Mixture Modeling for Longitudinal Data

Xiwei Tang and Annie Qu

In this article, we propose an unbiased estimating equation approach for a two-component mixture model with correlated response data. We adapt the mixture-of-experts model and a generalized linear model for component distribution and mixing proportion, respectively. The new approach only requires marginal distributions of both component densities and latent variables. We use serial correlations from subjects’ subgroup memberships, which improves estimation efficiency and classification accuracy, and show that estimation consistency does not depend on the choice of the working correlation matrix. The proposed estimating equation is solved by an expectation-estimating-equation (EEE) algorithm. In the E-step of the EEE algorithm, we propose a joint imputation based on the conditional linear property for the multivariate Bernoulli distribution. In addition, we establish asymptotic properties for the proposed estimators and the convergence property using the EEE algorithm. Our method is compared to an existing competitive mixture model approach in both simulation studies and an election data application. Supplementary materials for this article are available online.

Key Words: Expectation-estimating-equation algorithm; Latent variable; Model-based clustering; Multivariate Bernoulli distribution; Unbiased estimating equation.

1. INTRODUCTION

The mixture model has been extensively applied in many fields due to its flexibility to capture the heterogeneity arising from subgroups (components) in the whole population. The mixture model can also be viewed as a two-level hierarchical structure with incomplete data, where the first level consists of latent variables indicating subjects’ subgroup memberships, and the second level consists of outcome variables. However, the individual’s subgroup membership has to be inferred from the location of the outcome response due to unobservable latent variables.

Existing methods of mixture model with covariates for independent data include, but are not limited to, Jacobs et al.’s (1991) mixture-of-experts for a mixture of component regressions, Jordan and Jacobs’s (1994) extension on the generalized linear model for mixing proportion. For correlated or longitudinal outcome data, Rubin and Wu (1997) introduced a linear random-effects model for components’ densities under the mixture framework. Alternatively, Rosen, Jiang, and Tanner (2000) proposed a generalized estimating equation for
component distributions to incorporate correlations. These approaches all assume that the latent variables are independent. However, modeling correlation from outcome variables only is not sufficient to address correlations from subgroup membership over time.

One notable approach to model the correlated latent structure is the hidden Markov chain model (Qian and Titterington 1992). However, this is not applicable for longitudinal data since the Markov chain assumption does not hold or approximate some common correlation structure in longitudinal data such as exchangeable structure. Another well-known approach of mixture modeling for longitudinal data is related to growth-curve mixture modeling (Muthén and Muthén 2000), where the subject’s group membership is fixed over time, while different trajectory classes represent different mixture components. Huang et al. (2014) proposed a kernel smoothing method for a mixture of Gaussian processes incorporating both functional and heterogeneous types of dense longitudinal data. In addition, to incorporate the individual effects, Sun, Rosen, and Sampson (2007) proposed a multivariate Bernoulli mixture model by using random effects in the generalized linear model for mixing proportion.

Although the mixed-effects model is widely used to handle dependent data, incorporating serial correlation for latent variables in mixture modeling for longitudinal data is still limited. In this article, we are interested in developing an efficient method in mixture modeling of longitudinal data, where within-subject subgroup memberships could be correlated.

One challenge in formulating a mixture model for dependent data is that the joint likelihood function usually does not have an explicit form, because the latent variables are correlated categorical variables. In addition, the latent indicator variables are unobservable and require imputation. Although one can assume independent structure for the latent variables, the estimation efficiency will be compromised.

In this article, we allow group memberships to change over time in addition to taking serial correlation into account. We adopt the generalized estimating equation (GEE) approach (Liang and Zeger 1986) for both component distributions and mixing proportion in a two-component mixture model. This can be accomplished by treating the estimating equations for incomplete data as the conditional expectations of those for complete data. Specifically, we apply the expectation-estimating-equation (EEE) algorithm to solve the equations. To impute the latent variable in the E-step, we provide an approximation method to calculate the joint conditional expectation using the conditional linear property for the correlated binary variables (Qaqish 2003; Qu, Lindsay, and Lu 2010).

In contrast to the joint likelihood approach, the proposed method only requires the marginal distributions for both components’ densities and latent variables. In the estimating step, we fully use the serial correlation while treating it as a working structure. Allowing different working correlation structures enables one to incorporate various correlated latent structures, although the proposed method does not require to know the true correlation structure to produce consistent estimators of mean parameters. However, if the correlation structure is correctly or closely specified, we gain efficiency of the parameter’s estimation and improve the classification accuracy from a model-based clustering.

The article is organized as follows: Section 2 introduces some notation and background knowledge; Section 3 presents our method and provides some theoretical results, as well as the EEE algorithm and imputation methods; Section 4 provides simulation studies; Section
5 considers an application to the Election data; and Section 6 offers a brief summary and some further discussion.

