Received: 1 September 2021 | Accepted: 5 July 2022

DOI: 10.1111/biom.13718

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Grouped generalized estimating equations for longitudinal data analysis

Tsubasa Ito[1] | **Shonosuke Sugasawa**[2]

[1]Faculty of Economics and Business, Hokkaido University, Hokkaido, Japan

[2]Center for Spatial Information Science, University of Tokyo, Tokyo, Japan

Correspondence
Tsubasa Ito, Faculty of Economics and Business, Hokkaido University, Hokkaido, Japan.
Email: tito@econ.hokudai.ac.jp

Funding information
Japan Society for the Promotion of Science, Grant/Award Numbers: 18K12757, 19K23242, 22K13375

Abstract

Generalized estimating equation (GEE) is widely adopted for regression modeling for longitudinal data, taking account of potential correlations within the same subjects. Although the standard GEE assumes common regression coefficients among all the subjects, such an assumption may not be realistic when there is potential heterogeneity in regression coefficients among subjects. In this paper, we develop a flexible and interpretable approach, called grouped GEE analysis, to modeling longitudinal data with allowing heterogeneity in regression coefficients. The proposed method assumes that the subjects are divided into a finite number of groups and subjects within the same group share the same regression coefficient. We provide a simple algorithm for grouping subjects and estimating the regression coefficients simultaneously, and show the asymptotic properties of the proposed estimator. The number of groups can be determined by the cross validation with averaging method. We demonstrate the proposed method through simulation studies and an application to a real data set.

KEYWORDS
estimating equation, grouping, *k*-means algorithm, unobserved heterogeneity

1 | INTRODUCTION

Longitudinal data where response variables (repeated measurements) within the same subject are correlated widely appears in biomedical studies. For analyzing longitudinal data, it is typically difficult to correctly specify the underlying correlation structures among response variables within the same subject, and one of the standard approaches is the generalized estimating equations (GEE) developed by Liang and Zeger (1986), which uses "working" correlation structures specified by users. The advantage of the GEE approach is that the estimator is still consistent even when the working correlation is misspecified. However, the existing GEE methods assume homogeneous regression coefficients that are common to all the subjects, which could be restrictive in practical applications since there might be potential heterogeneity among subjects or clusters, as confirmed in several applications (Barban & Billari, 2012; Lin & Ng, 2012; Nagin et al., 2018). To address such heterogeneity, a crude approach is to apply a model separately to each subject, but the results are typically inaccurate and unstable as small subject-wise sample sizes often arise in real longitudinal data. Therefore, some compromised approach is required.

In this work, we extend the standard GEE analysis to take into account potential heterogeneity in longitudinal data. Specifically, we develop grouped GEE (GGEE) analysis by adopting the grouping approach that is widely adopted in literature for panel data analysis (Bonhomme & Manresa, 2015; Liu et al., 2020; Zhang et al., 2019). We assume that subjects in longitudinal data can be classified into a finite number of groups, and subjects within the same group share the same regression coefficients; that is,

the regression coefficients are homogeneous over subjects in the same groups. Since the grouping assignment of subjects is unknown, we treat it as unknown parameters and estimate them and the group-wise regression coefficient simultaneously. Given the grouping parameters, the standard GEE can be performed to obtain group-wise estimators of regression coefficients. On the other hand, given the group-wise regression coefficients, we consider estimating the grouping parameters using a kind of Mahalanobis distance between response variables and predictors taking account of potential correlations via a working correlation matrix. In other words, we employ the working correlation not only in performing GEE analysis in each group but also in estimating the grouping assignment. We will show that the GGEE method can be easily carried out by a simple iterative algorithm similar to the *k-means* algorithm that combines the existing algorithm for the standard GEE and simple optimization steps for grouping assignment. Moreover, we adopt the cross validation with the averaging method proposed in Wang (2010) to carry out a data-dependent selection of the number of groups.

We derive the statistical properties of the GGEE estimator in an asymptotic framework where both n (the number of subjects) and T (the number of repeated measurements) tend to infinity, but we here allow T to grow considerably slower than n, namely, $n/T^{\nu} \to 0$ for some $\nu > 0$. Hence, our method can be applied when T is much smaller than n as observed in many applications using longitudinal data. As theoretical difficulties of the grouped estimation in longitudinal data analysis, the true correlations within the same subject can be considerably high, so the existing theoretical argument assuming negligibly small correlations imposed typically by mixing conditions (Bonhomme & Manresa, 2015; Gu & Volgushev, 2019; Zhang et al., 2019) for the true underlying correlations are no more applicable. To overcome the limitation of the existing theoretical argument, we consider grouping assignment using a kind of Mahalanobis distance with working correlation. We will show that such a grouping strategy leads to the consistent estimation of the grouping parameters as long as the working correlation is reasonably close to the true one. Therefore, even when the underlying correlations within the same subject are not weak, we can successfully estimate the grouping parameters using a reasonable working correlation matrix. Then, we will establish consistency and asymptotic normality of the GGEE estimator of the regression coefficients and provide a consistent estimator of asymptotic variances.

In the context of longitudinal data or clustered data analysis, several methods to take account of the potential heterogeneity among subjects have been proposed. Ng and McLachlan (2014), Rubin and Wu (1997), Sugasawa et al. (2019), and Sun et al. (2007) proposed a mixture modeling based on random effects, but the estimation algorithms can be computationally very intensive since the algorithms include iteration steps that entail numerical integration. On the other hand, Rosen et al. (2000) and Tang and Qu (2016) proposed a mixture modeling based on the GEE, but the primary interest in these works is estimating the component distributions in the mixture rather than grouping subjects. Fokkema et al. (2018) and Hajjem et al. (2011, 2017) employed regression tree techniques for grouping observations, but the tree-based methods can handle grouping based on covariate information rather than regression coefficients. Moreover, Coffey et al. (2014), Vogt and Linton (2017), and Zhu and Qu (2018) proposed grouping methods for longitudinal curves, and Tang et al. (2021) developed covariate-specific grouping methods via regularization. Lastly, Zhu et al. (2021) is similar to our work, which proposed the GEE-type loss functions penalizing pairwise distance of heterogeneous fixed effects, but computational cost rapidly becomes much larger as the sample size increases compared to the k-means method. To the best of our knowledge, this paper is the first one to consider grouped estimation in the GEE analysis by the k-means algorithm with a quite small computational burden.

This paper is organized as follows. In Section 2, we illustrate the proposed GEE analysis and provide an iterative estimation algorithm. We also propose the averaging method for selecting the number of groups. In Section 3, we give the asymptotic properties of the GGEE estimator. In Section 4, we demonstrate the GGEE analysis through simulation studies and an application to a real longitudinal data set. We give some discussions in Section 6. All the technical details and the proofs of the theorems, additional numerical results, and data analyses are provided in the Supporting Information. R code implementing the proposed method is also available in the Supporting Information.

2 | GGEE ANALYSIS

2.1 | Grouped models for longitudinal data

For longitudinal data, let y_{it} be the response of interest and $\boldsymbol{x}_{it}$ be a p-dimensional vector of covariate information of subject i at time t, where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. For ease of notation, we set $T_i = T$ for all i, representing a balanced data case, but the extension to an unbalanced case is straightforward. We consider a generalized linear model for y_{it}, given by

$$f(y_{it}|\boldsymbol{x}_{it};\boldsymbol{\beta}_i,\phi) = \exp\left[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}/\phi\right], \quad (1)$$

where $a(\cdot)$ and $b(\cdot)$ are known functions, and $\theta_{it} = u(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i)$ for a known monotone function $u(\cdot)$. A commonly used link function is the canonical link function, that is, $u(x) = x$. Here $\boldsymbol{\beta}_i$ is the regression parameter of interest that can be heterogeneous among subjects, and ϕ is a known scale parameter common to all subjects. Under the model (1), the first two moments of y_{it} are given by $m(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i) = a'(\theta_{it})$ and $\sigma^2(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i) = a''(\theta_{it})\phi$, respectively. For example, under binary response, it follows that $a(x) = \log\{1 + \exp(x)\}$, leading to the logistic model given by $m(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i) = \{1 + \exp(-\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i)\}^{-1}$.

