_{i=1}^N p_i S_i^2(\boldsymbol{\beta}) \\ \vdots \\ \sum_{i=1}^N p_i S_i^J(\boldsymbol{\beta}) \end{pmatrix} = 0, \tag{1.2}$$

where $S_i^j(\boldsymbol{\beta}) = D_i^\tau A_i^{-1/2} R_j^{-1}(\alpha_j) A_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$. Then if one of the working correlation structures is correct, the hybrid GEE estimator is asymptotically efficient. Generally, it performs better than GEE methods using a single working correlation structure when the true correlation structure is unknown.

For small samples, however, our experiences suggest that the hybrid GEE method can be numerically unstable when some of the estimated correlation matrices are nearly singular, and its finite sample efficiency is sometimes unsatisfactory. The importance of the positive definiteness of the working correlation matrix to the efficiency of GEE estimates was reported by Sutradhar and Das (1999). Here, we propose an alternative method, called mix-GEE, by viewing the sample as from a finite mixture of distributions with different correlation structures. As a result, we obtain a single GEE with the working correlation matrix represented by a combination of a finite number of matrices, whereas the hybrid GEE method uses many GEEs with each given by a single working correlation structure. Under mild regularity conditions, we show that the mix-GEE estimator is consistent and asymptotically normal. It is asymptotically efficient if data are from a Gaussian mixture model and one of the mixture components gives the correct correlation structure. If either the AR(1) or exchangeable structure is included in the considered working correlation structures, the mix-GEE method guarantees that the estimated working correlation matrix is positive definite. This assures that it is numerically more stable and often displays a better finite sample efficiency than the hybrid GEE method. We also propose an iterative algorithm to solve the mix-GEE estimates that iterates between estimation of the nuisance parameters in the working correlation and estimation of the regression coefficients. Estimation of the nuisance parameters is solved by an EM-type algorithm using pseudo likelihood. A byproduct of our approach is that it can also serve as a tool for selecting the working correlation structures. For example, if estimates of some mixture proportions are close to 0, it indicates that the corresponding correlation structures should not be considered for the given data.

The rest of this paper is organized as follows. In Section 2, we propose the mix-GEE method and the computational algorithm, and then prove its asymptotic properties. Simulation studies and data examples are given in Section 3 and Section 4, respectively. Technical proofs are given in the Appendix.

2. The Mix-GEE Method

Our basic idea is to assume that $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^\tau$ is from an L-component mixture: $\boldsymbol{y}_i = z_{i1}\boldsymbol{y}_i^{(1)} + \cdots + z_{iL}\boldsymbol{y}_i^{(L)}$, where $\boldsymbol{y}_i^{(1)}, \ldots, \boldsymbol{y}_i^{(L)}$ are the L components, and the $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iL})^\tau$ are latent indicators that have only one component equal to one with the rest of its components equal to zero. Let $\pi_l = Pr(z_{il} = 1)$. Suppose that $\boldsymbol{y}_i^{(l)}$'s have the same variances but different correlation structures $Corr(\boldsymbol{y}_i^{(l)}) = Corr(\boldsymbol{y}_i | z_{il} = 1) = R^{(l)}(\alpha_l)$. The covariance matrix of $\boldsymbol{y}_i$ can be expressed as a weighted linear combination of L components,

$$Cov(\boldsymbol{y}_i) = A_i^{1/2}\left[\sum_{l=1}^{L}\pi_l R^{(l)}(\alpha_l)\right]A_i^{1/2}. \tag{2.1}$$

The mix-GEE estimate of $\boldsymbol{\beta}$ is then given by the solution to (1.1) with $R(\alpha) = \sum_{l=1}^{L}\pi_l R^{(l)}(\alpha_l)$. Direct computation of the mix-GEE estimates can be numerically challenging, and we use an iterative algorithm that iterates between estimating nuisance parameters $\alpha_1, \ldots, \alpha_L$ and $\pi_1, \ldots, \pi_L$ for given $\boldsymbol{\beta}$, and estimating $\boldsymbol{\beta}$ for given nuisance parameters.

2.1. Pseudo-likelihood estimation of nuisance parameters

Consider estimating the nuisance parameters given $\boldsymbol{\beta}$. Let $\boldsymbol{\psi} = (\alpha_l, \pi_l)_{l=1}^{L} \in \Psi$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\tau = A_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$. Parameter estimation in a mixture model has been well studied when the distribution of each component is known, and estimates are often obtained using EM algorithms. In our context, the distribution of $\boldsymbol{\varepsilon}_i$ is not specified and we consider a pseudo-likelihood (PL) approach. The PL approach has been considered by many researchers for longitudinal data or repeated measured data, for example Crowder (1985), Carroll and Ruppert (1988), Davidian and Giltinan (1995), and Sun, Shults, and Leonard (2009). Essentially, the nuisance parameters in the working correlation matrix can be estimated by maximizing a Gaussian log-likelihood. Let

$$\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l) = (1/\sqrt{2\pi})|R^{(l)}(\alpha_l)|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}_i{}^\tau[R^{(l)}(\alpha_l)]^{-1}\boldsymbol{\varepsilon}_i\right\}. \tag{2.2}$$

A pseudo-log-likelihood function based on the mixture model (2.1) is

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{N}\log\left(\sum_{l=1}^{L}\pi_l\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l)\right), \tag{2.3}$$

and the maximum PL estimate (MPLE) of $\boldsymbol{\psi}$ can be obtained by maximizing (2.3). We propose an PL-EM algorithm to solve this optimization problem, with each iteration consisting of an E-step and an M-step. A thorough discussion on EM algorithms can be found in McLachlan and Krishnan (2008).

Suppose $\boldsymbol{\psi}^{(t)}$ is the value of $\boldsymbol{\psi}$ at the t-th step. Then $\omega_{il}^{(t)} = \pi_l^{(t)}\phi_l(\boldsymbol{\varepsilon}_{\mathbf{i}}, \alpha_l^{(t)})/\sum_{l=1}^{L}\pi_l^{(t)}\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l^{(t)})$ is the current estimated posterior probability of $\boldsymbol{\varepsilon}_i$ belonging to the l-th component. The E-step is to compute the expectation of the complete-data pseudo-log-likelihood conditional on $\boldsymbol{\psi}^{(t)}$ and the observed data,

$$Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(t)}) = \sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(t)}\log[\pi_l\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l)].$$

In the M-step, we maximize $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(t)})$ with respect to $\boldsymbol{\psi}$ to get $\boldsymbol{\psi}^{(t+1)}$,

$$\hat{\pi}_l{}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N}\omega_{il}^{(t)} \text{ and } \hat{\alpha}_l{}^{(t+1)} = \arg\max_{\alpha_l \in \Upsilon_l}\sum_{i=1}^{N}\omega_{il}^{(t)}\log\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l), \tag{2.4}$$

where Υ_l is the region of α_l in which $R^{(l)}$ is positive definite. It is easy to show that, for AR(1), $\Upsilon_{AR} = \{\alpha | -1 < \alpha < 1\}$; for CS, $\Upsilon_{CS} = \{\alpha | -1/(T-1) < \alpha < 1\}$; and for MA(1), $\Upsilon_{MA} = \{\alpha | -1/[2\cos(\pi/(T+1))] \le \alpha \le -1/[2\cos(T\pi/(T+1))]\}$.

The PL-EM algorithm stops when the difference between $\boldsymbol{\psi}^{(t+1)}$ and $\boldsymbol{\psi}^{(t)}$ is sufficiently small, and the MPLE $\hat{\boldsymbol{\psi}} = (\hat{\alpha}_l, \hat{\pi}_l)_{l=1}^L$ is taken as the value at the last step.

Theorem 1. *In the PL-EM algorithm, $\ell(\boldsymbol{\psi}^{(t+1)}) > \ell(\boldsymbol{\psi}^{(t)})$ for every t.*

Theorem 1 shows that if the pseudo-likelihood function ℓ has a unique maximum, the PL-EM algorithm converges to this global maximum.

Theorem 2. *The estimated working correlation matrix $\sum_{l=1}^L \hat{\pi}_l R^{(l)}(\hat{\alpha}_l)$ is positive definite if at least one of $R^{(1)}, \ldots, R^{(L)}$ is AR(1) or exchangeable.*

Theorem 2 shows an important feature of our mix-GEE method. Most existing methods have no guarantee on the positive definiteness of the working correlation matrix, and this may cause serious loss of efficiency (Sutradhar and Das (1999)).

Next we show the consistency of the MPLE $\hat{\boldsymbol{\psi}}$ for given $\boldsymbol{\beta}$. We need some technical assumptions.

(i) $\boldsymbol{\varepsilon}_i (i = 1, \ldots, N)$ have a joint distribution function G with density g.

(ii) The support of $\boldsymbol{\psi}$, Ψ, is a compact subset of a Euclidean space.

(iii) $|\log\sum_{l=1}^L \pi_l\phi_l(u, \alpha_l)| \le M(u)$ for all $\boldsymbol{\psi} = (\pi_l, \alpha_l)_{l=1}^L \in \Psi$ and all $u \in \Omega$, where M is integrable with respect to G.