2. BACKGROUND AND NOTATION

For longitudinal data, let \( y_{it} \) be the response of subject \( i \) at time \( t \), and \( x_{it} \) be a \( p \)-dimensional covariate vector, where \( i = 1, \ldots, n \) and \( t = 1, \ldots, T_i \). For ease of notation, we first assume \( T_i = T \) for all \( i \) representing a balanced data case, where an unbalanced data case will be discussed later. Denote \( y_i = (y_{i1}, \ldots, y_{iT}) \) as a \( T \times 1 \) response vector, \( x_i = (x_{i1}, \ldots, x_{iT}) \) as a \( p \times T \) covariate matrix. In a two-component mixture model, let \( z_{it} \) denote the binary latent variable associated with \( y_{it} \). Let \( \mu_r(\cdot) \) be an inverse link function satisfying \( E[y_{it}|x_{it}, z_{it} = 2 - r] = \mu_r(\beta_r^t x_{it}) \) \((r = 1, 2)\), where \( \beta_r \) is a \( p \)-dimensional parameter, and \( \mu_{zi} = E[y_{it}|x_{it}, z_{it} = 2 - r] \).

In this article, we consider a two-component mixture model. The outcome variable \( y_{it} \) is from either one of the two subgroup populations and thus is assumed to follow a mixture distribution \( \pi_{it} f_1(y_{it} | x_{it}, \beta_1, \phi_1) + (1 - \pi_{it}) f_2(y_{it} | x_{it}, \beta_2, \phi_2) \), where \( \pi_{it} \) is a mixing proportion defined as the probability of a response \( y_{it} \) from a specified subgroup, \( f_1(y_{it} | x_{it}, \beta_1, \phi_1) \) and \( f_2(y_{it} | x_{it}, \beta_2, \phi_2) \) are the components’ distributions, and \( \phi = (\phi_1, \phi_2) \) is the dispersion parameter. This two-component mixture model can be regarded as a hierarchical structure. Specifically,

\[
z_{it} \sim \text{Bernoulli}(\pi_{it}), \quad y_{it} | (z_{it} = 1) \sim f_1(y_{it} | x_{it}, \beta_1, \phi_1)
\]

and

\[
y_{it} | (z_{it} = 0) \sim f_2(y_{it} | x_{it}, \beta_2, \phi_2),
\]

where a latent variable \( z_{it} \) represents a subgroup membership indicator.

The most common assumption for this hierarchical model is that the outcome variable \( y_{it} \)'s are conditionally independent given the subgroup membership label \( z_{it} \), then the joint likelihood density function of complete data \((y_i, z_i)\) can be written as

\[
f_c(y_i, z_i) = f(y_i | z_i, \beta_1, \beta_2, \phi, x_i) p(z_i | \pi_i, x_i)
\]

\[
= \left( \prod_{t=1}^{T} f_1(y_{it} | \beta_1, \phi_1, x_{it})^{z_{it}} f_2(y_{it} | \beta_2, \phi_2, x_{it})^{1-z_{it}} \right) p(z_i | \pi_i, x_i).
\]

Under the mixture framework, the latent variable \( z_{it} \) is missing. The simplest way to get around the latent structure is to assume independence of subgroup memberships at different times within each subject and thus \( p(z_i | \pi_i) = \prod_{t=1}^{T} \pi_i^{z_{it}} (1 - \pi_i)^{1-z_{it}} \). In the independent mixture model, the correlations among longitudinal observations and their subgroup memberships are not fully used.

Jordan and Jacobs (1994) proposed a hierarchical mixture-of-experts model which takes covariates into consideration for both latent variable and component distributions, where the latent variable \( z_{it} \) is assumed to follow a logistic model: \( \pi_{it} = \exp(\eta^t x_{it})/(1 + \exp(\eta^t x_{it})) \), and components’ densities have a mean regression form. Here, we denote \( \theta = (\eta^t, \beta_1^t, \beta_2^t)^t \) as a grand mean parameter vector, then the log of joint likelihood for
complete data is
\[
L_c(y, z, \theta, \phi) = \sum_{i=1}^{n} \sum_{t=1}^{T} \left( z_{it} \{ \log \pi_{it} + \log f_1(y_{it} | x_{it}, \beta_1, \phi_1) \} + (1 - z_{it}) \{ \log (1 - \pi_{it}) \right.
\]
\[
+ \log f_2(y_{it} | x_{it}, \beta_2, \phi_2) \right),
\]
(2)
under the independence assumption within subject.

For classical mixture models where the latent variable $z_{it}$ is missing, the maximum likelihood estimator (MLE) is typically computed through the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). The EM algorithm proceeds iteratively in two steps. In the E-step, we take the conditional expectation of complete-data log-likelihood given observed outcome response $y$ and current parameters’ estimates ($\theta^{(k)}, \phi^{(k)}$).