In the standard GEE analysis, the regression parameters are homogeneous, that is, $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, but we allow potential heterogeneity among the subjects. However, the number of $\boldsymbol{\beta}_i$ increases with the number of subjects, so $\boldsymbol{\beta}_i$ cannot be estimated with reasonable accuracy as long as T is not large, which is the standard situation in longitudinal data analysis. Hence, we consider a grouped structure for the subjects, that is, the n subjects are divided into G groups, and subjects within the same group share the same regression coefficients. Specifically, we introduce an unknown grouping variable $g_i \in \{1, \dots, G\}$, which determines the group that ith subject belongs to. Then, we define $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{g_i}$ under which the unknown regression parameters are $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G$. Therefore, if G is not large compared with n and T, then $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G$ can be accurately estimated. Moreover, due to the grouping nature, the estimation results of g_i give grouping of subjects in terms of regression coefficients, so the estimation result is easily interpretable for users. We also treat G as an unknown parameter, but we assume that G is known for a while. The estimation will be discussed in Section 2.3.

2.2 | Estimation algorithm

Define $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})^\top$ as a T-dimensional response vector, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{iT})^\top$ as a $T \times p$ covariate matrix. We also define $m(\boldsymbol{X}_i \boldsymbol{\beta}_g) = (m(\boldsymbol{x}_{i1}^\top \boldsymbol{\beta}_g), \dots, m(\boldsymbol{x}_{iT}^\top \boldsymbol{\beta}_g))^\top$, $\boldsymbol{A}_i(\boldsymbol{\beta}_g) = \mathrm{diag}(\sigma^2(\boldsymbol{x}_{i1}^\top \boldsymbol{\beta}_g), \dots, \sigma^2(\boldsymbol{x}_{iT}^\top \boldsymbol{\beta}_g))$, $\boldsymbol{\Delta}_i(\boldsymbol{\beta}_g) = \mathrm{diag}(u'(\boldsymbol{x}_{i1}^\top \boldsymbol{\beta}_g), \dots, u'(\boldsymbol{x}_{iT}^\top \boldsymbol{\beta}_g))$, where $\mathrm{diag}(\boldsymbol{a})$ is a diagonal matrix with a vector $\boldsymbol{a}$ as the diagonal elements, and $\boldsymbol{D}_i(\boldsymbol{\beta}_g) = \boldsymbol{A}_i(\boldsymbol{\beta}_g)\boldsymbol{\Delta}_i(\boldsymbol{\beta}_g)\boldsymbol{X}_i$. In what follows, we might abbreviate the explicit dependence on the parameters for notational simplicity when there seems to be no confusion. We here introduce "working" correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ to approximate the true underlying correlation matrix of $\boldsymbol{y}_i$, which is assumed to be common across different subjects for simplicity. This assumption can be easily extended to the heterogeneous correlation structures among different subjects. The working correlation matrix can be chosen freely, where it might include the nuisance unknown parameter $\boldsymbol{\alpha}$. Then, we define working covariance matrix $\boldsymbol{V}_i(\boldsymbol{\beta})$ as $\boldsymbol{V}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})\widehat{\boldsymbol{R}}\boldsymbol{A}_i^{1/2}(\boldsymbol{\beta})$ with $\widehat{\boldsymbol{R}} = \boldsymbol{R}(\widehat{\boldsymbol{\alpha}})$. If $\widehat{\boldsymbol{R}}$ is consistent to the true correlation matrix $\boldsymbol{R}^0$, $\boldsymbol{V}_i(\boldsymbol{\beta}^0)$ with the true parameter $\boldsymbol{\beta}^0$ is also consistent to the true covariance matrix of $\boldsymbol{y}_i$.

Given the grouping parameter $\boldsymbol{\gamma} = (g_1, \dots, g_n)$, we can estimate $\boldsymbol{\beta}_g$ by performing the standard GEE estimation (Liang & Zeger, 1986) for each group, namely, solving the following estimating equation:

$$\begin{aligned} &\boldsymbol{S}_g(\boldsymbol{\beta}_g) \equiv \sum_{i=1}^{n} 1(g_i = g)\boldsymbol{S}_i(\boldsymbol{\beta}_g) = 0, \\ &\text{s.t} \quad \boldsymbol{S}_i(\beta_g) \equiv \boldsymbol{D}_i^\top(\boldsymbol{\beta}_g)\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_g)\{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_g)\}, \end{aligned} \tag{2}$$

which is the GEE based on the subjects classified to the gth group. We can employ an existing numerical algorithm for the standard GEE to obtain the solution of (2). On the other hand, given $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_G^\top)^\top$, it is quite reasonable to classify the subjects into groups having the most suitable regression structures to explain the variation of $\boldsymbol{y}_i$. Thus, we propose estimating the unknown $\boldsymbol{\gamma}$ based on the following minimization problem:

$$\widehat{g}_i(\boldsymbol{\beta}) = \underset{g=1,\dots,G}{\mathrm{argmin}} \ \{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_g)\}^\top \widehat{\boldsymbol{R}}^{-1}\{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_g)\}. \tag{3}$$

The objective function in (3) can be seen as a kind of the Mahalanobis distance with taking the working correlation structure into account. Such an estimation strategy for the grouping variable has not been paid attention to very much, but the use of the working correlation in the grouping step is shown to be quite important to expand our theoretical argument given in Section 3. Note that the above minimization problem can be carried out separately for each subject; thus (3) can be easily solved by simply evaluating all the values of the objective function over $g \in \{1, \dots, G\}$.

Regarding the estimation of the nuisance parameter $\boldsymbol{\alpha}$ in the working correlation, we suggest using a moment-based method. Given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, one can estimate $\boldsymbol{\alpha}$ by solving the following minimization problem:

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}} \Big\| \boldsymbol{R}(\boldsymbol{\alpha}) - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i^{-1/2}\{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_{g_i})\}\{\boldsymbol{y}_i \\ - m(\boldsymbol{X}_i\boldsymbol{\beta}_{g_i})\}^\top \boldsymbol{A}_i^{-1/2} \Big\|_F, \end{aligned} \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. This method can be easily extended to the heterogeneous correlation structures among different groups. Let $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_G$ be different

correlation parameters. Then, $\boldsymbol{\alpha}_g$ can be estimated by minimizing

$$\left\| \boldsymbol{R}(\boldsymbol{\alpha}_g) - n_g^{-1} \sum_{i=1}^{n} 1(g_i = g) \boldsymbol{A}_i^{-1/2} \{\boldsymbol{y}_i - m(\boldsymbol{X}_i \boldsymbol{\beta}_{g_i})\} \{ \boldsymbol{y}_i - m(\boldsymbol{X}_i \boldsymbol{\beta}_{g_i})\}^\top \boldsymbol{A}_i^{-1/2} \right\|_F,$$

where n_g is the number of subjects classified to the gth group.

The estimating equation (2) and two optimization problems (3) and (4) define the GGEE estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the estimator can be easily computed by the following iterative algorithm 1:

ALGORITHM 1 (GGEE estimation)

Starting from some initial values $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\alpha}^{(0)}$, we repeat the following procedure until algorithm converges:

- Update $\boldsymbol{\gamma}^{(r)}$ to get $\boldsymbol{\gamma}^{(r+1)}$ by solving (3) with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(r)}$.
- Update $\boldsymbol{\beta}^{(r)}$ to get $\boldsymbol{\beta}^{(r+1)}$ by solving (2) with $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(r)}$.
- Update $\boldsymbol{\alpha}^{(r)}$ to get $\boldsymbol{\alpha}^{(r+1)}$ by solving (4) with $\boldsymbol{\beta} = \boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(r+1)}$.