Theorem 3. *Let $KL(f, g) = \int f\log(f/g)$ denote Kullback-Leibler distance and suppose that $KL(g(\boldsymbol{\varepsilon}_i), \sum_{l=1}^L \pi_l\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l))$ has its unique minimum at $\boldsymbol{\psi}^* = (\pi_l^*, \alpha_l^*)_{l=1}^L$ for given $\boldsymbol{\beta}$. Under Assumptions (i), (ii) and (iii), the MPLE $\hat{\boldsymbol{\psi}} = (\hat{\alpha}_l, \hat{\pi}_l)_{l=1}^L$ converges to $\boldsymbol{\psi}^*$ almost surely as $N \to \infty$. Further, if $E[(\partial^2\ell(\boldsymbol{\psi}))/(\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau)]$ is continuous in Ψ, $E\{\partial\ell(\boldsymbol{\psi})/\partial\boldsymbol{\psi}\} = \partial E\{\ell(\boldsymbol{\psi})\}/\partial\boldsymbol{\psi}$, and there is a neighborhood $\mathcal{N}$ of $\boldsymbol{\psi}$ such that $E[\sup_{\boldsymbol{\psi}\in\mathcal{N}}\partial^2\ell(\boldsymbol{\psi})/(\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau)] < \infty$ and $E[(\partial\log\sum_{l=1}^L \pi_l\phi_l(\varepsilon_{\mathbf{i}}, \alpha_l)/\partial\boldsymbol{\psi})(\partial\log\sum_{l=1}^L \pi_l\phi_l(\varepsilon_{\mathbf{i}}, \alpha_l)/\partial\boldsymbol{\psi})^T\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}] < \infty$, then $\sqrt{N}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) = O_p(1)$.*

Since $|\log\sum_{l=1}^{L}\pi_l\phi_l(u,\alpha_l)| \le |\log[\max\phi_l(u,\alpha_l)]| + |\log[\min\phi_l(u,\alpha_l)]| \le \sum_{l=1}^{L}|\log\phi_l(u,\alpha_l)|$, (iii) is generally satisfied. More generally, even when G is a non-normal distribution, as long as it has a finite fourth moment, the assumptions of Theorem 3 are satisfied.

Note that $\boldsymbol{\psi}^*$ and $\hat{\boldsymbol{\psi}}$ are actually functions of $\boldsymbol{\beta}$, $\boldsymbol{\psi}^* = \boldsymbol{\psi}^*(\beta), \hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\psi}}(\beta)$. Given $\boldsymbol{\beta}$, Theorem 3 implies that the estimated working correlation matrix based on the MPLE converges to some constant matrix. Results in Theorem 3 are similar to those in White (1982) regarding properties of maximum likelihood estimators under model misspecification. The next theorem has it that if the responses are from a multivariate Gaussian mixture distribution, $\hat{\boldsymbol{\psi}}$ converges to the true parameter in the true correlation matrix.

Theorem 4. *If $g(\boldsymbol{\varepsilon}_i) = \sum_{l=1}^{L}\pi_l^0\phi_l(\boldsymbol{\varepsilon}_i,\alpha_l^0)$, where ϕ_l is defined in (2.2), then $\boldsymbol{\psi}^* = \boldsymbol{\psi}^0$, where $\boldsymbol{\psi}^0 = (\pi_l^0,\alpha_l^0)_{l=1}^{L}$.*

2.2. Mix-GEE estimation of regression parameters

Given estimated nuisance parameters $\boldsymbol{\psi}$ we can estimate regression parameters $\boldsymbol{\beta}$ by plugging the estimated working correlation matrix into (1.1), leading to the following iterative algorithm.

1. Obtain an initial estimate $\hat{\boldsymbol{\beta}}^0$ using a $N^{1/2}$-consistent estimator (for example, the GEE estimator based on the independent working correlation structure).
2. At the m-th step, compute $\hat{\boldsymbol{\varepsilon}}_i^m = A_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{m-1}}$, where $\hat{\boldsymbol{\beta}}^{m-1}$ is the estimation of $\boldsymbol{\beta}$ at the $(m-1)$-th step.
3. Apply the PL-EM algorithm in Section 2.1 to obtain the MPLE $(\hat{\pi}_l^m,\hat{\alpha}_l^m)_{l=1}^{L}$.
4. Let $R^m = \sum_{l=1}^{L}\hat{\pi}_l^m R^{(l)}(\hat{\alpha}_l^m)$. Compute $\hat{\boldsymbol{\beta}}^m$ by solving the equation $U(\boldsymbol{\beta}, R^m) = 0$ given in (2.5).

Repeating steps 2−4 until the algorithm converges, we denote the final estimate by $\hat{\boldsymbol{\beta}}^M$ and call it the mix-GEE estimator. It follows immediately from Lemma E.4 in Appendix E that for, any integers k and m, the difference between updates $\hat{\boldsymbol{\beta}}^{k+m}$ and $\hat{\boldsymbol{\beta}}^m$, vanishes in probability as N tends to infinity. If $U_i(\boldsymbol{\beta}, M) = D_i^{\tau}A_i^{-1/2}MA_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ $(i = 1,\ldots,N)$, then $\hat{\boldsymbol{\beta}}^M$ is the solution to

$$U(\boldsymbol{\beta}, M) = \sum_{i=1}^{N}U_i(\boldsymbol{\beta}, M) = \sum_{i=1}^{N}D_i^{\tau}A_i^{-1/2}M^{-1}A_i^{-1/2}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0, \tag{2.5}$$

where $M = M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta})) = \sum_{l=1}^{L}\hat{\pi}_l R^{(l)}(\hat{\alpha}_l)$. The next theorem gives the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}^M$ under some assumptions on $U_i(\boldsymbol{\beta}, M)$.

Theorem 5. *Assume the following.*

1. *There exists* $H(\boldsymbol{Y},\boldsymbol{\beta}) = O_p(1)$ *such that* $|\partial\hat{\boldsymbol{\psi}}/\partial\boldsymbol{\beta}| \le H(\boldsymbol{Y},\boldsymbol{\beta}) = O_p(1)$, *where* $\boldsymbol{Y} = \{\boldsymbol{y}_i, i = 1,\ldots,N\}$.
2. $\partial^2 U_i(\boldsymbol{\beta}, M(\boldsymbol{\psi}))/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau$ *is continuous at* $\boldsymbol{\psi}^*$ *with probability one, and there is a neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\psi}$ *such that* $E\left[\sup_{\boldsymbol{\psi}\in\mathcal{N}} \partial^2 U_i(\boldsymbol{\beta}, M(\boldsymbol{\psi}))/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau\right] < \infty$.
3. *There is a neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\psi}^*$ *such that* $\left[(1/N)\sum_{i=1}^N \partial U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}/\partial\boldsymbol{\beta}\right]^{-1}$ *is bounded in* $\mathcal{N}$ *with probability one.*
4. $E[U_i(\boldsymbol{\beta}, M(\boldsymbol{\psi}^*))] = 0$ *has a unique solution* $\boldsymbol{\beta}_0$.
5. *The support of* $\boldsymbol{\beta}$ *is compact.*
6. $[\partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau]$ *is continuous at* $\boldsymbol{\beta}_0$ *with probability one, and there is a neighborhood* $\mathcal{N}$ *of* $\boldsymbol{\beta}$ *such that* $E[\sup_{\boldsymbol{\beta}\in\mathcal{N}} \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau] < \infty$.

If $R^* = \sum_{l=1}^L \pi_l^* R^{(l)}(\alpha_l^*)$ *and* $\tilde{R}$ *is the true correlation matrix of* $\boldsymbol{y}_i$, *then* $N^{1/2}(\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}) \xrightarrow{d} N(0, V^M)$ *with*

$$V^M = \lim_{N\to\infty} N\Big(\sum_{i=1}^N H_i\Big)^{-1}\Big(\sum_{i=1}^N G_i\Big)\Big(\sum_{i=1}^N H_i\Big)^{-1}, \tag{2.6}$$

where $H_i = {D_i}^\tau A_i^{-1/2}(R^*)^{-1}A_i^{-1/2}D_i$, $G_i = {D_i}^\tau A_i^{-1/2}(R^*)^{-1}\tilde{R}(R^*)^{-1}A_i^{-1/2}D_i$.

The conditions of Theorem 5 are mild. We provide some concrete examples satisfying the conditions in a supplementary file at `http://www.math.wustl.edu/$\sim$nlin/mixGEEsup.pdf`.

To simplify notation, we put $\boldsymbol{\psi}^* \equiv \boldsymbol{\psi}^*(\beta_0)$ in Theorem 5. If the data are indeed from a Gaussian mixture model, Theorem 4 implies that R^* is the true correlation matrix $\tilde{R} = \sum_{l=1}^L \pi_l^0 R^{(l)}(\alpha_l^0)$. Consequently, in Theorem 5, we have

$$V^M = \lim_{N\to\infty}\Big(\frac{1}{N}\sum_{i=1}^N {D_i}^\tau A_i^{-1/2}(\tilde{R})^{-1}A_i^{-1/2}D_i\Big)^{-1},$$

which shows that $\hat{\boldsymbol{\beta}}^M$ is asymptotically efficient (Chaganty and Joe (2004)).