Since the complete-data log-likelihood $L_c(y, z, \theta, \phi)$ in (2) is a linear function of latent variable $z_{it}$, the E-step only requires us to impute $z_{it}$ by its conditional expectation $w_{it}^{(k)} = E[z_{it} | y_{it}, x_{it}, \theta^{(k)}, \phi^{(k)}]$ denoted as the mixing weight. In the M-step, we obtain the $k + 1$th updates of parameters’ estimations by maximizing
\[
L(\theta, \phi | \theta^{(k)}, \phi^{(k)}) = \sum_{i=1}^{n} \sum_{t=1}^{T} \left( w_{it}^{(k)} \{ \log \pi_{it} + \log f_1(y_{it} | x_{it}, \beta_1, \phi_1) \}
\]
\[
+ (1 - w_{it}^{(k)}) \{ \log (1 - \pi_{it}) + \log f_2(y_{it} | x_{it}, \beta_2, \phi_2) \} \right),
\]
(3)
It has been shown that the obtained series ($\theta^{(k)}, \phi^{(k)}$) in the EM algorithm converges to the MLE estimator of the mixture distribution $f(y, \theta, \phi) = \int_z f_c(y, z, \theta, \phi) dz$ (Dempster, Laird, and Rubin 1977).

3. UNBIASED ESTIMATING EQUATIONS FOR MIXTURE MODELING

3.1 UNBIASED ESTIMATING EQUATIONS

Due to the correlated nature of longitudinal data, it is important to incorporate correlation for mixture modeling as it provides more efficient estimation and benefits the longitudinal subgrouping. In this section, we first propose an efficient unbiased estimating equation for complete data, and then extend it to the incomplete-data mixture model.

To account for within-subject correlation among group memberships in (1), we have to deal with the joint likelihood function $p(z_i | \pi_i)$ of the latent multivariate Bernoulli variable $z_i$. Unlike the multivariate Gaussian distribution, there is no explicit form of the likelihood function for the multivariate Bernoulli distribution. The unknown membership in the mixture models makes the incomplete-data likelihood function even more complicated since the latent $z_i$ needs to be integrated out from the complete-data likelihood function. One possible approximation is to apply the Bahadur representation (Bahadur 1961) for the multivariate Bernoulli distribution by ignoring high-order moments. However, there are several drawbacks such that the correlations are constrained through marginal means and
covariates in a complicated way (Diggle, Liang, and Zeger 1994). In addition, the dimension of the correlation parameters could be very high when the repeated measurement size $T$ is large, which leads to increasing computational demands.

To address this problem, we first introduce an unbiased estimating equation approach and establish it for the complete-data model. The proposed unbiased estimating equation regarding the mean parameter $\theta$ is formulated through the log-likelihood function $L_c$ in (2). Maximizing $L_c$ with respect to $\theta$ by solving the corresponding score equation is equivalent to solving the quasi-likelihood equation (Wedderburn 1974; McCullagh 1983) as follows:

$$S^c(\theta) = \sum_{i=1}^{n} S^c_i(\theta) = \left( \sum_{i=1}^{n} \left( \frac{\partial \pi_i}{\partial \theta} \right)^T V_i^{-1} (z_i - \pi_i) \right) = 0,$$

where $\mu_r = (\mu_{r1}, \ldots, \mu_{rT})^T$, $\mu_{rit} = E[y_{it}|z_{it} = 2 - r] = \mu_r(\beta_r^T x_{it})$ ($r = 1, 2$) are the corresponding mean functions for components’ densities, $Z_i = \text{diag}(z_{i1}, \ldots, z_{iT})$ is a diagonal matrix of corresponding latent labels, and $U_{1i}$ and $U_{2i}$ are the diagonal covariance matrices of component densities respectively: $U_{ri} = \text{var}(y_i|z_i = 2 - r)$. The covariance matrices $U_{1i}$ and $U_{2i}$ could be functions of both mean parameters $\mu_{1i}$, $\mu_{2i}$ and dispersion parameter $\phi$. As a result, given $\theta$, the dispersion parameter $\phi$ can be consistently estimated via the second moment conditions of component distributions. For the independent model, $V_i = \text{diag}(\text{var}(z_{i1}), \ldots, \text{var}(z_{iT}))$.

To account for serial correlation induced by latent variable $z_i$, we assume $V_i = A_i^T R(\rho) A_i$ as analogue to the generalized estimating equation (GEE) (Liang and Zeger 1986), where $A_i = \text{diag}(\pi_{i1}(1 - \pi_{i1}), \ldots, \pi_{iT}(1 - \pi_{iT}))$ is a diagonal matrix of marginal variance of $z_i$, and $R(\rho)$ is a working correlation matrix. If $R$ is the true correlation matrix for $z_i$, then $V_i = \text{cov}(z_i)$.