Since there might be multiple solutions for the GGEE estimator, the above algorithm might be sensitive to the setting of initial values. A reasonable starting value for $\boldsymbol{\alpha}$ would induce an independent correlation matrix of $\boldsymbol{R}$, for example, $\boldsymbol{\alpha} = 0$ in the exchangeable working correlation. Regarding $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we suggest two simple methods to determine their initial values. First method is to apply the finite mixture models with G components of the form: $y_{it}|(z_{it} = k) \sim h_k(y_{it}; \boldsymbol{x}_{it}^\top \boldsymbol{\beta}_k)$ and $P(z_{it} = k) = \pi_k$, for $k = 1, \ldots, G$, where h_k is the distribution having mean $m(\boldsymbol{x}_{it}^\top \boldsymbol{\beta})$. Then, we set the initial values of $\boldsymbol{\beta}_k$ and g_i to the estimates of $\boldsymbol{\beta}_k$ and the maximizer of $\sum_{t=1}^{T} P_*(z_{it} = k)$ over $k \in \{1, \ldots, G\}$, respectively, where $P_*(z_{it} = k)$ is the conditional probability that y_{it} belongs to the kth group. The second approach is separately fitting the regression model with mean structure $m(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i)$ for each subject. Based on the estimates $\widehat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$, we apply the k-means clustering algorithm with G clusters to the n-points $\{\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_n\}$, and set the initial values of $\boldsymbol{\beta}_k$ and g_i to the center of the resulting clusters and clustering assignment, respectively. Note that the second method is only applicable when T is sufficiently larger than p to get stable estimates of $\boldsymbol{\beta}_i$.

2.3 | Selecting the number of groups

Since the number of groups is typically unknown in practice, we need to estimate it based on appropriate criteria. One possible strategy is to adopt a criterion using quasi-likelihood (Wedderburn, 1974) and to use a penalty term in view of Bayesian-type information criterion in GEE analysis (Wang & Qu, 2009). However, the theoretical asymptotic properties of such approaches are not necessarily clear even under the standard GEE settings so that the theoretical investigation would be more complicated under the grouping structure. Instead, we here adopt the cross-validation with averaging (CVA) method proposed in Wang (2010), which is shown to have the selection consistency when the clusters are properly separated into subgroups. The same strategy is adopted in Zhang et al. (2019) in the context of quantile regression for panel data.

The CVA criterion is concerned with clustering instability under given G. For $c = 1, \ldots, C$, we randomly divide n subjects into three subsets: two training data sets with sizes M and one testing set with size $n - 2M$, where the subject indices included in the three subsets are denoted by Z_1^c, Z_2^c, and Z_3^c, respectively, that is, $|Z_1^c| = |Z_2^c| = M$, $|Z_3^c| = n - 2M$, $Z_h^c \cap Z_{h'}^c = \emptyset$ for $h \neq h'$ and $\cup_{h=1,2,3} Z_h^c = \{1, \ldots, n\}$. We first apply the proposed GGEE method to the two training data sets, which gives us the estimates of regression coefficients and working correlation matrices. Then, we can compute the optimal grouping assignment in the test data as

$$\widehat{g}_i^{(h)} = \underset{g=1,\ldots,G}{\operatorname{argmin}} \{\boldsymbol{y}_i - m(\boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_g^{(h)})\}^\top \{\widehat{\boldsymbol{R}}^{(h)}\}^{-1} \{\boldsymbol{y}_i - m(\boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_g^{(h)})\}, \quad i \in Z_3^c,$$

where $\widehat{\boldsymbol{\beta}}_g^{(h)}$ and $\widehat{\boldsymbol{R}}^{(h)}$ are estimates of regression coefficients and working correlation based on the hth training data for $h = 1, 2$. Based on the grouping assignment, grouping instability can be quantified as

$$\widehat{s}^c(G) = \sum_{i,j \in Z_3^c} 1\Big\{ 1(\widehat{g}_i^{(1)} = \widehat{g}_j^{(1)}) + 1(\widehat{g}_i^{(2)} = \widehat{g}_j^{(2)}) = 1 \Big\},$$

since the summand of $\widehat{s}^c(G)$ takes the value 1 when the ith and jth subjects in the testing set are classified into the same group if we use the estimators based on one training data, but they are classified into the different groups if we use the estimators based on the other training data, which implies that the grouping results are more unstable as $\widehat{s}^c(G)$ is large. By averaging the above values over $c = 1, \ldots, C$, we have $\widehat{s}(G) = C^{-1} \sum_{c=1}^{C} \widehat{s}^c(G)$, and we select G as the minimizer of the criterion among some candidates of G. Finally, regarding the choice of M, we set $M = \lfloor n/3 \rfloor$, so the three subsets have almost the same number of subjects.

3 | ASYMPTOTIC PROPERTIES

We here provide the asymptotic properties of the GGEE estimators, that is, the grouping parameter $\boldsymbol{\gamma}$ can be consistently estimated, and $\widehat{\boldsymbol{\beta}}_g$ admits both consistency and asymptotic normality. Our asymptotic framework is that both n and T tend to infinity, but we allow T to grow considerably slower than n, as discussed later.

We first prepare some notations before assumptions. Let $\boldsymbol{M}_g(\boldsymbol{\beta}_g) = \mathrm{Cov}(\boldsymbol{S}_g(\boldsymbol{\beta}_g)) = \sum_{i=1}^n 1\{g_i = g\}\boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i(\boldsymbol{\beta}_g)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i$ and $\boldsymbol{H}_g(\boldsymbol{\beta}_g) = -\mathrm{E}[\partial \boldsymbol{S}_g(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g] = \sum_{i=1}^n 1\{g_i = g\}\boldsymbol{D}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{D}_i$. We here denote the working correlation matrix as $\boldsymbol{R}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ to emphasize its dependence on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, and let $\widehat{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{R}(\widehat{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. We also let $\overline{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{R}(\overline{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ be a constant positive definite matrix, where $\overline{\alpha}$ is a nonrandom constant to which $\widehat{\alpha}$ converges. We do not require $\overline{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to be the true correlation matrix $\boldsymbol{R}^0$. Next, we denote $\overline{\boldsymbol{V}}_i(\boldsymbol{\beta}_g)$ by replacing $\widehat{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with $\overline{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in $\boldsymbol{V}_i(\boldsymbol{\beta}_g)$. $\overline{\boldsymbol{S}}_i(\boldsymbol{\beta}_g)$, $\overline{\boldsymbol{S}}_g(\boldsymbol{\beta}_g)$, $\overline{\boldsymbol{M}}_g(\boldsymbol{\beta}_g)$, and $\overline{\boldsymbol{H}}_g(\boldsymbol{\beta}_g)$ are defined similarly. To facilitate the Taylor expansion of the estimating function of GEE, we denote the negative gradient function of $\boldsymbol{S}_i(\boldsymbol{\beta}_g)$ as $\mathcal{D}_i(\boldsymbol{\beta}_g) = -\partial \boldsymbol{S}_i(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g^\top$. $\overline{\mathcal{D}}_i(\boldsymbol{\beta}_g)$ is defined as $\overline{\boldsymbol{V}}_i(\boldsymbol{\beta}_g)$. For $g = 1, \ldots, G$, let $\boldsymbol{\beta}_g^0$ be a true value of $\boldsymbol{\beta}_g$ and g_i^0 be a group variable, which ith cluster actually belongs to. Then, we also define the oracle score function for $\boldsymbol{\beta}_g$ under the true grouping assignment as $\boldsymbol{S}_g^*(\boldsymbol{\beta}_g) = \sum_{i=1}^n 1\{g_i^0 = g\}\boldsymbol{S}_i(\boldsymbol{\beta}_g)$. $\overline{\boldsymbol{S}}_g^*(\boldsymbol{\beta}_g)$, $\overline{\boldsymbol{M}}_g^*(\boldsymbol{\beta}_g)$, and $\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g)$ are similarly defined. As discussed in Xie and Yang (2003), to prove the existence and weak consistency of the clustered GEE estimators, we need assumptions given later in Assumption (A3), that is, for all $g = 1, \ldots, G$, $\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)$s or $\lambda_{\min}(\overline{\boldsymbol{H}}^*) \equiv \min_{1\le g\le G} \lambda_{\min}(\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0))$ are divergent at a rate faster than $\tau \equiv \sup_{\boldsymbol{\beta},\boldsymbol{\gamma}} \lambda_{\max}(\{\overline{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})\}^{-1}\boldsymbol{R}^0)$. To make further assumptions, we need to introduce some notations similar to those in Wang (2011) and Xie and Yang (2003). We denote a local neighborhood of $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0\top}, \ldots, \boldsymbol{\beta}_G^{0\top})^\top$ as $\mathcal{B}_{nT} = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top : \max_{g=1,\ldots,G} ||\{\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)\}^{1/2}(\boldsymbol{\beta}_g - \boldsymbol{\beta}_g^0)|| \le C\tau^{1/2}\}$. Lastly, we denote $\varepsilon_{it} = A_{it}^{-1/2}(\boldsymbol{\beta}_{g_i^0}^0)\{y_{it} - m(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_{g_i^0}^0)\}$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\top$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