3. Simulation Studies

In this section, we compare our mix-GEE method with the hybrid GEE method (Leung, Wang, and Zhu (2009)) through simulation studies. As reference, we also include the maximum likelihood estimator (MLE) and two other GEE estimators that use a single prespecified correlation structure. One is the 'independent' GEE estimator using the independent correlation matrix with all off diagonal entries zero, the other is the 'PL-GEE' (Crowder (1985); Carroll and Ruppert (1988); Davidian and Giltinan (1995); Sun, Shults, and Leonard (2009)) that is based on the classic GEE method with nuisance parameters in an assumed single correlation structure estimated by pseudo-likelihood. For PL-GEE, we consider the correlation structures CS, AR(1), and MA(1). The PL-GEE estimator is efficient when the working correlation structure is correctly specified and its efficiency can be seriously affected under misspecification.

In each simulation study, 1,000 Monte Carlo samples were generated.

3.1. Continuous response

We generated the response y_{ik} at time k for the i th subject from $N(\mu_{ik} = \beta_0 + x_{ik}\beta_1, 1)$, $k = 1, \ldots, 10$, $i = 1, \ldots, n$. The true value of (β_0, β_1) was (1, -1); covariates x_{ik}'s were randomly sampled from $N(k, 1)$; the sample size n was 20, 50, or 100.

We compare the performance of several estimation methods according to their simulated relative efficiency (SRE) under different correlation structures. Let $\hat{\boldsymbol{\beta}}^{true}$ be the GEE estimator obtained from plugging the true correlation matrix into (1.1). The SRE is taken as the ratio of the sample mean squared error (MSE) of an estimator and that of $\hat{\boldsymbol{\beta}}^{true}$. For an estimator $(\hat{\beta}_0, \hat{\beta}_1)$, its sample MSE is obtained by averaging $(\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 - \beta_1)^2$ over all Monte Carlo samples.

3.1.1. Three component mixture

Consider the case in which the true correlation is given by a three component mixture of AR(1), CS, and MA(1). Let $R^{(1)}(\alpha_1) = R_{AR(1)}(\alpha_1)$, $R^{(2)}(\alpha_2) = R_{CS}(\alpha_2)$ and $R^{(3)}(\alpha_3) = R_{MA(1)}(\alpha_3)$. The true values of the correlation parameter are $\alpha_1, \alpha_2, \alpha_3$ and the relative component proportion π_1, π_2, π_3 are given in Table 1. The PL-GEE method uses a single correlation structure, and we consider PL-GEE estimators with correlation structure CS, AR(1), and MA(1), respectively. In Table 1, 'Ind', 'hybrid' and 'mix' refer to the independent GEE estimator, the hybrid GEE estimator, and the mix-GEE estimator, respectively. The estimated component proportions from the mix-GEE method are given by $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$. Numbers in brackets are mean squared errors (MSEs) of the mix-GEE estimator. MSEs of other estimators can then be inferred using their SREs. Under this setup, independent GEE and PL-GEE both use misspecified correlation structures, so we can see that their efficiency was always lower than the mix-GEE estimator. Surprisingly, for small samples, the hybrid GEE estimator performed even worse than the independent GEE and the PL-GEE estimator even when a misspecified correlation structure was used. Our experience suggests that the hybrid GEE estimator requires a very large sample to achieve the claimed asymptotic efficiency. Across different finite mixture correlation models,

Table 1. Comparison for continuous data generated from three-component mixture distributions.

	true parameter						SRE							mixture proportion		
n	α_1	α_2	α_3	π_1	π_2	π_3	Ind	PL_{CS}	PL_{AR1}	PL_{MA1}	hybrid	MLE	mix (MSE)	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$
5	0.7	0.7	0.4	0.3	0.3	0.4	1.1033	1.0925	1.0380	1.0334	7.8143	0.9876	0.9969(0.1450)	0.334	0.334	0.332
5	0.4	0.4	0.4	0.3	0.3	0.4	1.0506	1.0423	1.0186	1.0053	1.2426	1.0566	1.0081(0.1481)	0.257	0.347	0.396
10	0.7	0.7	0.4	0.3	0.3	0.4	1.0669	1.0613	1.0324	1.0113	2.1068	0.9070	1.0049(0.0773)	0.345	0.315	0.340
10	0.4	0.4	0.4	0.3	0.3	0.4	1.0510	1.0467	1.0306	1.0249	1.7227	1.0283	1.0146(0.0668)	0.288	0.310	0.402
20	0.7	0.7	0.4	0.3	0.3	0.4	1.1037	1.0852	1.0309	1.0268	1.4988	0.8552	1.0043(0.0390)	0.338	0.311	0.351
20	0.4	0.4	0.4	0.3	0.3	0.4	1.0692	1.0629	1.0033	1.0038	1.5312	0.9826	1.0010(0.0303)	0.302	0.301	0.394
20	0.7	0.7	0.4	0.5	0.5	0	1.1132	1.1071	1.0487	1.0562	1.4355	0.8759	1.0023(0.0474)	0.460	0.498	0.042
50	0.7	0.7	0.4	0.3	0.3	0.4	1.0597	1.0555	1.0382	1.0239	1.1850	0.8054	1.0034(0.0148)	0.325	0.301	0.374
50	0.4	0.4	0.4	0.3	0.3	0.4	1.0511	1.0515	1.0077	1.0004	1.0822	0.9417	1.0030(0.0139)	0.305	0.282	0.413
50	0.7	0.7	0.4	0.5	0.5	0	1.1362	1.1166	1.0443	1.0660	1.1893	0.8495	0.9974(0.0168)	0.484	0.493	0.123
50	0.7	0.7	0.4	0.2	0.6	0.2	1.0315	1.0283	1.0752	1.0218	1.1137	0.8126	1.0046(0.0174)	0.220	0.600	0.180
100	0.7	0.7	0.4	0.3	0.3	0.4	1.0798	1.0836	1.0196	1.0149	1.0561	0.7888	1.0005(0.0081)	0.314	0.300	0.386
100	0.4	0.4	0.4	0.3	0.3	0.4	1.0587	1.0485	1.0205	1.0043	1.0802	0.9624	1.0017(0.0064)	0.302	0.293	0.405
5	0.7	0.7	0.4	0.2	0	0.8	1.1471	1.1492	1.0337	1.0419	1.3395	1.0759	1.0474(0.1370)	0.270	0.136	0.594
10	0.7	0.7	0.4	0.2	0	0.8	1.1816	1.1748	1.0064	1.0155	1.9327	1.0142	1.0028(0.0643)	0.278	0.087	0.635
20	0.7	0.7	0.4	0.2	0	0.8	1.1173	1.1241	1.0365	0.9944	1.5203	0.9803	1.0013(0.0353)	0.272	0.055	0.673
50	0.7	0.7	0.4	0.2	0	0.8	1.1309	1.1334	1.0071	1.0111	1.1322	0.9950	1.0064(0.0142)	0.247	0.037	0.716
100	0.7	0.7	0.4	0.2	0	0.8	1.0880	1.0939	1.0284	0.9926	1.0769	0.9813	1.0014(0.0061)	0.233	0.023	0.744
5	0.4	0.4	0.4	0.2	0	0.8	1.1150	1.1173	1.0375	1.0350	1.3727	1.0923	1.0430(0.1287)	0.223	0.1520	0.625
10	0.4	0.4	0.4	0.2	0	0.8	1.1514	1.1485	1.0092	1.0163	1.7506	1.0453	1.0107(0.0605)	0.240	0.091	0.669
20	0.4	0.4	0.4	0.2	0	0.8	1.0862	1.0893	1.0171	0.9981	1.4223	1.0164	0.9971(0.0310)	0.234	0.059	0.707
50	0.4	0.4	0.4	0.2	0	0.8	1.1176	1.1203	1.0275	0.9991	1.1607	1.0081	1.0008(0.0128)	0.217	0.033	0.75
100	0.4	0.4	0.4	0.2	0	0.8	1.1360	1.1380	1.0242	1.0003	1.0404	0.9989	1.0008(0.0068)	0.222	0.024	0.754

the SRE of the mix-GEE estimator was always the smallest and close to 1. Meanwhile, we also see that the component proportions are consistently estimated by the mix-GEE method.

3.1.2. Single component correlation structure

Here we consider data from a single correlation structure, either AR(1), CS or MA(1). For each, we consider different values of the nuisance parameter α. In Table 2, the first column gives the true correlation structure while the other columns are defined similarly as in Table 1.

Table 2 has the PL-GEE estimators as most efficient if the prespecified correlation structure is correct, but they have low efficiency under misspecification. On the other hand, the mix-GEE estimator has SREs close to 1 under all scenarios, and its efficiency is close to the PL-GEE estimator when the correlation structure is correctly specified. Furthermore, similarly as in Table 1, the mix-GEE estimator always gave smaller MSEs than the hybrid GEE and the independent GEE. And most of the time, the estimated mixture proportions given by the mix-GEE method correctly identified the true correlation structure, except when the true correlation matrix is difficult to identify. For example, $0.2AR(0.4) + 0.8MA(0.4)$ is very close to $MA(0.4)$.