In the following, we extend the proposed estimating equation to the incomplete-data model. To handle the missing latent indicator $z_i$, we take the conditional expectations for the equations in (4) given the outcome observation $y$, and therefore construct the unbiased estimating equations for the incomplete data as

$$S(\theta) = \mathbb{E}_z \left[ \sum_{i=1}^{n} S^c_i(\theta)|y, \theta \right]$$

$$= \sum_{i=1}^{n} S_i(\theta|\theta) = \left( \sum_{i=1}^{n} \left( \frac{\partial \pi_i}{\partial \theta} \right)^T V_i^{-1} (w_i - \pi_i) \right) = 0,$$

where $w_i = \mathbb{E}_z[z_i|y, \theta]$ is the mixing weight representing the inferred probability of the subgroup memberships, and $W_i$ is a diagonal matrix of corresponding mixing weights associated with $y_i$. Similar to the GEE method, our approach can handle unbalanced data if the missing mechanism is missing completely at random (MCAR; Rubin 1976). If the missing mechanism is not MCAR, then the estimating equation estimators could be biased and inefficient, and weighted generalized estimating equations (WGEE) can be employed to deal with the missing not completely at random cases (Robins, Rotnitzky, and Zhao 1995; Rotnitzky, Robins, and Zhao 1998).
To solve the estimating Equation (5), we present a two-step expectation-estimating-equation (EEE) algorithm. Analog to the EM algorithm, the EEE algorithm follows an iterative estimating process. Specifically, at the \((k+1)\)th step, based on the current estimation of parameter \(\theta^{(k)}\), we impute \(w_i^{(k)} = E_z[z_i|y, \theta^{(k)}]\), and then update \(\theta^{(k+1)}\) by solving the estimating equation

\[
S(\theta | \theta^{(k)}) = \sum_{i=1}^{n} S_i(\theta | \theta^{(k)}) = \left( \begin{array}{c}
\sum_{i=1}^{n} \left( \frac{\partial \pi_i}{\partial \eta_i} \right)^T V_i^{-1} (u_i^{(k)} - \pi_i) \\
\sum_{i=1}^{n} \left( \frac{\partial \mu_1}{\partial \beta_1} \right)^T U_{1i}^{-1} W_i^{(k)} (y_i - \mu_1) \\
\sum_{i=1}^{n} \left( \frac{\partial \mu_2}{\partial \beta_2} \right)^T U_{2i}^{-1} (1 - W_i^{(k)}) (y_i - \mu_2)
\end{array} \right) = 0. \tag{6}
\]

A more detailed EEE algorithm is provided in Section 3.3.

The main difference between the above method and Rosen, Jiang, and Tanner (2000) approach is that we incorporate an additional estimating equation associated with the mixing proportion. This enables us to use the longitudinal correlations among the unobservable subgroup memberships from the same subject. Furthermore, in our estimating Equation (5), we can also use the correlation information from the outcome variable \(y_i\) simultaneously through the working correlation structures of \(U_{1i}\) and \(U_{2i}\). In this article, we focus on using the subject’s group membership correlation over time. We assume that \(y_{it}\)’s given \(z_{it}\)’s are conditionally independent within the same subject, and thus \(U_{1i}\) and \(U_{2i}\) are set to be diagonal.

Indeed, the proposed model assuming working correlation structure has the same identifiability problem as the independent model, which is equivalent to the mixtures-of-experts model. We can refer to Jiang and Tanner’s (1999) discussion for the identifiability problem of the mixtures-of-experts model.

### 3.2 Asymptotic Properties

In this article, we use the latent variable’s serial correlation information to improve the parameter estimation and the subject’s group membership prediction. We first examine the optimization properties for the complete-data equations.

Once we establish the optimal estimating equation \(S'(z, y, \theta)\) for complete data, then the conditional expectation \(S(y, \theta) = E[S'|y]\) is the best estimating equation for the incomplete data such that the \(L_2\) norm \(\|S'(z, y, \theta) - S(y, \theta)\|\) is minimized. In other words, if the proposed \(S'(z, y, \theta)\) is efficient enough, then \(S(y, \theta)\) will inherit some efficiency from the complete-data model. In the following proposition, we establish the optimality of the unbiased estimating equation (4) for complete data.

**Proposition 1.** If the working correlation structure is correctly specified, that is, \(V_i = \text{var}(z_i)\), the estimating equations in (4) are the optimal linear estimating equations with respect to \((z_i, y_i)\), in the sense that the asymptotic variance of the estimator solved by (4) reaches the minimum.

The proof is provided in the Appendix (A.2).