We here give some regularity assumptions, and the other technical assumptions are given in Supporting Information Section S.1.

Assumption 1.

(A1) (i) For all $g = 1, \ldots, G$, the unknown parameter $\boldsymbol{\beta}_g$ belongs to a compact subset $\mathcal{B} \in \mathbb{R}^p$, the true parameter value $\boldsymbol{\beta}_g^0$ lies in the interior of $\mathcal{B}$, (ii) the covariates $\{\boldsymbol{x}_{it}, i = 1, \ldots, n, t = 1, \ldots, T\}$ are in a compact set $\mathcal{X}$.

(A2) (i) For all $g = 1, \ldots, G$, $\lim_{n\to\infty}(1/n)\sum_{i=1}^n 1\{g_i^0 = g\} = \pi_g > 0$, and (ii) for all $g, g' = 1, \ldots, G$ such that $g \ne g'$ and $c > 0$, $\min_{1\le g,g'\le G} ||\boldsymbol{\beta}_g^0 - \boldsymbol{\beta}_{g'}^0|| > c$.

(A3) $\tau\lambda_{\min}^{-1}(\overline{\boldsymbol{H}}^*) \to 0$.

(A4) For all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, $E[\varepsilon_{it}^{2+2/\zeta}] \le M$ for some $0 < \zeta \le 1$.

(A5) The eigenvalues of the true correlation matrix $\boldsymbol{R}^0$ are bounded away from 0, and the eigenvalues of $\overline{\boldsymbol{R}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are bounded away from 0 uniformly for any $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. All off-diagonal elements of $\boldsymbol{R}^0$ are uniformly bounded away from 1.

Assumption (A1) seems to be slightly strict. However, the compactness of the parameter space and the set of all possible covariates is required because in the proof of the consistency of our GGEE estimators, we need to bound $a''(\theta_{it})$ and $u'_{it}(\eta_{it})$ uniformly on the whole parameter space for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Assumption (A2) is typically imposed in the literature on the grouping approach in panel data models (Bonhomme & Manresa, 2015), which ensures that the G subgroups are well separated so that the parameters $\boldsymbol{\beta}_g$s and $\boldsymbol{\gamma}$ can be identifiable. Assumption (A3) is the same as the condition (L*) in Xie and Yang (2003). Assumption (A4) is slightly stronger than the condition in Lemma 2 of Xie and Yang (2003) since we require the fourth moment of error terms to be finite. Assumption (A5) is the same assumption imposed well in the literature on GEE with large cluster sizes. Assumption (A5) is a much weaker assumption than the one typically adopted in the existing literature on the grouped estimation (Bonhomme & Manresa, 2015; Gu & Volgushev, 2019; Zhang et al., 2019) in which $\{\varepsilon_{it}\}_{t=1,\ldots,T}$ is assumed to satisfy some strong mixing conditions with a faster-than-polynomial decay rate. Such assumptions are quite unrealistic in longitudinal data analysis, so we do not impose any restriction on the correlation strength of $\{\varepsilon_{it}\}_{t=1,\ldots,T}$, which is essentially related to the use of a kind of Mahalanobis distance for grouping assignment given in (3). Moreover, since we assume that true correlations are uniformly bounded away from 1, we can estimate each $\boldsymbol{\beta}_i$ consistently by solving $\boldsymbol{S}_i(\boldsymbol{\beta}_i) = 0$ from the assumptions in Supporting Information Section S.1, as argued in Xie and Yang (2003).

We now give our main theorems. We first establish the existence and weak consistency of the GGEE estimators and the classification consistency of the grouping variables.

Theorem 1. *Suppose the Assumptions (A1)–(A5) and the assumptions in Supporting Information Section S.1 hold.*

For all $g = 1, \dots, G$, $\boldsymbol{S}_g(\boldsymbol{\beta}_g) = 0$ *has a root* $\widehat{\boldsymbol{\beta}}_g$ *such that* $\widehat{\boldsymbol{\beta}}_g \to \boldsymbol{\beta}_g^0$ *in probability. Moreover, as* n *and* T *tend to infinity such that* $n/T^\nu \to 0$ *for some* $\nu > 0$*, it holds that* $P(\max_{1\le i\le n} |\widehat{g}_i(\widehat{\boldsymbol{\beta}}) - g_i^0| > 0) = o(1) + O(nT^{-\delta})$ *for all* $\delta > 0$ *for* $\widehat{g}_i(\widehat{\boldsymbol{\beta}})$*s are obtained by (3).*

Since the second part of Theorem 1 holds for all $\delta > 0$, the probability of miss-clustering vanishes if we take δ larger than ν in Assumption (A9) (iv) in Supporting Information Section S.1.

We next establish the asymptotic normality of $\widehat{\boldsymbol{\beta}}_g$ for $g = 1, \dots, G$. The following notations are similar to Xie and Yang (2003): $c^* = \max_{1\le g\le G} \lambda_{\max}(\overline{\boldsymbol{M}}_g^*(\boldsymbol{\beta}_g^0)^{-1}\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0))$ and

$$\gamma^* = \max_{1\le g\le G} \max_{i:g_i^0=g} \lambda_{\max} (\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)^{-1/2}\boldsymbol{D}_i^\top(\boldsymbol{\beta}_g^0)\overline{\boldsymbol{V}}_i^{-1}(\boldsymbol{\beta}_g^0)\boldsymbol{D}_i(\boldsymbol{\beta}_g^0)\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)^{-1/2}).$$

The following result is a direct consequence of Theorem 4 in Xie and Yang (2003) combined with Lemma S.9 in Supporting Information Section S.1.