Table 2. Comparison for continuous data generated from a single correlation structure.

			SRE							mixture proportion		
$R(\alpha)$	n	α	Ind	PL_{CS}	PL_{AR1}	PL_{MA1}	hybrid	MLE	mix (MSE)	$\hat{\pi}_{ar1}$	$\hat{\pi}_{cs}$	$\hat{\pi}_{ma1}$
CS	10	0.7	1.0285	1.0000	1.1814	1.0611	2.2181	1.0000	1.0034(0.0832)	0.04	0.95	0.01
CS	10	0.4	1.0202	1.0004	1.0909	1.0471	1.7439	1.0004	1.0027(0.0717)	0.08	0.83	0.09
AR	10	0.7	1.3490	1.3469	0.9951	1.1259	2.3519	0.9951	1.0034(0.0786)	0.85	0.08	0.07
AR	10	0.4	1.0977	1.1003	1.0133	1.0213	1.6165	1.0133	1.0252(0.0750)	0.48	0.16	0.36
AR	10	-0.4	1.0814	1.0870	1.0058	1.0007	1.6589	1.0058	1.0071(0.0229)	0.58	0.09	0.33
AR	10	-0.7	1.6467	1.7952	0.9999	1.0490	2.0795	0.9999	1.0046(0.0097)	0.91	0.01	0.08
MA	10	0.4	1.0644	1.0739	1.0103	1.0183	1.7444	1.0183	1.0206(0.0590	0.17	0.08	0.75
MA	10	-0.4	1.2702	1.2682	1.0393	1.0056	1.6860	1.0056	1.0084(0.0152)	0.20	0.09	0.81
CS	50	0.7	1.0153	1.0000	1.1646	1.0379	1.1464	1.0000	1.0003(0.0177)	0.01	0.98	0.01
CS	50	0.4	1.0159	0.9998	1.0876	1.0403	1.0892	0.9998	1.0015(0.0140)	0.03	0.94	0.03
AR	50	0.7	1.4096	1.3949	1.0026	1.1385	1.1067	1.0026	1.0077(0.0162)	0.94	0.03	0.03
AR	50	0.4	1.1059	1.1108	0.9969	1.0195	1.1159	0.9969	0.9997(0.0135)	0.65	0.08	0.27
AR	50	-0.4	1.1292	1.1168	1.0018	1.0091	1.1465	1.0018	1.0033(0.0047)	0.72	0.04	0.24
AR	50	-0.7	1.5913	1.6796	1.0004	1.0705	1.1561	1.0004	1.0042(0.0020)	0.97	0.00	0.03
MA	50	0.4	1.1333	1.1337	1.0193	1.0019	1.1189	1.0019	1.0019(0.0124)	0.13	0.20	0.67
MA	50	-0.4	1.2469	1.2625	1.0346	1.0031	1.0971	1.0031	1.0042(0.0029)	0.15	0.03	0.82
CS	100	0.7	1.0348	1.0000	1.2559	1.1030	1.0339	1.0000	1.0005(0.0081)	0.01	0.99	0.00
CS	100	0.4	1.0226	0.9998	1.0895	1.0411	1.0744	0.9998	1.0011(0.0063)	0.02	0.96	0.02
AR	100	0.7	1.3061	1.3048	1.0003	1.0958	1.0292	1.0003	0.9992(0.0087)	0.96	0.02	0.02
AR	100	0.4	1.1258	1.1287	1.0043	1.0280	1.0768	1.0043	1.0079(0.0069)	0.69	0.06	0.25
AR	100	-0.4	1.1161	1.1284	1.0014	1.0010	1.0554	1.0014	1.0008(0.0023)	0.75	0.03	0.22
AR	100	-0.7	1.5308	1.5800	0.9992	1.0512	1.0666	0.9992	0.9994(0.0010)	0.98	0.00	0.02
MA	100	0.4	1.1431	1.1390	1.0282	0.9996	1.0176	0.9996	0.9992(0.0060)	0.11	0.01	0.88
MA	100	-0.4	1.2590	1.2656	1.0595	0.9990	1.0456	0.9990	1.0036(0.0016)	0.12	0.02	0.86

3.2. Binary response

3.2.1. Probit model

We generated the binary responses from the probit regression model $y_{it} = I[z_{it} \leq (\beta_0 + \beta_1 x_{it})]$, where the $Z_i = (z_{i1}, \ldots, z_{iT})$ are multivariate normal with mean 0, variance 1, and correlation structure CS, MA(1), or AR(1). The covariates x_{it} were randomly sampled from $U(-1,1)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. We considered $T = 4$ and $N = 50, 100$ or 200. The true values were set at $(\beta_0, \beta_1) = (0, 0.5)$.

For this setup, the explicit form of the true correlation structure of the responses y_{it}'s is difficult to obtain even though the correlation structure of the latent variable z_{it}'s is known. So, as in Chaganty and Joe (2004), we compared all estimators to the MLE for the case that the latent variable is observed, and obtained their SRE as the ratio between the sample MSE of an estimator and that of the MLE. Table 3 again has the mix-GEE estimator with high efficiency

Table 3. Comparison for binary data from probit model.

			SRE						mixture proportion		
$R(\alpha)$	n	α	Ind	PL_{CS}	PL_{AR1}	PL_{MA1}	hybrid	mix (MSE)	$\hat{\pi}_{ar1}$	$\hat{\pi}_{cs}$	$\hat{\pi}_{ma1}$
CS	50	0.7	1.2333	1.0285	1.0904	1.1226	1.1560	1.0304(0.0392)	0.162	0.466	0.372
CS	50	0.9	1.4330	1.0927	1.2079	1.2551	1.2273	1.1030(0.0426)	0.073	0.596	0.331
AR	50	0.7	1.2023	1.0638	1.0282	1.0583	1.1295	1.0321(0.0366)	0.279	0.295	0.426
AR	50	-0.7	1.2940	1.1887	1.0310	1.0795	1.1780	1.0322(0.0256)	0.367	0.226	0.407
AR	50	0.9	1.3378	1.0856	1.0955	1.1355	1.1734	1.0794(0.0433)	0.111	0.515	0.374
MA	50	-0.3	1.0567	1.0280	1.0004	0.9943	1.1167	1.0163(0.0314)	0.283	0.354	0.363
MA	50	0.3	1.0211	1.0007	0.9932	0.9981	1.1467	0.9994(0.0360)	0.305	0.311	0.384
AR	100	0.7	1.2311	1.0566	1.0265	1.0583	1.0811	1.0272(0.0187)	0.284	0.276	0.440
CS	100	0.7	1.2364	1.0196	1.0881	1.1188	1.0648	1.0229(0.0201)	0.146	0.475	0.379
MA	100	0.3	1.0230	1.0137	0.9905	0.9893	1.0401	0.9914(0.0171)	0.313	0.306	0.381
CS	200	0.7	1.2125	1.0034	1.0632	1.0877	1.0222	1.0028(0.0101)	0.147	0.481	0.372
AR	200	0.7	1.2294	1.0326	1.0019	1.0308	1.0130	1.0014(0.0087)	0.290	0.248	0.462
MA	200	0.3	1.0149	0.9963	0.9731	0.9759	1.0003	0.9740(0.0085)	0.318	0.306	0.376

across different correlation models, whereas the hybrid GEE estimator is always less efficient.

4. Applications to Data

4.1. Six cities data set: Children's wheezing status

In this section, we apply our mix-GEE method to a respiratory infection data set from Ware et al. (1984). It has been previously analyzed by Fitzmaurice and Laird (1993) and Chaganty and Joe (2004). In this, a group of 537 children from Steubenville, was examined annually at ages 7 through 10 to study the relationship between health and air pollution. The response variable is a binary variable indicating a child's wheezing status. We fit a probit regression model using the three predictors *Age*, with value -2,-1,0,1 at ages 7,8,9 and 10; *Maternal*, with value 1 if the mother smoked regularly at the first year of study and 0 if she did not smoke; *Age*Maternal* indicates the interaction effect between *Age* and *Maternal*. In addition to our mix-GEE estimator, we also computed the independent GEE estimator, the PL-GEE estimator with working correlation structure CS, AR(1), or MA(1), the hybrid GEE estimator, and the MLE based on multivariate probit model. For the hybrid-GEE and the mix-GEE method, the correlation structures $R^1(\alpha_1) = AR(1)$, $R^2(\alpha_2) = CS$, and $R^3(\alpha_3) = MA(1)$ were considered. Results are shown in Table 4. Numbers in brackets are standard errors of the estimator. Estimates from all methods are quite similar. It is worth noting that the component proportion estimates from our mix-GEE estimator are $\hat{\pi}_{AR} = 0.062$, $\hat{\pi}_{cs} = 0.694$, and $\hat{\pi}_{ma} = 0.244$, which indicates a preference to the exchangeable structure. Interestingly, Chaganty and Joe (2004) also reported that the pattern of dependence is closer to exchangeable than to AR(1).