Indeed, incorporating the serial correlation information is more important for the mixture model since \(z_i\) cannot be observed. For the complete-data estimating equations in (4), the
first set of equations with respect to the group membership and the other two equations
associated with component densities are uncorrelated with each other. Therefore, the induced
serial correlations only affect the estimation of mixing proportion parameters in \( \pi(\eta) \), but
have no influence on estimating the component parameters in \( \mu_1(\beta_1) \) and \( \mu_2(\beta_2) \). However,
for the incomplete data, the three equations in (5) are correlated because the imputed term
\( w_i \) is a function of \( y_i \). For the longitudinal data of the subjects, taking advantage of the
information at the other time points through accounting for the serial correlation improves
the estimation efficiency of both the component parameters and the mixing proportion
parameters.

It is known that the consistency of the above estimator only depends on correct spec-
ification of the mean functions, but does not rely on correct specification of the working
correlation structure nor on the estimation of the correlation parameters \( \rho \). This is one
desirable property for the proposed approach since the serial correlation induced from un-
observable latent variables could be difficult to model and estimate precisely. In contrast
to the full joint likelihood approach, we only require the second moment approximation.
Furthermore, the following theorem establishes the asymptotic properties of the proposed
estimators.

**Theorem 1.** Let \( \gamma = (\theta', \phi')' \) denote a grand parameter vector, in the proposed unbiased
estimating Equation (5), assuming that the following regularity conditions are satisfied:

(i) \( \hat{\phi} \) is consistently estimated given \( \theta \) via some unbiased estimating equation
\( \sum_{i=1}^{n} H_i(\phi|\theta) = 0 \);

(ii) \( \hat{\rho} \) is \( n^{1/2} \)-consistent given \( \gamma \), denoted as \( \hat{\rho}(\gamma) \);

(iii) \( \| \partial \hat{\rho}/\partial \gamma \| \leq O_p(1) \).

Let \( G_i(\gamma) = (S_i(\theta)', H_i(\phi|\theta)\gamma)' \) be the augmented unbiased estimating equation, then the
asymptotic distribution of \( \hat{\gamma} \) obtained from \( \sum_{i=1}^{n} G_i = 0 \) is: \( n^{1/2}(\hat{\gamma} - \gamma) \rightarrow N(0, V_{\gamma}) \),
where the asymptotic covariance matrix \( V_{\gamma} \) is given by

\[
V_{\gamma} = \lim_{n \to \infty} n \left( \sum_{i=1}^{n} \nabla G_i(\gamma) \right)^{-1} \left( \sum_{i=1}^{n} \text{cov}(G_i(\gamma)) \right) \left( \sum_{i=1}^{n} \nabla G_i(\gamma) \right)^{-T},
\]

\( \nabla G_i \) is the gradient of \( G_i \) with respect to \( \gamma \).

Theorem 1 is established based on the following unbiased estimating equations:

\[
E[(w_i - \pi_i)] = E[E[z_i|y_i] - \pi_i] = E[z_i] - \pi_i = 0;
\]

\[
E[W_i(y_i - \mu_{1i})] = E[E[Z_i|y_i](y_i - \mu_{1i})] = E[Z_i(y_i - \mu_{1i})] = 0;
\]

\[
E[(1 - W_i)(y_i - \mu_{2i})] = E[(1 - E[Z_i|y_i])(y_i - \mu_{2i})] = E[(1 - Z_i)(y_i - \mu_{2i})] = 0.
\]

A sketch of the proof of Theorem 1 is given in the Appendix. Noting that even if \( w_i \) is
imputed through the marginal expectation \( w_{ii} = E[z_{ii}|y_{ii}] \), the estimating equations in (5)
are still unbiased. The estimating equation \( H_i \) for the dispersion parameter is established
based on the second moment condition provided in the Appendix (A.1).
In practice, a consistent variance estimator of \( \hat{\gamma} \) can be obtained by replacing \( \text{cov}(G_i(y)) \) by \( \hat{G}_i\hat{G}_i \) and \( \nabla G_i \) by \( \nabla \hat{G}_i \) at the estimates of \( \hat{\theta}, \hat{\rho}, \hat{\phi} \). One practical difficulty arises from getting an explicit form of the gradient \( \nabla \hat{G}_i \) in the calculation of mixing weights \( w_i \). The imputation term \( w_i \) is the posterior probability of \( z_i \) containing both the response value \( y_i \) and all the parameters \( \theta, \phi, \rho \). The complicated form of \( w_i \) makes it difficult to calculate the exact derivatives, especially when the conditional expectation \( w_i = E[z_i|y_i] \) requires specifying the joint likelihood function for \( z_i \). We present some numerical approximation methods to calculate \( w_i \) and \( \nabla \hat{G}_i \). More details are provided in Section 3.3.