Theorem 2. *Suppose the Assumptions (A1)–(A5) and the assumptions in Supporting Information Section S.1 hold. Moreover, suppose that, for all* $g = 1, \dots, G$*, there exists a constant* ζ *such that* $(c^*T)^{1+\zeta}\gamma^* \to 0$ *as* $n \to \infty$*. Moreover, suppose the marginal distribution of each observation has a density of the form from (1). Then, as* n *and* T *tend to infinity such that* $n/T^\nu \to 0$ *for some* $\nu > 0$*, we have* $\overline{\boldsymbol{M}}_g^*(\boldsymbol{\beta}_g^0)^{-1/2}\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)(\widehat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_g^0) \to N(0, \boldsymbol{I}_p)$ *in distribution.*

From Theorem 2, it can be easily shown that for all n, $\{\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)\}^{-1}\overline{\boldsymbol{M}}_g^*(\boldsymbol{\beta}_g^0)\{\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)\}^{-1}$ is minimized in the matrix sense when $\overline{\boldsymbol{V}}_i = \boldsymbol{\Sigma}_i$ for all i. This implies that the group GEE estimator becomes most efficient when we can specify the working correlation matrix correctly, and the corresponding asymptotic variance of $\widehat{\boldsymbol{\beta}}_g$ is given by $\lim_{n\to\infty}[\sum_{i=1}^n 1\{g_i^0 = g\}\boldsymbol{D}_i^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{D}_i]^{-1}$.

Moreover, $\{\boldsymbol{H}_g(\widehat{\boldsymbol{\beta}}_g)\}^{-1}\boldsymbol{M}_g(\widehat{\boldsymbol{\beta}}_g)\{\boldsymbol{H}_g(\widehat{\boldsymbol{\beta}}_g)\}^{-1}$ can be used as the estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_g$. Since this estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_g$ involves $\boldsymbol{M}_g(\widehat{\boldsymbol{\beta}}_g)$ depending on the unknown covariance matrix $\widehat{\boldsymbol{\Sigma}}_i = \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\beta}}_{\widehat{g}_i})$ for $\widehat{g}_i = g$, following Liang and Zeger (1986), we suggest obtaining $\boldsymbol{M}_g(\widehat{\boldsymbol{\beta}}_g)$ by

$$\sum_{i=1}^n 1(\widehat{g}_i = g)\boldsymbol{D}_i^\top(\boldsymbol{\beta}_g)\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_g)\{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_g)\} \{\boldsymbol{y}_i - m(\boldsymbol{X}_i\boldsymbol{\beta}_g)\}^\top\boldsymbol{V}_i^{-1}(\boldsymbol{\beta}_g)\boldsymbol{D}_i(\boldsymbol{\beta}_g)\Big|_{\boldsymbol{\beta}_g=\widehat{\boldsymbol{\beta}}_g},$$

which is consistent to $\overline{\boldsymbol{M}}_g^*(\boldsymbol{\beta}_g^0)$ as $n \to \infty$ from Lemma 1 in Supporting Information Section S.2. Similarly, we can show that $\boldsymbol{H}_g(\widehat{\boldsymbol{\beta}}_g)$ is consistent to $\overline{\boldsymbol{H}}_g^*(\boldsymbol{\beta}_g^0)$, which implies that $\{\boldsymbol{H}_g(\widehat{\boldsymbol{\beta}}_g)\}^{-1}\boldsymbol{M}_g(\widehat{\boldsymbol{\beta}}_g)\{\boldsymbol{H}_g(\widehat{\boldsymbol{\beta}}_g)\}^{-1}$ converges to the asymptotic variance of $\widehat{\boldsymbol{\beta}}_g$. Although the variability in the estimation of grouping parameters can be ignored according to Theorems 1 and 2, it can be considerable under finite sample sizes. As an alternative method, we also suggest using clustered bootstrap (eg, Field & Welsh, 2007). This approach generates the bootstrap sample $y_1^*, \dots, y_n^*$ from the distribution placing probability $1/n$ on each of $y_i = (y_{i1}, \dots, y_{iT})$. Letting $\widehat{\boldsymbol{\beta}}_g^*$ be the estimator obtained from the bootstrap sample $y_1^*, \dots, y_n^*$, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_g$ can be approximated by the sample variance of replications of $\widehat{\boldsymbol{\beta}}_g^*$.

4 | SIMULATION STUDIES

We investigate the finite sample performance of the proposed GGEE method through simulation studies. First, we consider the estimation and classification accuracy of the GGEE estimator. To this end, we generated a two-dimensional covariate vector (x_{1it}, x_{2it}) from a two-dimensional normal distribution with mean 0, marginal variance 1, and correlation 0.4, for $i = 1, \dots, n$ and $t = 1, \dots, T$. We considered the logistic model for the marginal expectation of Y_{it}, namely, $Y_{it} \sim \text{Ber}(\pi_{it})$ and $\text{logit}(\pi_{it}) = \boldsymbol{X}_{it}^\top\boldsymbol{\beta}_{g_i}$, where $\boldsymbol{X}_{it} = (1, x_{1it}, x_{2it})^\top$, $g_i \in \{1, \dots, G\}$, and $\boldsymbol{\beta}_g = (\beta_{g0}, \beta_{g1}, \beta_{g2})^\top$ is a vector of unknown regression coefficients. Here, we set $G = 3$ and $\boldsymbol{\beta}_1 = (0, -2, 0)^\top$, $\boldsymbol{\beta}_2 = (1, 1, 2)^\top$, and $\boldsymbol{\beta}_3 = (-1, 1, -2)^\top$. For the true grouping assignment, we set $g_i = 1$ for $i = 1, \dots, n/3$, $g_2 = 2$ for $i = n/3 + 1, \dots, 2n/3$ and $g_i = 3$ for $i = 2n/3 + 1, \dots, n$. Based on the probability π_{it}, we generated $(Y_{i1}, \dots, Y_{iT})$ from a correlated binary vector using R package "bindata" with two scenarios of correlation matrix, exchangeable correlation matrix with 0.5 correlation parameter, and AR(1) correlation matrix with 0.7 correlation parameter. We then applied the proposed GGEE method with $G = 3$ and four options of correlation matrices, independent (ID), exchangeable correlation (EX), AR(1) correlation (AR), and unstructured correlation (US) matrices, and unknown parameters in these correlation matrices were also estimated. For comparison, we also applied the naive grouping (NG) method that first separately fits the logistic regression to each subject to estimate subject-specific regression coefficients, then group them via k-means clustering, and re-estimate group-wise regression coefficients.

TABLE 1 Average values of squared error loss of the regression coefficients in three groups based on the proposed grouped GEE method with independent (ID), exchangeable correlation (EX), first-order autoregressive (AR), and unstructured (US) working correlation matrices. The results of the naive grouping (NG) method using the subject-wise estimation of regression coefficients are also given for comparison. The reported values are averaged over 5000 Monte Carlo replications and are multiplied by 100

		True correlation: EX					True correlation: AR				
(n,T)	Group	ID	EX	AR	US	NG	ID	EX	AR	US	NG
(180,10)	1	8.8	7.8	9.0	8.7	12.9	7.9	7.4	7.2	7.3	10.2
	2	9.3	8.3	9.1	8.6	12.7	8.2	7.6	7.5	7.4	10.1
	3	9.3	7.8	9.3	8.7	12.7	8.0	7.4	7.6	7.7	10.3
(180,20)	1	4.4	3.7	4.4	5.0	5.0	3.1	2.9	2.8	3.2	3.1
	2	4.3	3.8	4.3	5.1	4.9	3.1	2.9	2.8	3.2	3.2
	3	4.3	3.8	4.4	5.3	5.0	3.1	3.0	2.8	3.3	3.0
(270,10)	1	6.5	5.2	6.2	5.6	10.3	5.8	5.0	5.0	4.7	7.4
	2	6.4	5.4	6.1	5.6	10.1	5.7	4.9	4.8	5.0	7.5
	3	6.8	5.2	6.3	5.5	10.4	5.7	4.8	4.8	4.8	7.5
(270,20)	1	2.9	2.5	2.8	3.0	3.4	2.1	1.9	1.8	2.0	2.1
	2	2.9	2.5	2.9	2.9	3.4	2.0	2.0	1.9	1.9	2.0
	3	2.8	2.5	2.9	3.0	3.4	2.1	2.0	1.9	2.0	2.1