Table 4. Parameter estimates for Child's wheeze status data.

parameter	ind	PL_{cs}	PL_{ar1}	PL_{ma1}	hybrid	MLE	mix
Intercept	-1.1259	-0.1258	-1.1387	-1.1391	-1.1288	-1.1192	-1.1264
	(0.0634)	(0.0634)	(0.0637)	(0.0637)	(0.0684)	(0.0611)	(0.0635)
Age	-0.0768	-0.0768	-0.0805	-0.0858	-0.0783	-0.0781	-0.0771
	(0.0313)	(0.0313)	(0.0318)	(0.0321)	(0.0340)	(0.0302)	(0.0313)
Maternal	0.1709	0.1708	0.1561	0.1653	0.1708	0.1611	0.1695
	(0.1016)	(0.1028)	(0.1035)	(0.1032)	(0.1115)	(0.1001)	(0.1029)
Age* Maternal	0.0367	0.0367	0.0438	0.0526	0.0392	0.0382	0.0373
	(0.0487)	(0.0486)	(0.0496)	(0.0501)	(0.0531)	(0.0491)	(0.0486)

Table 5. Parameter estimates for the Sitka spruce data.

parameter	independent	PL_{cs}	PL_{ar1}	PL_{ma1}	GEE_{EL}	GEE_{mix}
Intercept	-8.738	-8.736	-8.545	-8.115	-7.826	-8.505
	(0.402)	(0.417)	(0.408)	(0.408)	(0.359)	(0.407)
ozone	-0.215	-0.215	-0.232	-0.216	-0.498	-0.228
	(0.158)	(0.157)	(0.162)	(0.168)	(0.149)	(0.161)
times	2.593	2.592	2.545	2.444	2.424	2.539
	(0.068)	(0.070)	(0.067)	(0.067)	(0.061)	(0.067)

4.2. Growth of Sitka spruces

We consider the Sitka spruce data set, from Diggle et al. (2002), in which 79 spruces were observed at 5 time points. Among them, 54 trees were grown in ozone-enriched chambers and 25 were grown in the general environment. The experimental design was balanced with the same number of trees at each time point. The responses y_{ij}'s were computed as $y_{ij} = \log(\text{height}_{ij} \times \text{diameter}^2_{ij})$ to denote the growth status of the trees. The explanatory variables are the ozone environment indicator (0 = normal, 1=ozone) and log(observation day time), where observation day time is the number of days from the start of the study.

We fit this data using the linear model $E(y_{ij}) = \beta_0 + \beta_1 ozone_{ij} + \beta_2 times_{ij}$. The same set of GEE estimators as in Section 4.1 was considered with results given in Table 5. Here, the component proportion estimates were $\hat{\pi}_{AR} = 0.96$, $\hat{\pi}_{cs} = 0.03$ and $\hat{\pi}_{ma} = 0.01$, Thus a strong indication that a single AR(1) correlation structure is perhaps sufficient. The mix-GEE estimates are very similar to the naive GEE estimates with the AR(1) working correlation.

Acknowledgement

We are grateful to the Editor, an associate editor, and a referee for comments leading to an improved presentation. This work was partially supported by the NSF of China (No.10871037 and No.10931002), NCET of China (NCET-09-0284), the PhD Programs Foundation of Ministry of Education of China (20100043110002), and NSF grant DMS0906023.

Appendix

Appendix A

Proof of Theorem 1. According to the M-step (2.4), we have $Q(\boldsymbol{\psi}^{(t+1)}, \boldsymbol{\psi}^{(t)}) > Q(\boldsymbol{\psi}^{(t)}, \boldsymbol{\psi}^{(t)})$, or $\sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(t)}\log\pi_l^{(t+1)}\phi_l(\varepsilon_i, \alpha_l^{(t+1)}) > \sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(t)}\log\pi_l^{(t)}$ $\phi_l(\varepsilon_i, \alpha_l^{(t)})$. Let $H(\omega^{(t)}, \boldsymbol{\psi}^{(t+1)}) = Q(\boldsymbol{\psi}^{(t+1)}, \boldsymbol{\psi}^{(t)}) - \ell(\boldsymbol{\psi}^{(t+1)})$ and $H(\omega^{(t)}, \boldsymbol{\psi}^{(t)}) = Q(\boldsymbol{\psi}^{(t)}, \boldsymbol{\psi}^{(t)}) - \ell(\boldsymbol{\psi}^{(t)})$. It is easy to show that

$$H(\omega^{(t)}, \boldsymbol{\psi}^{(t)}) - H(\omega^{(t)}, \boldsymbol{\psi}^{(t+1)}) = \sum_{i=1}^{N}\sum_{l=1}^{L}\omega_{il}^{(t)}\log\left\{\frac{\omega_{il}^{(t)}}{\omega_{il}^{(t+1)}}\right\}.$$

Since $\sum_{l=1}^{L}\omega_{il}^{(t+1)} = 1$, it follows from Jensen's inequality that

$$\sum_{l=1}^{L}\omega_{il}^{(t)}\log\left\{\frac{\omega_{il}^{(t+1)}}{\omega_{il}^{(t)}}\right\} \le \log\sum_{l=1}^{L}\omega_{il}^{(t)}\cdot\frac{\omega_{il}^{(t+1)}}{\omega_{il}^{(t)}} = \log\sum_{l=1}^{L}\omega_{il}^{(t+1)} = 0.$$

Hence, $H(\omega^{(t)}, \boldsymbol{\psi}^{(t)}) > H(\omega^{(t)}, \boldsymbol{\psi}^{(t+1)})$ and it then follows that the change from $\boldsymbol{\psi}^{(t+1)}$ to $\boldsymbol{\psi}^{(t)}$ increases the pseudo-likelihood ℓ.

Appendix B

Proof of Theorem 2. In (2.4), let

$$\begin{aligned} f(\alpha_l, \boldsymbol{\psi}^{(t)}) &= \sum_{i=1}^{N}\omega_{il}^{(t)}\log\phi_l(\boldsymbol{\varepsilon}_i, \alpha_l) \\ &= -\frac{1}{2}\sum_{i=1}^{N}\omega_{il}^{(t)}\log(|R^{(l)}(\alpha_l)|) - \frac{1}{2}\sum_{i=1}^{N}\omega_{il}^{(t)}\boldsymbol{\varepsilon}_i^{\tau}\{R^{(l)}(\alpha_l)\}^{-1}\boldsymbol{\varepsilon}_i. \end{aligned}$$

Then, we have

$$\begin{aligned} &\frac{\partial f(\alpha_l, \boldsymbol{\psi}^{(t)})}{\partial\alpha_l}\Big|_{\alpha_l=\hat{\alpha_l}^{(t+1)}} \\ &= \sum_{i=1}^{N}\omega_{il}^{(t)} tr\left(\frac{\partial(R^{(l)}(\alpha_l))^{-1}}{\partial\alpha_l}\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^{\tau} - R^{(l)}(\alpha_l)\frac{\partial(R^{(l)}(\alpha_l))^{-1}}{\partial\alpha_l}\right)\Big|_{\alpha_l=\hat{\alpha_l}^{(t+1)}} = 0. \end{aligned}$$

(i) Suppose one of the $R^{(l)}(\alpha_l)$ is $AR(1)$. And without loss of generality, let $R^{(1)}(\alpha_1) = R_{AR}(\alpha_1)$. Then the derivative of f can be written as

$$\begin{aligned}\frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1} &= \sum_{i=1}^{N} \omega_{il}^{(t)} tr\left[\frac{\partial (R_{AR}(\alpha_1))^{-1}}{\partial \alpha_1}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\tau} - R_{AR}(\alpha_1))\right] \\ &\propto (T-1)\sum_{i=1}^{N} \omega_{il}^{(t)} \alpha_1 (1-\alpha_1^2) \\ &\quad - \sum_{i=1}^{N} \omega_{il}^{(t)} \left\{\alpha_1 \left(\sum_{j=1}^{T-1} \varepsilon_{ij}^2 + \sum_{j=2}^{T} \varepsilon_{ij}^2\right) - (1+\alpha_1^2) \sum_{j=1}^{T-1} \varepsilon_{ij}\varepsilon_{ij+1}\right\},\end{aligned}$$

a cubic polynomial of α_1. It is easy to show that

$$\lim_{\alpha_1 \to -\infty} \frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1} = -\infty, \quad \lim_{\alpha_1 \to +\infty} \frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1} = +\infty,$$
$$\left.\frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1}\right|_{\alpha=-1} = \sum_{i=1}^{N} \omega_{il}^{(t)} [\sum_{j=1}^{T-1} (\varepsilon_{ij} + \varepsilon_{ij+1})^2] > 0,$$

and

$$\left.\frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1}\right|_{\alpha=1} = -\sum_{i=1}^{N} \omega_{il}^{(t)} [\sum_{j=1}^{T-1} (\varepsilon_{ij} - \varepsilon_{ij+1})^2] < 0.$$