### 3.3 Algorithm and Implementation

In this section, we provide the two-step iterative EEE algorithm and its theoretical properties. In addition, the details of marginal imputation and the joint imputation methods are provided.

---

**Algorithm 1: Expectation-Estimating-Equation (EEE)**

**Step 1**: Set the initial values of the mean regression parameter \( \theta \) and the dispersion parameter \( \phi \):

\[
\gamma^{(0)} = (\theta^{(0)}, \phi^{(0)});
\]

**Step 2 (E-Step)**: Impute the mixing weights \( w_{it}^{(k)} = E[z_{it}|y_i] \) given current estimates \( \gamma^{(k)} \);

**Step 3 (M-Step)**: Given the current imputed mixing weights \( w_{it}^{(k)} \),

(i) update \( \theta^{(k+1)} \) by solving the estimating equation in (6), and

(ii) update \( \phi^{(k+1)} \) given \( \theta^{(k+1)} \).

**Step 4**: Return to Step 2 if \( \| \gamma^{(k+1)} - \gamma^{(k)} \| > \epsilon \), where \( \epsilon \) is the chosen tolerance level.

---

In Step 3 of Algorithm 1, with the current mixing weights, we apply the Newton-Raphson method to solve the estimating equations in (6) to obtain the updates \( \theta^{(k+1)} \) for the mean parameter. The updated dispersion parameter \( \phi^{(k+1)} \) can be estimated using the second moment conditions given \( \theta^{(k+1)} \). See Appendix (A.1) for more details. In the Newton-Raphson algorithm, the first derivative of the estimating function can be calculated as a simple diagonal matrix:

\[
\begin{align*}
\text{diag} & \left( \sum_{i=1}^{n} \left( \frac{\partial \pi_i}{\partial \eta} \right)^T V_i^{-1} \left( \frac{\partial \pi_i}{\partial \eta} \right), \sum_{i=1}^{n} \left( \frac{\partial \mu_{1i}}{\partial \beta_1} \right)^T U_{1i}^{-1} W_i^{(k)} \left( \frac{\partial \mu_{1i}}{\partial \beta_1} \right), \\
& \sum_{i=1}^{n} \left( \frac{\partial \mu_{2i}}{\partial \beta_2} \right)^T U_{2i}^{-1} (1 - W_i^{(k)}) \left( \frac{\partial \mu_{2i}}{\partial \beta_2} \right) \right).
\end{align*}
\]

For the EEE Algorithm 1, the estimating function in (5) can be regarded as a bivariate function \( S(\theta|\lambda) \) restricted on the subspace \( \theta = \lambda \), which is denoted as \( S(\theta|\theta) \). In fact, the first argument \( \theta \) of \( S(\theta|\theta) \) comes from the mean regression part, while the second argument comes from the conditional expectations. The EEE Algorithm 1 is simply updating the estimation of \( \theta \) by solving \( S(\theta|\theta^{(k)}) = 0 \) in (6). If the iterative sequence of the estimator obtained from the EEE algorithm converges, it will converge to the solution of the original
equation $S(\theta|\theta) = 0$ which is guaranteed by the continuity. The following lemma provides a local convergence of the estimator based on the Algorithm 1 in a more general form.

**Lemma 1.** Suppose that $S(\theta|\lambda)$ is a bivariate function on the space $\Theta \times \Theta$ satisfying the following regularity conditions:

(a) $S(\theta|\lambda)$ is continuous and both $\frac{\partial S}{\partial \theta}$ and $\frac{\partial S}{\partial \lambda}$ exist;

(b) in a neighborhood of $(\theta_0, \theta_0)$, where $S(\theta_0|\theta_0) = 0$, we have $\| (\frac{\partial S}{\partial \theta})^{-1} \cdot \frac{\partial S}{\partial \lambda} \|_2 < 1$ hold;

then the iterative sequence of estimator $\{\theta^{(k)}\}$ solved by $S(\theta^{(k+1)}|\theta^{(k)}) = 0$ converges to $\theta_0$.

The proof of Lemma 1 is provided in the Appendix. Lemma 1 provides a sufficient but not necessary condition for the algorithm’s convergence. Indeed, the algorithm convergence problem is also associated with general convergence theory regarding the iterative solutions to nonlinear equations. More detailed discussion of this type of algorithm can be found in Ortega and Rheinboldt (1970).

In practice, it is possible to have multiple roots for the proposed estimating equations. One suggested method is to choose the root closest to the independent estimator which is consistent (Small, Wang, and Yang 2000). Seo and Kim (2012) proposed to exclude all singular solutions and therefore reduce the risk of selecting spurious roots of the likelihood equation. Their method is powerful in selecting reasonable roots for the independent finite normal mixture model. In addition, we may also need to provide a “good” initial value for the iterations to converge. One way is to select different initial values randomly until the algorithm converges; another choice is to set the initial value as the estimator obtained from the existing independent mixture model.