TABLE 2 Average values of classification error (%) of the grouping parameters in the grouped GEE analysis with independent (ID), exchangeable correlation (EX), first-order autoregressive (AR), and unstructured (US) working correlation matrices, averaged over 5000 Monte Carlo replications

	True correlation: EX				True correlation: AR			
(n,T)	ID	EX	AR	US	ID	EX	AR	US
(180, 10)	9.6	4.4	6.6	5.3	6.5	4.8	4.0	4.8
(180, 20)	4.3	1.5	2.3	1.8	1.9	1.6	1.2	1.5
(270, 10)	8.5	4.3	6.0	4.9	6.1	4.6	4.0	4.4
(270, 20)	3.7	1.5	2.1	1.4	1.8	1.4	1.3	1.4

We evaluated the performance of the estimation of $\boldsymbol{\beta}_g$ by using the squared error loss defined as $\mathrm{SEL}_g = \sum_{k=0}^{2}(\widehat{\beta}_{gk} - \beta_{gk})^2$, and assessed the classification accuracy via the classification error given by $\mathrm{CE} = n^{-1}\sum_{i=1}^{n} 1(\widehat{g}_i \neq g_i)$. In Tables 1 and 2, we reported the average values of SEL and CE using 5000 Monte Carlo replications, respectively, under four combinations of (n,T).

From Table 1, we can see that the correct specification of working correlation matrices induces the most efficient estimation of the regression coefficient. In contrast, using the other working correlations that are not necessarily equal to the true correlation structures can still provide a more efficient estimation than the independent working structure. We also note that the US working correlation includes both EX and AR, although the number of unknown parameters is much larger than these structures. Hence, the estimation performance under the moderate sample size such as $(n,T) = (180, 10)$ is not very satisfactory, but the performance improves as the sample size increases. Regarding NG, the performance is comparable when T is not small (eg, $T = 20$), while the performance gets worse as T decreases. This would be because the subject-wise fitting does not perform well when T is not large, leading to poor grouping results. From Table 2, it is observed that introducing working correlation structures in the classification step (3) achieves a more accurate classification than the common classification strategy using the standard sum of squared residuals as adopted in existing literature when observations within the same subject are correlated. Moreover, the results reveal that the correct specification of the working correlation leads to the most accurate classification. In the Supporting Information, we provide simulation results for 95% confidence intervals of β_1, β_2, and β_3.

We next investigate the performance of the CVA selection strategy given in Section 2.3 by adopting the same data-generating process with an exchangeable correlation structure. For the simulated data set, we selected the number of components G using the CVA criteria from the candidate $G \in \{2, 3, \ldots, 7\}$, noting that the true number of components is 3. We employed four working correlations, ID, EX, AR, and US, to carry out the GGEE analysis for each G. Based on Monte Carlo replications, we obtained selection probabilities of each G, which are reported in Table 3.

It is observed that the use of independent working correlations under significant correlations within the same individual does not necessarily provide satisfactory selection performance when the number of samples is limited. We can also see that the selection probabilities of the true

TABLE 3 Selection probabilities (%) of the number of groups (G) obtained from the CVA criteria in Section 2.3 with independent (ID), exchangeable (EX), first-order autoregressive (AR), and unstructured (US) working correlation matrices, based on 200 Monte Carlo replications

(n,T)	Working correlation	G 2	3	4	5	6	7
(180,10)	ID	0.5	61.0	8.0	10.0	3.5	17.0
	EX	3.0	95.0	0.5	1.0	0.5	0.0
	AR	3.0	78.0	5.0	7.5	2.5	4.0
	US	0.0	89.5	2.0	2.5	1.0	5.0
(180,20)	ID	0.0	94.0	3.0	1.0	0.5	1.5
	EX	0.0	100.0	0.0	0.0	0.0	0.0
	AR	0.0	100.0	0.0	0.0	0.0	0.0
	US	0.0	93.5	5.5	0.0	0.0	1.0
(270,10)	ID	3.0	77.5	3.5	9.5	0.5	6.0
	EX	2.0	98.0	0.0	0.0	0.0	0.0
	AR	2.0	97.0	0.0	0.5	0.0	0.5
	US	0.5	98.5	0.5	0.0	0.5	0.0
(270,20)	ID	0.0	96.5	1.0	1.5	0.0	1.0
	EX	0.5	99.5	0.0	0.0	0.0	0.0
	AR	0.0	100.0	0.0	0.0	0.0	0.0
	US	0.0	100.0	0.0	0.0	0.0	0.0

number of components based on EX and US working correlations tend to be larger than those of using the AR working correlation structure since the true correlation is EX. Moreover, when the sample sizes are large, such as $(n,T)=(270,20)$, the adopted CVA strategy can select the true number of components with a probability of almost 1, which would be compatible with the selection consistency of the strategy.

Finally, we compare the proposed GGEE method with some existing methods under situations where the subjects do not necessarily admit perfect grouping. To this end, we considered the following underlying scenarios for the subject-specific regression coefficients:

$$\text{(S1)}\ \boldsymbol{\beta}_i \sim \pi_1\boldsymbol{\delta}(0,-2,0)+\pi_2\boldsymbol{\delta}(1,1,2)+\pi_3\boldsymbol{\delta}(-1,1,-2),$$
$$\pi_1=\pi_2=\pi_3=\frac{1}{3},$$
$$\text{(S2)}\ \boldsymbol{\beta}_i=(0,-2,0)1(g_i=1)+(1,1,2)1(g_i=2)$$
$$+(-1,1,-2)1(g_i=3)+U([-0.5,0.5]^3),$$
$$\text{(S3)}\ \boldsymbol{\beta}_{i0}\sim U([-0.2,0.2]),$$
$$\beta_{i1}\sim U([-2,2]),\quad \beta_{i2}\sim U([0,2]),$$

where $\boldsymbol{\delta}(a_1,a_2,a_3)$ denotes a Dirac distribution on (a_1,a_2,a_3), $U(A)$ denotes the uniform distribution on the region A, and g_i is the grouping variable defined as $g_i=1$ for $i=1,\dots,n/3$, $g_2=2$ for $i=n/3+1,\dots,2n/3$ and $g_i=3$ for $i=2n/3+1,\dots,n$. Note that scenario (S1) is quite similar to the one used in the previous simulation study. On the other hand, in scenarios (S2) and (S3), the subjects do not admit complete classification since the regression coefficients are different among subjects. We also note that in scenario (S2), the subjects may admit approximate classification based on g_i, but there seems to be no trivial classification in scenario (S3) as the regression coefficients are completely random. The binary response variable Y_{it} is generated in the same way as the previous study with the exchangeable correlation structure with 0.5 correlation parameter. We generated a new vector of covariates $\boldsymbol{X}_{i,T+1}$ from the same data-generating process, and the target to be estimated is the success probability of future observations, $\mu_i\equiv \text{logit}^{-1}(\boldsymbol{X}_{i,T+1}^\top\boldsymbol{\beta}_i)$. For the simulated data set, we applied the proposed GGEE method with the estimated number of groups to estimate $\boldsymbol{\beta}_i$ by $\widehat{\boldsymbol{\beta}}_{\widehat{g}_i}$. For comparison, we applied random coefficient models (RC), growth mixture models (eg, Ram & Grimm, 2009), denoted by GMM, and pairwise penalization approaches (Zhu et al., 2021), denoted by PWL, to estimate the subject-specific coefficient $\boldsymbol{\beta}_i$, where the details of each method are provided in the Supporting Information. Then, μ_i is estimated by $\boldsymbol{X}_{i,T+1}^\top\widehat{\boldsymbol{\beta}}_i$. Furthermore, we also applied the generalized linear mixed model tree (Fokkema et al., 2018; Hajjem et al., 2017), denoted by GLMMT, to directly estimate μ_i, for which we used the R package "glmertree" (Fokkema et al., 2018).