Hence $\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})/\partial \alpha_1 = 0$ has a unique solution in $(-1, 1)$, so $\hat{\alpha_1}^{(t+1)} \in (-1, 1)$ and $R_{AR}(\hat{\alpha}_1)$ is positive definite. Then because $\hat{\pi}_1 > 0$, we have $\sum_{l=1}^{L} \hat{\pi}_l R^{(l)}(\hat{\alpha}_l)$ $= \hat{\pi}_1 R_{AR}(\hat{\alpha}_1) + \hat{\pi}_2 R_{(2)}(\hat{\alpha}_2) + \cdots + \hat{\pi}_L R_{(L)}(\hat{\alpha}_L)$ is a positive-definite matrix.
(ii) Suppose one of the $R^{(l)}(\alpha_l)$ is CS. Let $R^{(1)}(\alpha_1) = R_{CS}(\alpha_1)$. We have

$$\begin{aligned}&\frac{\partial f(\alpha_1, \boldsymbol{\psi}^{(t)})}{\partial \alpha_1} \\ &\quad = \sum_{i=1}^{N} \omega_{il}^{(t)} tr\left[\frac{\partial (R_{CS}(\alpha_1))^{-1}}{\partial \alpha_1}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{\tau} - R_{CS}(\alpha_1))\right] \\ &\quad \propto \sum_{i=1}^{N} \omega_{il}^{(t)} \left\{\alpha_1 [(T-2)\alpha_1 + 2](T-1) \sum_{j=1}^{n} \varepsilon_{ij}^2 - [1 + (T-1)\alpha_1^2] \sum_{j \neq h}^{T} \varepsilon_{ij}\varepsilon_{ih}\right\} \\ &\qquad - \sum_{i=1}^{N} \omega_{il}^{(t)} \left\{T(T-1)\alpha_1 [(T-2)\alpha_1 + 2] - [1 + (T-1)\alpha_1^2] T(T-1)\alpha_1\right\}.\end{aligned}$$

Then we can show that

$$\lim_{\alpha_1\to-\infty}\frac{\partial f(\alpha_1,\boldsymbol{\psi}^{(t)})}{\partial\alpha_1}=+\infty,\ \lim_{\alpha_1\to+\infty}\frac{\partial f(\alpha_1,\boldsymbol{\psi}^{(t)})}{\partial\alpha_1}=-\infty,$$

$$\frac{\partial f(\alpha_1,\boldsymbol{\psi}^{(t)})}{\partial\alpha_1}\bigg|_{\alpha=1}=\sum_{i=1}^{N}\omega_{il}^{(t)}T\Big\{(T-2)\sum_{j=1}^{T}\varepsilon_{ij}^2+\frac{1}{2}\sum_{j\neq h}^{T}(\varepsilon_{ij}-\varepsilon_{ih})^2\Big\}>0,$$

and

$$\frac{\partial f(\alpha_1,\boldsymbol{\psi}^{(t)})}{\partial\alpha_1}\bigg|_{\alpha=\frac{-1}{T-1}}=-\sum_{i=1}^{N}\omega_{il}^{(t)}\Big\{\frac{T}{2(T-1)}\sum_{j\neq h}^{T}(\varepsilon_{ij}+\varepsilon_{ih})^2\Big\}<0.$$

Therefore, $\partial f(\alpha_1,\boldsymbol{\psi}^{(t)})/\partial\alpha_1=0$ has a unique solution $\hat{\alpha_1}^{(t+1)}$ in $(-1/(T-1),1)$, $R_{CS}(\hat{\alpha}_1)$ is positive definite, and $\sum_{l=1}^{L}\hat{\pi}_l R^{(l)}(\hat{\alpha}_l)$ is also positive-definite.

Appendix C

Proof of Theorem 3. To prove Theorem 3, we need Theorem 2.2 of White (1982) and state it as the follows.

Lemma C.1. *Let $U_1,\ldots,U_N$ be i.i.d. with joint distribution function G and density g.*

(i) *Suppose that a family of distribution functions $F_\theta(u)$ has densities $f_\theta(u)$ that are measurable in u for every $\theta\in\Theta$, and continuous in θ for every $u\in\Omega$, with Θ a compact subset of space.*

(ii) *$E(\log g(U_i))$ exists and $|\log f(u,\theta)|\le M(u)$ for all θ in Θ, where M is integrable with respect to G.*

Then if $KL(g;f,\theta)$ has a unique minimum at θ_* in Θ, we have $\hat{\theta}_N\to\theta_*$ almost surely as $N\to\infty$, where $\hat{\theta}_N=\arg\max_{\theta\in\Theta}\sum_{i=1}^{N}\log f(U_i,\theta)$.

Denote the MPLE by

$$\hat{\boldsymbol{\psi}}=(\hat{\alpha}_l,\hat{\pi}_l)_{l=1}^{L}=\arg\max_{\boldsymbol{\psi}}\ell(\boldsymbol{\psi})=\arg\max_{\boldsymbol{\psi}}\Big\{\sum_{i=1}^{N}\log\sum_{l=1}^{L}\pi_l\phi_l(\boldsymbol{\varepsilon}_i,\alpha_l)\Big\}.$$

Apply Lemma 1 to $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N$ and $\sum_{l=1}^{L}\pi_l\phi_l(\boldsymbol{\varepsilon}_i,\alpha_l)$. Given Assumptions 1-3 of Theorem 3, the conditions of Lemma C.1 are all satisfied. It then follow from Lemma C.1 that $\hat{\boldsymbol{\psi}}\overset{a.s.}{\to}\boldsymbol{\psi}^*$.

To prove the second part of Theorem 3, we use Lemma 4.3 of Newey and McFadden (1994, p.2156) as stated below.

Lemma C.2. *Given an i.i.d. random sample $z_1,\ldots,z_n$, if $a(z,\theta)$ is continuous at θ_0 with probability one, and there is a neighborhood $\mathcal{N}$ of θ_0 such that $E[sup_{\theta\in\mathcal{N}}|a(z,\theta)|]<\infty$, then for any $\tilde{\theta}\overset{a.s.}{\to}\theta_0$, $(1/n)\sum_{i=1}^{n}a(z_i,\tilde{\theta})\overset{P}{\to}E[a(z,\theta_0)]$.*

For any given $\boldsymbol{\beta}$, consider the Taylor expansion

$$\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} = \frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} + \frac{\partial^2 \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}(\boldsymbol{\psi}^* - \hat{\boldsymbol{\psi}}),$$

where $\tilde{\boldsymbol{\psi}}$ is a point on the segment connecting $\hat{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}^*$. It is obvious that $\tilde{\boldsymbol{\psi}} \xrightarrow{a.s.} \boldsymbol{\psi}^*$. Then because $\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}} = 0$, we have

$$\sqrt{N}(\boldsymbol{\psi}^* - \hat{\boldsymbol{\psi}}) = \left[\frac{1}{N}\frac{\partial^2 \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}\right]^{-1} \sqrt{N}\left[\frac{1}{N}\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right].$$

Since $E\left[\partial^2 \ell(\boldsymbol{\psi})/\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau\right]$ is continuous on Ψ and $E\left[\sup_{\boldsymbol{\psi}\in\mathcal{N}} \partial^2 \ell(\boldsymbol{\psi})/\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau\right] < \infty$, it follows from Lemma C.2 that

$$\frac{1}{N}\frac{\partial^2 \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}} \xrightarrow{P} E\left[\frac{1}{N}\frac{\partial^2 \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}\partial \boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right].$$

Because $KL\left(g(\varepsilon_i), \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)\right) = E\{\log g(\boldsymbol{\varepsilon}_i)\} - E\{\log \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)\}$, we have $\boldsymbol{\psi}^* = \arg\max_{(\pi_l,\alpha_l)} E\left(\log \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)\right) = \arg\max_{\boldsymbol{\psi}} E\{(1/N)\ \ell(\boldsymbol{\psi})\}$, and hence $E\{(1/N)\partial \ell(\boldsymbol{\psi})/\partial \boldsymbol{\psi}|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\} = 0$. In addition,

$$\begin{aligned} &Var\left(\sqrt{N}\left[\frac{1}{N}\frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right]\right) \\ &\quad = E\left(\left(\frac{\partial \log \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)}{\partial \boldsymbol{\psi}}\right)\left(\frac{\partial \log \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)}{\partial \boldsymbol{\psi}}\right)^T \Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right) < \infty. \end{aligned}$$

By the Central Limit Theorem, we then have $\sqrt{N}\left[(1/N)\partial \ell(\boldsymbol{\psi})/\partial \boldsymbol{\psi}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right] = O_p(1)$. Therefore, $\sqrt{N}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) = O_p(1)$.