In Algorithm 1, the E-step (Step 2) requires one to impute the missing latent variable $z_i$ through its conditional expectation $w_i = E[z_i|y, \theta, \phi]$. However, this joint conditional expectation involves the joint density of the multivariate binary variable $z_i$, which is very difficult to calculate. One possible way is to use the marginal imputation $w_{it} = E[z_{it}|y_{it}, \theta, \phi]$ since we only assume a marginal distribution for the individual observation $(y_{it}, z_{it})$ in our approach. The marginal imputation

$$w_{it} = \frac{\pi_{it} f_1(y_{it}, \theta_1, \phi_1)}{\pi_{it} f_1(y_{it}, \theta_1, \phi_1) + (1 - \pi_{it}) f_2(y_{it}, \theta_2, \phi_2)}$$

guarantees the estimating equations in (5) to be unbiased, where $f_1$ and $f_2$ are component densities. In fact, the imputation $w_{it}$ provides the inferred subgroup membership of the outcome observation $y_{it}$. The drawback of the marginal imputation is that it only uses local information to infer the group membership. If two subgroups are well-separated, that is, most $w_{it}$ are either close to 1 or 0, then the marginal imputation is sufficient since the local information $y_{it}$ dominates the subgroup membership’s prediction. However, if two subgroups are not well-separated, then the marginal imputation does not fully use the within-subject correlation to improve the subgroup classification.

Therefore, we provide an alternative approximation of the joint imputation $w_{it}^* = E[z_{it}|y_{it}]$ which relies only on the second moment condition of $z_i$. To predict the subgroup membership of $y_{it}$, we consider the conditional expectation $E[z_{it}|y_{it}, z_{it(-t)}]$ based
on local observation $y_{it}$ and all latent group labels $z_{it(-t)}$ at other time points for the $i$th subject, where $z_{it(-t)} = (z_{i1}, \ldots, z_{i(t-1)}, z_{i(t+1)}, \ldots, z_{iT})$. Since

$$P(z_{it}|y_{it}, z_{it(-t)}) \propto f(y_{it}|z_{it})P(z_{it}|z_{it(-t)}),$$

then $w_{it}^s = \pi_{it}^s f_1(y_{it}, \beta_1, \phi_1)/[\pi_{it}^s f_1(y_{it}, \beta_1, \phi_1) + (1 - \pi_{it}^s) f_2(y_{it}, \beta_2, \phi_2)]$, where $\pi_{it}^s = E[z_{it}|z_{it(-t)}]$. By taking account of $z_{it(-t)}$, this imputation $w_{it}^s$ allows us to use the group membership information from other time points within the same subject as well.

However, the exact value of $E[z_{it}|z_{it(-t)}]$ still requires the joint likelihood function for $z_i$. Motivated by Qaqish’s (2003) conditional linear family for the multivariate binary distribution in generating correlated binary data, we consider a linear approximate:

$$\pi_{it}^s = E[z_{it}|z_{it(-t)}] = \pi_{it} + b_{it}^T(z_{it(-t)} - \pi_{it(-t)}),$$

where $\pi_{it} = E[z_{it}]$, $\pi_{it(-t)} = E[z_{it(-t)}]$, and $b_{it}$ is a $(T - 1)$-dimensional coefficient vector. Based on the fact that for any two random variable $X$ and $Y$, $\text{cov}(X, Y) = \text{cov}(X, E[Y|X])$, we have:

$$\text{cov}(z_{it(-t)}, z_{it}) = \text{cov}(z_{it(-t)}, \pi_{it}^s) = \text{cov}(z_{it(-t)})b_{it}.$$

Denote $\text{cov}(z_{it(-t)}, z_{it})$ as $s_{it}$ and $\text{cov}(z_{it(-t)})$ as $V_{it(-t)}$, then $b_{it} = V_{it(-t)}^{-1}s_{it}$. Here, both $s_{it}$ and $V_{it(-t)}$ could be obtained from the second moment condition $\text{cov}(z_{it})$.

To implement Algorithm 1, at the $(k + 1)$th step, we use the current estimators $\theta^{(k)}$ and $V_{it(-t)}^{(k)}$ to obtain $\pi_{it}^{(k+1)}$ and thus impute $w_{it}^{(k+1)}$. Here, the true latent label $z_{it(-t)}$ is replaced by the last prediction $\hat{z}_{it(-t)}^{(k)}$. A similar technique is also used by Qian and Titterington (1992) to approximate $P(z_{it}|\hat{z}_{it(-t)}^{(k)})$ in a hidden Markov field modeling.