TABLE 4 Squared root of mean squared error (RMSE) of estimators of the success probability of future observations, averaged over 1000 Monte Carlo replications, for the proposed method with two working correlation matrices (GGEE-EX and GGEE-US), and four competing methods

Method	(S1) $T=10$	(S1) $T=20$	(S2) $T=10$	(S2) $T=20$	(S3) $T=10$	(S3) $T=20$
CGEE-EX	12.6	6.3	15.0	9.5	29.0	23.8
CGEE-US	13.8	8.9	16.1	11.4	29.1	24.2
RC	22.3	21.3	22.5	21.6	24.2	24.3
LCM	13.7	10.2	15.8	12.7	24.5	24.1
MT	32.7	33.3	34.0	35.4	20.5	21.3
PWL	19.9	15.9	20.8	16.8	21.3	21.5

The performance of estimating μ_i is measured by the square root of mean squared errors (RMSE), defined as $\{n^{-1}\sum_{i=1}^n(\widehat{\mu}_i-\mu_i)^2\}^{1/2}$. The averaged values of RMSE based on 1000 Monte Carlo replications are presented in Table 4.

In scenario (S1), since the subject-specific regression coefficients can be perfectly grouped, the proposed methods provide better estimation accuracy than the other

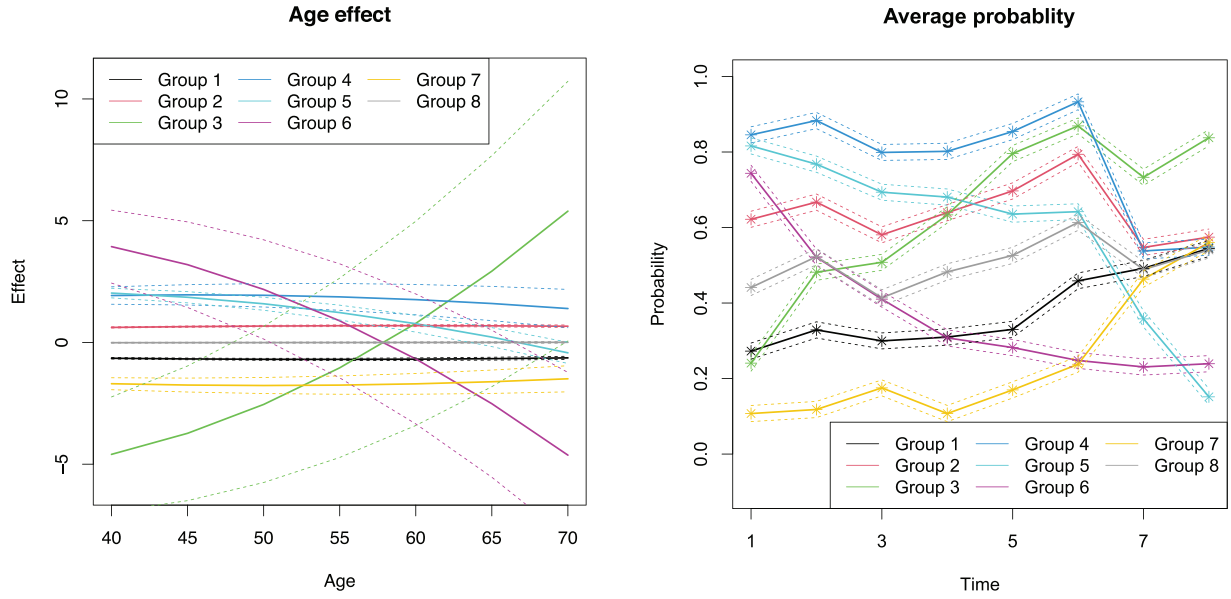

FIGURE 1 Estimates (solid line) and 95% point-wise confidence intervals (dotted line) of group-wise quadratic effects of age (left) and group-wise probability being "unhealthy" (right) in detected eight heterogeneous groups. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

methods except for LCM. Moreover, in scenario (S2), the subjects do not hold exact grouping structures but can be approximately grouped, and the proposed method still works better than the other methods except for LCM. On the other hand, the regression coefficients are completely random in scenario (S3), and the results show that MT and PWL are appealing. It should be noted that the difference between the GGEE and RC methods are relatively comparable, which would indicate that the proposed GGEE method can reasonably approximate the subject-specific random coefficients by grouping subjects having similar regression coefficients. Finally, comparing the two working correlations, the EX correlation provides better performance than the US correlation since the EX is the true underlying correlation structure within the same subject. In contrast, the US correlation is quite comparable with EX.

5 | APPLICATION TO THE HEALTH AND RETIREMENT STUDY (HRS)

We apply the proposed method to the HRS data, which come from the study conducted by the University of Michigan. This longitudinal panel study surveys adults over the age of 50 in the United States through detailed interviews once every 2 years for each participant and provides information on their health and economic circumstances. For more details, see Juster and Suzman (1995). The main goal of the study is to investigate the change in participants' health conditions in the HRS study over time and the relevant factors associated with their condition. We used the data set from the HRS study, which can be obtained from an R Package "LMest." The sample includes $n = 7074$ individuals followed at $T = 8$ approximately equally spaced occasions without missing responses or dropouts. The response variable is the self-reported health status (named SHLT), in which five categories of statuses: "poor," "fair," "good," "very good," and "excellent," are recorded as an ordinal response variable from 1 to 5, noting that a smaller value corresponds to a high level of health condition. We then dichotomized the response by setting values of 1 or 2 to "healthy" (1) and the other values to "unhealthy" (0). As auxiliary information, we adopted indicator variables of gender (1:male, 0:female), indicators of black and others, respectively, indicators of two education levels, "some college" (SC) and "college and above" (CAA), and age, which is measured in years for each time occasion. We also included a quadratic term age and seven time effects for $t = 2, \dots, 8$. Among the individuals, it would be reasonable to assume that different types of individuals exist, that is, some individuals are always healthy, whereas some individuals are not, or their health condition changes during the term. Therefore, instead of focusing on population-averaged regression coefficients, we here focus on such potential heterogeneity in the population to apply the proposed GGEE approach.

Let y_{it} be the binary response variable, and x_{it} be the vector of five covariates and an intercept, for $i = 1, \dots, n(= 7074)$ and $t = 1, \dots, T(= 8)$. We consider the mean structure $E[y_{it}|\boldsymbol{x}_{it}] = m(\boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_{g_i})$ with $m(x) = \exp(x)/\{1 + \exp(x)\}$ and $g_i \in \{1, \dots, G\}$. In this analysis, we

TABLE 5 Point estimates (PE) and standard errors (SE) of group-specific regression coefficients, where the values of PE and SE are multiplied by 1000

		Group							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Group size		1478	1650	191	686	310	117	559	2083
Intercept	PE	−0.87	0.79	−7.64	2.87	4.33	7.02	−2.29	−0.02
	SE	0.01	0.01	0.83	0.13	0.07	0.54	0.09	0.01
Gender	PE	4.25	−1.50	1138.89	−11.86	195.25	−1455.80	34.66	2.33
	SE	2.07	1.66	75.91	27.53	10.40	75.20	29.57	1.54
Black	PE	0.67	−0.28	−200.84	−1.34	37.33	−481.21	2.17	0.23
	SE	0.27	0.26	17.67	3.08	1.93	30.07	1.87	0.15
Other	PE	−0.16	0.06	42.70	−0.10	0.54	−41.30	−1.64	−0.03
	SE	0.04	0.01	4.10	0.30	0.02	2.25	1.10	0.02
SC	PE	0.63	−0.81	−31.44	0.36	95.86	339.82	5.53	0.32
	SE	0.36	0.69	50.64	0.82	3.76	45.74	5.07	0.22
CAA	PE	−1.78	1.40	84.76	6.20	−145.79	406.11	−11.38	−0.70
	SE	0.65	0.83	25.22	9.68	7.17	47.82	7.99	0.45
Age	PE	−25.82	23.28	−370.78	85.88	126.41	317.68	−70.48	−0.75
	SE	0.31	0.24	24.73	3.89	2.12	15.64	2.68	0.19
Age^2	PE	0.24	−0.20	6.40	−0.94	−1.89	−5.48	0.70	0.01
	SE	0.01	0.00	0.43	0.06	0.03	0.27	0.04	0.00

use unstructured working correlation. We first selected the number of groups, G, from candidates $\{2, 3, \ldots, 10\}$, using the CVA value. The CVA value for each G is shown in Figure 1 in Supporting Information Section S.5.3, and the CVA value is minimized at $G = 8$. Thus, we carried out the GGEE analysis with $G = 8$ in what follows. The estimated regression coefficients and their standard errors in eight groups are shown in Table 5.