Appendix D

Proof of Theorem 4. Because $g(\boldsymbol{\varepsilon}_i) = \sum_{l=1}^L \pi_l^0 \phi_l(\boldsymbol{\varepsilon}_i, \alpha_l^0)$, we have

$$(\pi_l^*, \alpha_l^*)_{l=1}^L = \arg\min_{(\pi_l,\alpha_l)} KL\left(\sum_{l=1}^L \pi_l^0 \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l^0), \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)\right)$$

By Jensen's inequality, it is obvious that

$$\begin{aligned} (\pi_l^0, \alpha_l^0)_{l=1}^L &= \arg\min_{(\pi_l,\alpha_l)} \int \left[\log \frac{\sum_{l=1}^L \pi_l^0 \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l^0) d\boldsymbol{\varepsilon}_i}{\sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)}\right]\left[\sum_{l=1}^L \pi_l^0 \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l^0)\right] d\boldsymbol{\varepsilon}_i \\ &= \arg\min_{(\pi_l,\alpha_l)} KL\left(\sum_{l=1}^L \pi_l^0 \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l^0), \sum_{l=1}^L \pi_l \phi_l(\boldsymbol{\varepsilon}_\mathbf{i}, \alpha_l)\right). \end{aligned}$$

Because $KL(g(\boldsymbol{\varepsilon}_i), \sum_{l=1}^{L} \pi_l \phi_l(\boldsymbol{\varepsilon}_i, \alpha_l))$ has its unique minimum at $\boldsymbol{\psi}^* = (\pi_l^*, \alpha_l^*)_{l=1}^{L}$, Theorem 4 is established.

Appendix E

Proof of Theorem 5. From (2.5), we know that $\hat{\boldsymbol{\beta}}^M$ is the solution to the estimating equation $\sum_{i=1}^{N} U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\} = 0$.

(i) We first prove the consistency of $\hat{\boldsymbol{\beta}}^M$ ($\hat{\boldsymbol{\beta}}^M \xrightarrow{P} \boldsymbol{\beta}_0$) by induction. It is obvious that $\hat{\boldsymbol{\beta}}^0 \xrightarrow{P} \boldsymbol{\beta}_0$ for $m = 0$. Under the assumption that $\hat{\boldsymbol{\beta}}^m \xrightarrow{P} \boldsymbol{\beta}_0$, we prove $\hat{\boldsymbol{\beta}}^{m+1} \xrightarrow{P} \boldsymbol{\beta}_0$.

By the algorithm 2-4 steps, we have $\sum_{i=1}^{N} U_i\{\hat{\boldsymbol{\beta}}^{m+1}, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\} = 0$. The Taylor expansion of $\boldsymbol{\beta}$ at $\boldsymbol{\beta}_0$ gives

$$
\begin{aligned}
0 &= \frac{1}{N} \sum_{i=1}^{N} U_i\{\hat{\boldsymbol{\beta}}^{m+1}, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\} \\
&= \frac{1}{N} \sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\} + \frac{1}{N} \sum_{i=1}^{N} \frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\}}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{m+1}} (\hat{\boldsymbol{\beta}}^{m+1} - \boldsymbol{\beta}_0) \\
&= S_1 + S_2(\hat{\boldsymbol{\beta}}^{m+1} - \boldsymbol{\beta}_0),
\end{aligned}
\tag{E.1}
$$

where

$$
S_1 = \frac{1}{N} \sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\} \ \text{ and } \ S_2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\}}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^{m+1}},
$$

with $\tilde{\boldsymbol{\beta}}^{m+1}$ as a point on the segment connecting $\hat{\boldsymbol{\beta}}^{m+1}$ and $\boldsymbol{\beta}_0$.

Lemma E.1. *Under conditions of Theorem 5, we have* $S_1 = A_N + [B_N^* + o_p(1)]$ $(\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0)$, *where* $A_N = (1/N) \sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}$ *and* $B_N^* = \Big((1/N) \sum_{i=1}^{N} \partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta}\Big)_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$.

Proof of Lemma E.1. Taylor expansion of $\boldsymbol{\beta}$ at β_0 gives

$$
\begin{aligned}
S_1 =& \sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\} + \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}}{\partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} (\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0) \\
&+ R_N^m (\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0) \\
=& A_N + [B_N^* + R_N^m](\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0),
\end{aligned}
$$

where $R_N^m = (\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0)^\tau \left((1/N) \sum_{i=1}^{N} \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^m} \right)$ with $\tilde{\boldsymbol{\beta}}^m$ as a point on the segment connecting $\hat{\boldsymbol{\beta}}^m$ and $\boldsymbol{\beta}_0$. Next, we prove $R_N^m = o_p(1)$.

It is obvious that $\tilde{\boldsymbol{\beta}}^m \xrightarrow{P} \boldsymbol{\beta}_0$. Furthermore, since $[\partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau]$ is continuous at $\boldsymbol{\beta}_0$ with probability one and $E[\sup_{\boldsymbol{\beta}\in\mathcal{N}} \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau] < \infty$, it follows from Lemma 2 that $(1/N)\sum_{i=1}^N \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^m} \xrightarrow{P} E[\partial^2 U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau]$. Hence, $R_N^m = (\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0)^\tau\Big((1/N)\sum_{i=1}^N \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\tau\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^m}\Big) = o_p(1)$. Proof of Lemma E.1 is then completed.

Lemma E.2. *Under the conditions of Theorem* 5, $\sqrt{N}A_N \xrightarrow{d} N(0, \lim_{N\to\infty}(1/N)\sum_{i=1}^N G_i)$.

Proof. For a given $\boldsymbol{\beta}$, the Taylor expansion at $\boldsymbol{\psi}^*$ gives

$$\begin{aligned}
\sqrt{N}A_N &= N^{-1/2}\sum_{i=1}^N U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\} \\
&= N^{-1/2}\sum_{i=1}^N U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi}^*(\boldsymbol{\beta}))\} \\
&\quad + \left[N^{-1}\sum_{i=1}^N \frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}})\}}{\partial\hat{\boldsymbol{\psi}}}\right]\left[\sqrt{N}(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}) - \boldsymbol{\psi}^*(\boldsymbol{\beta}))\right] + R_N \\
&= C_N + D_N E_N + R_N, \qquad \text{(E.2)}
\end{aligned}$$

where

$$R_N = N^{-1/2}\sqrt{N}(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}) - \boldsymbol{\psi}^*(\boldsymbol{\beta}))^\tau\Big(\frac{1}{N}\sum_{i=1}^N \frac{\partial^2 U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}}\Big)\sqrt{N}(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}) - \boldsymbol{\psi}^*(\boldsymbol{\beta}))$$

for some $\tilde{\boldsymbol{\psi}}(\boldsymbol{\beta})$ on the segment connecting $\hat{\boldsymbol{\psi}}(\boldsymbol{\beta})$ and $\boldsymbol{\psi}^*(\boldsymbol{\beta})$.

We prove a) $E_N = O_p(1)$ and $R_N = o_p(1)$ for given $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ equals the true value $\boldsymbol{\beta}_0$, we further prove b) $D_N = o_p(1)$, and c) the asymptotic distribution of C_N is $N(0, \lim_{N\to\infty}(1/N)\sum_{i=1}^N G_i)$.

(a) From Theorem 3, we have $E_N = \sqrt{N}(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}) - \boldsymbol{\psi}^*(\boldsymbol{\beta})) = O_p(1)$ for given $\boldsymbol{\beta}$. It is obvious that $\tilde{\boldsymbol{\psi}}(\boldsymbol{\beta}) \xrightarrow{a.s.} \boldsymbol{\psi}^*(\boldsymbol{\beta})$. Furthermore, since $\left[\partial^2 U_i(\boldsymbol{\beta}, M(\boldsymbol{\psi}))/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau\right]$ is continuous at $\boldsymbol{\psi}^*$ with probability one and $E\left[\sup_{\boldsymbol{\psi}\in\mathcal{N}} \partial^2 U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau\right] < \infty$, it follows from Lemma 2 that

$$\frac{1}{N}\sum_{i=1}^N \frac{\partial^2 U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\tilde{\boldsymbol{\psi}}} \xrightarrow{P} E\left[\frac{\partial^2 U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\tau}\Big|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*}\right].$$

Therefore $R_N = N^{-1/2}O_p(1)O_p(1)O_p(1) = o_p(1)$.

(b)By the Law of Large Numbers, we have

$$\begin{aligned} D_N &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\hat{\boldsymbol{\psi}}} \\ &= \frac{1}{N}\sum_{i=1}^{N} D_i{}^{\tau} A_i^{-1/2}\frac{\partial M^{-1}(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))}{\partial\hat{\boldsymbol{\psi}}} A_i^{-1/2}(\mathbf{y_i}-\mu_{\mathbf{i}}) \\ &\rightarrow^{a.s.} E[D_i{}^{\tau} A_i^{-1/2}\frac{\partial M^{-1}(\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}} A_i^{-1/2}(\mathbf{y_i}-\mu_{\mathbf{i}})] = 0. \end{aligned}$$

(c) By the Central Limit Theorem, since $E[U_i\{\boldsymbol{\beta}_0, M(\boldsymbol{\psi}^*)\}] = E[D_i^{\tau} A_i^{-1/2}(R^*)^{-1} A_i^{-1/2}(\mathbf{y_i}-\boldsymbol{\mu}_i)] = 0$ and $Var(N^{-1}\sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\boldsymbol{\psi}^*)\}) = (1/N)\sum_{i=1}^{N} G_i$, it follows that the asymptotic distribution of C_N is $N(0, \lim_{N\to\infty}(1/N)\sum_{i=1}^{N} G_i)$.