It is observed that estimated regression coefficients in the eight groups are very different from each other. To visualize the difference, we computed the estimated quadratic function of the age effect in Figure 1, which indicates that some groups have representative shapes of the age effect.

For example, the probability of "health" of individuals classified in group 3 increases according to their age, while the opposite tendency is confirmed in group 6. Although clear differences among four groups (groups 1, 2, 7, and 8) are not observed from Figure 1, the regression coefficients of other covariates reported in Table 5 are quite different. Moreover, in each group, we computed average values of y_{it} for $t = 1, \ldots, T$, where the results are presented in the right panel in Figure 1. From the result, we can more directly understand the characteristics of the eight groups. For example, individuals in groups 3 and 7 have a low probability of being "healthy" at the earlier period, and the probability increases with the period. On the other hand, the probability in groups 5 and 6 decreases according to the period, but there is a difference in the shape of the decrease. Therefore, we can conclude that the classical GEE analysis assuming homogeneity in the regression coefficients is not an appropriate strategy for the data set. In contrast, the proposed GGEE analysis can successfully capture the potential heterogeneity among individuals.

6 | CONCLUDING REMARKS

This paper developed a new statistical approach to analyzing longitudinal data. The proposed method called GGEE analysis carries out grouping subjects and estimating the regression coefficients simultaneously to take account of potential heterogeneity. We employed working correlations in estimation and grouping steps and provided a simple iterative algorithm to obtain GGEE estimator. We also developed asymptotic properties of the proposed method. The simulation studies and an application to the HRS suggest the usefulness of the proposed approach.

The proposed method has some useful extensions. First, we can introduce a penalty term in the grouping step as considered in Sugasawa (2021), which can make subjects have similar characteristics or covariates tend to be classified to the same group. This might make the estimation results more interpretable. Second, it would be possible to extend the proposed GGEE method for incomplete longitudinal data. Since the GGEE separately applies the standard GEE to each group, we can employ existing methodology to handle missing data in the standard GEE

method. Moreover, when the dimension of the regression coefficients is large, it would be better to conduct variable selection, which can be done by introducing a penalty function in the estimating equation as considered in Wang et al. (2012). Finally, instead of using working correlation matrices, it would be beneficial to consider quadratic inference functions (Qu et al., 2000), and develop the GGEE method with theoretical justifications. We leave the detailed investigation of these issues for interesting future works.

ACKNOWLEDGMENTS

This work is partially supported by the Japan Society for the Promotion of Science (JSPS KAKENHI), grant numbers: 18K12757, 19K23242, and 22K13375.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available in R code provided in the Supporting Information of this paper.

ORCID

Tsubasa Ito https://orcid.org/0000-0001-8021-9907
Shonosuke Sugasawa https://orcid.org/0000-0002-9495-4280

REFERENCES

Barban, N. & Billari, F.C. (2012) Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society, Series C*, 61, 765–784.

Bonhomme, S. & Manresa, E. (2015) Grouped pattern of heterogeneity in panel data. *Econometrica*, 83, 1147–1184.

Coffey, N., Hinde, J. & Holian, E. (2014) Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data. *Computational Statistics & Data Analysis*, 71, 14–29.

Field, C.A. & Welsh, A.H. (2007) Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3), 369–390.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T. & Kelderman, H. (2018) Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, 50(5), 2016–2034.

Gu, J. & Volgushev, S. (2019) Panel data quantile regression with grouped fixed effects. *Journal of Econometrics*, 213, 68–91.

Hajjem, A., Bellavance, F. & Larocque, D. (2011) Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, 81, 451–459.

Hajjem, A., Larocque, D. & Bellavance, F. (2017) Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126, 114–118.

Juster, F.T. & Suzman, R. (1995) An overview of the health and retirement study. *Journal of Human Resources*, 30, S7–S56.

Liang, L. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13–22.

Lin, C.C. & Ng, S. (2012) Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1, 42–55.

Liu, R., Shang, Z., Zhang, Y. & Zhou, Q. (2020) Identification and estimation in panel models with overspecified number of groups. *Journal of Econometrics*, 215(2), 574–590.

Nagin, D.S., Jones, B.L., Passos, V.L. & Tremblay, R.E. (2018) Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, 27, 2015–2023.

Ng, S.K. & McLachlan, G.J. (2014) Mixture models for clustering multilevel growth trajectories. *Computational Statistics & Data Analysis*, 71, 43–51.

Qu, A., Lindsay, B.G. & Li, B. (2000) Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4), 823–836.

Ram, N. & Grimm, K.J. (2009) Methods and measures: growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, 33(6), 565–576.

Rosen, O., Jiang, W. & Tanner, M.A. (2000) Mixtures of marginal models. *Biometrika*, 87, 391–404.

Rubin, D.B. & Wu, Y. (1997) Modeling schizophrenic behavior using general mixture components. *Biometrics*, 53, 243–261.

Sugasawa, S. (2021) Grouped heterogeneous mixture modeling for clustered data. *Journal of the American Statistical Association*, 116(534), 999–1010.

Sugasawa, S., Kobayashi, G. & Kawakubo, Y. (2019) Latent mixture modeling for clustered data. *Statistics and Computing*, 29, 537–548.

Sun, Z., Rosen, O. & Sampson, A.R. (2007) Multivariate bernoulli mixture models with application to postmortem tissue studies in schizophrenia. *Biometrics*, 63, 901–909.

Tang, X. & Qu, A. (2016) Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25, 1117–1137.

Tang, X., Xue, F. & Qu, A. (2021) Individualized multidirectional variable selection. *Journal of the American Statistical Association*, 116(535).

Vogt, M. & Linton, O. (2017) Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B*, 79, 5–27.

Wang, J. (2010) Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97, 893–904.

Wang, L. (2011) GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39, 389–417.

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

Wang, L., Zhou, J. & Qu, A. (2012) Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometreics*, 68, 353–360.

Wedderburn, R.W. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-newton method. *Biometrika*, 61, 439–447.

Xie, M. & Yang, Y. (2003) Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, 31(1), 310–347.

Zhang, Y., Wang, J. & Zhu, Z. (2019) Quantile-regression-based clustering for panel data. *Journal of Econometrics*, 213, 54–67.

Zhu, X. & Qu, A. (2018) Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, 12, 171–193.

Zhu, X., Tang, X. & Qu, A. (2021) Longitudinal clustering for heterogeneous binary data. *Statistica Sinica*, 31, 603–624.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Section 1-5 are available with this paper at the Biometrics website on Wiley Online Library.

R code implementing the proposed method is also available in the online library.

How to cite this article: Ito, T. & Sugasawa, S. (2023) Grouped generalized estimating equations for longitudinal data analysis. *Biometrics*, 79, 1868–1879. https://doi.org/10.1111/biom.13718