Following (a), (b), (c), and E.2, by Slutsky's theorem the asymptotic distribution of $\sqrt{N}A_N = N^{-1/2}\sum_{i=1}^{N} U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}$ is $N(0, \lim_{N\to\infty}(1/N)\sum_{i=1}^{N} G_i)$. The proof of Lemma E.2 is complete.

Lemma E.3. *Under the conditions of Theorem* 5, $B_N \xrightarrow{P} \lim_{N\to\infty}(1/N)\sum_{i=1}^{N} H_i$.

Proof. By (b), $D_N = (1/N)\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\hat{\boldsymbol{\psi}}} = o_p(1)$. Since $|\partial\hat{\boldsymbol{\psi}}/\partial\boldsymbol{\beta}| \le H(\mathbf{Y}, \boldsymbol{\beta})$, we have

$$B_N^* = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\hat{\boldsymbol{\psi}}}\right]\left[\frac{\partial\hat{\boldsymbol{\psi}}}{\partial\boldsymbol{\beta}}\right] = D_N O_p(1) = o_p(1).$$

By the chain rule, we have

$$\frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}}{\partial\boldsymbol{\beta}} = \frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}})\}}{\partial\boldsymbol{\beta}} + \left[\frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}})\}}{\partial\hat{\boldsymbol{\psi}}}\right]\left[\frac{\partial\hat{\boldsymbol{\psi}}}{\partial\boldsymbol{\beta}}\right].$$

It is easy to see that

$$\begin{aligned} B_N = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\boldsymbol{\beta}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\ &+ \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\hat{\boldsymbol{\psi}}}\right]\left[\frac{\partial\hat{\boldsymbol{\psi}}}{\partial\boldsymbol{\beta}}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + o_p(1) \xrightarrow{P} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} H_i. \end{aligned}$$

Lemma E.4. *Under the conditions of Theorem* 5, $\hat{\boldsymbol{\beta}}^m \xrightarrow{P} \boldsymbol{\beta}_0$ *as* $N\to\infty$*, and so* $\hat{\boldsymbol{\beta}}^{m+1} \xrightarrow{P} \boldsymbol{\beta}_0$ *as* $N\to\infty$.

Proof. By (E.1), we have $(\hat{\boldsymbol{\beta}}^{m+1} - \boldsymbol{\beta}_0) = -S_2^{-1} S_1$. By Slutsky's theorem, the proof that $S_1 = o_p(1)$ follows from (E.2), $\hat{\boldsymbol{\beta}}^m \xrightarrow{P} \boldsymbol{\beta}_0$, Lemma E.2, and Lemma E.3.

Since $\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m) - \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) = \left[\partial \hat{\boldsymbol{\psi}} / \partial \boldsymbol{\beta}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}^m} (\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}_0) = O_p(1) o_p(1) = o_p(1)$ and $|\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) - \boldsymbol{\psi}^*(\boldsymbol{\beta}_0)| = o_p(1)$, we have

$$|\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m) - \boldsymbol{\psi}^*(\boldsymbol{\beta}_0)| \leq |\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m) - \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)| + |\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0) - \boldsymbol{\psi}^*(\boldsymbol{\beta}_0)| = o_p(1).$$

Therefore, there is a neighborhood $\mathcal{N}$ of $\boldsymbol{\psi}^*$ such that $\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m)$ is contained in $\mathcal{N}$ for sufficiently large N. Then, since $\left[(1/N)\sum_{i=1}^N \partial U_i\{\boldsymbol{\beta}, M(\boldsymbol{\psi})\}/\partial \boldsymbol{\beta}\right]^{-1}$ is bounded in $\mathcal{N}$ with probability one, we have $S_2^{-1} = \left[(1/N)\sum_{i=1}^N \partial U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^m))\}/\partial \boldsymbol{\beta}\right]^{-1} = O_p(1)$. Thus $\hat{\boldsymbol{\beta}}^{m+1} \xrightarrow{P} \boldsymbol{\beta}_0$ as $N \to \infty$.

(ii) We prove the asymptotic normality of $\hat{\boldsymbol{\beta}}^M$.

Lemma E.5. *Under the conditions of Theorem* 5*,*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0) = [-B_N + o_p(1)]^{-1} \sqrt{N} A_N. \tag{E.3}$$

Proof of Lemma E.5. Taylor expansion of $\boldsymbol{\beta}$ at β_0 gives

$$\begin{aligned}
0 &= N^{-1/2} \sum_{i=1}^N U_i\{\hat{\boldsymbol{\beta}}^M, M(\hat{\boldsymbol{\psi}}(\hat{\boldsymbol{\beta}}^M))\} \\
&= N^{-1/2} \sum_{i=1}^N U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\} + \left(\frac{1}{N}\sum_{i=1}^N \frac{\partial U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}}{\partial \boldsymbol{\beta}}\right) \sqrt{N}(\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0) \\
&\quad + R_N^* \sqrt{N}(\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0) \\
&= \sqrt{N} A_N + [B_N + R_N^*] \sqrt{N}(\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0),
\end{aligned}$$

where $R_N^* = (\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0)^\tau \left((1/N)\sum_{i=1}^N \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}\right)$ with $\tilde{\boldsymbol{\beta}}$ as a point on the segment connecting $\hat{\boldsymbol{\beta}}^M$ and $\boldsymbol{\beta}_0$. We prove $R_N^* = o_p(1)$.

It is obvious that $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$. Furthermore, since $[\partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau]$ is continuous at $\boldsymbol{\beta}_0$ with probability one and $E[\sup_{\boldsymbol{\beta} \in \mathcal{N}} \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau] < \infty$, it follows from Lemma 2 that $(1/N)\sum_{i=1}^N \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \xrightarrow{P} E[\partial^2 U_i\{\boldsymbol{\beta}_0, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau]$. Hence, $R_N^* = (\hat{\boldsymbol{\beta}}^M - \boldsymbol{\beta}_0)^\tau ((1/N)\sum_{i=1}^N \partial^2 U_i\{\boldsymbol{\beta}, M(\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}))\}/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\tau \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}}) = o_p(1)$. Proof of Lemma E.5 is then completed.

By Slutsky's theorem, the proof for the asymptotic normality of $\hat{\boldsymbol{\beta}}^M$ follows from (E.3), Lemma E.2, and Lemma E.3.

References

Carroll, R. J. and Ruppert, D. (1988). *Transformations and Weighting in Regression.* Chapman Hall, London.

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plann. Inference* **63**, 39-54.

Chaganty, N. R. and Joe, H.(2004). Efficiency of generalized estimating equations for binary responses. *J. Roy. Statist. Soc. Ser. B* **66**, 851-860.

Chaganty, N. R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *J. Statist. Plann. Inference* **76**, 145-161.

Crowder, M. (1985). Gaussian estimation for correlated binary data. *J. Roy. Statist. Soc. Ser. B***47**, 229-237.

Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **82**, 407-410.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measures Data.* Chapman Hall, London.

Diggle, P. J., Heagerty, P., Liang, K. L. and Zeger, S. L. (2002). *Analysis of Longitudinal Data.* 2nd edition. Oxford University Press, Oxford.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**, 141-151.

Hall, D. B. and Severini, T. A.(1998). Extended generalized estimating equations for clustered data. *J. Amer. Statist. Assoc.* **93**, 1365-1375.

Leung, D. H. Y., Wang, Y. G. and Zhu, M. (2009). Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. *Biostatistics* **10**, 436-445.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Liu, T. Q., Lin, N. and Zhang, B. X. (2008). Pseudolikelihood for Correlation Matrix Estimation in Longitudinal Data Analysis. *Technical Report, School of Mathematics and Statistics, Northeast Normal University.*

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions.* 2nd Edition. Wiley, New York.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (Edited by R. Engle and D. McFadden) **4**, 2111-2245. North-Holland, Amsterdam.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalized estimating equations using quadratic inference functions. *Biometrika* **87**, 823-836.

Sun, W., Shults, J. and Leonard, M. (2009). A note on the use of unbiased estimating equations to estimate correlation in analysis of longitudinal trials. *Biometrical J.* **51**, 5-18.

Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**, 459-465.

Ware, J. H., Dockery, D. W., Spiro, A. III, Speizer, F. E. and Ferris, B. G. Jr. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *Am. Rev. Respir. Dis* **129**, 366-374.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Wang, Y. G. and Carey, V. J. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika* **90**, 29-41.

Wang, Y. G. and Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *J. Amer. Statist. Assoc.* **99**, 845-853.

Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China.

E-mail: xull.stat@gmail.com

Department of Mathematics, Washington University in Saint Louis, Saint Louis, MO 63130, USA.

E-mail: nlin@math.wustl.edu

Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China.

E-mail: bxzhang@nenu.edu.cn

Key Laboratory for Applied Statistics of MOE and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China.

E-mail: shinz@nenu.edu.cn

(Received April 2010; accepted March 2011)