In Section 3.2, we mentioned that to obtain the robust variance estimator of $\hat{y}$, it requires us to calculate the gradient of $G_i(y)$. Here, we take the numerical approximation of the gradient $\nabla G_i(y)$ by $\frac{G_i(y^{(k+1)}) - G_i(y^{(k)})}{y^{(k+1)} - y^{(k)}}$, where $y^{(k)}$ and $y^{(k+1)}$ are obtained from the previous two adjacent iterations, and the $i$th component of $\nabla G_i(y)$ corresponds to the ratio between the $i$th and $j$th components of $S_i(y^{(k+1)}) - S_i(y^{(k)})$ and $y^{(k+1)} - y^{(k)}$, respectively.

## 4. SIMULATION STUDY

In this section, we conduct two simulation studies to illustrate the performance of the proposed method on mixture modeling for longitudinal data. We are particularly interested in comparing the performance of our method to other approaches when the serial correlation is induced by latent variables. Our simulation results show that we can gain efficiency on parameter estimation by using correlation information.

In the first simulation study, we consider a two-component mixture of univariate normal distribution, and mainly focus on the estimating performance of the mixing proportion parameters with different dependence structures. In the second simulation study, we consider a mixture of two regression models, where the separation levels of two subgroups will change over time. The proposed method has extra power in both estimation and prediction, especially at poorly separated time points.
4.1 Study 1: Two-Component Mixture of Univariate Normal Densities

In this simulation study, we first generate component indicator variables $Z_i$ from Bernoulli distributions for subjects $i = 1, \ldots, n$. For each $Z_i$, we assume there are $T$ repeated measurements over time, where $Z_i = (Z_{i1}, \ldots, Z_{iT})$, and $Z_{it}$ is a binary variable following a logistic regression with time covariate. That is, $E[Z_{it}] = \pi_{it}$ and $\log(\pi_{it}/(1-\pi_{it})) = \eta_0 + \eta_1 t$. Conditional on $Z_{it}$, we generate outcome response $Y_{it}$ from two univariate normal distributions: $Y_{it}(Z_{it} = 1) \sim N(\mu_1, \sigma_1^2)$ and $Y_{it}(Z_{it} = 0) \sim N(\mu_2, \sigma_2^2)$.

For latent variable $Z_i$, we assume independence among different subjects, but a common correlation structure (either AR-1 or exchangeable) within a subject over time. For outcome response $Y_i$, we assume independence among different subjects and conditional independence within a subject given component indicator $Z_i$. In our simulation studies, we generate binary responses through the R package “mvtBinaryEP”.

We let the sample size $n = 100$ and time points $T = 10$, and the mixing proportion parameters in the logistic model are set to be $(\eta_0, \eta_1) = (-0.3, 0.5)$, which allows certain correlations among multivariate binary variables. In addition, we set variance parameters as $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

The separation between two normal homoscedastic components could be accessed by $\Delta = |\mu_1 - \mu_2|/\sigma$, defined as the Mahalanobis distance between two normal mixture distributions (McLachlan and Peel 2000). We investigate two settings with component means $(\mu_1, \mu_2) = (-1.5, 1.5)$ and $(\mu_1, \mu_2) = (-1.2, 1.2)$ to represent well-separated and poorly separated bimodal densities, respectively. In addition, we also simulate a heterogeneous case with $(\mu_1, \mu_2) = (-1.5, 1.5)$ and $(\sigma_1, \sigma_2) = (1, 1.5)$. The simulation results are similar to those for the homogeneous case, and are thus omitted due to space limitations.

In the following simulation studies, we choose the initial value randomly in the neighborhood of the independent estimator until the EEE algorithm converges. The estimation results are summarized based on 1000 replicates. In our simulations, the empirical standard errors are quite close to the standard errors calculated from the robust sandwich variance, and therefore we only provide the empirical standard errors in our tables. In practice, label-switching issues might arise, especially for Bayesian mixture models. Yao (2012) and Yao and Lindsay (2009) discussed many feasible labeling methods. In our simulation study, we solve the labeling problem by imposing an ordered constraint on components’ mean parameters.

We compare the estimators based on the proposed unbiased estimating equation with either joint imputation (UEE\text{Joint}) or marginal imputation (UEE\text{Mar}), with working correlation structure of either exchangeable, AR-1 or unstructured, the mixtures-of-experts model (Jacobs et al. 1991) which is the same as the independent estimating equation (UEE\text{Ind}), and the oracle estimators (Oracle) assuming the true values of latent variable $Z_i$ are known. The oracle estimator serves as a benchmark estimator, where the generalized estimating equation (GEE) uses the correlation structure for the binary data $Z_i$, and the MLE estimators of component parameters are obtained given $Z_i$. In the following tables, we denote the correlation structure in the superscript such as UEE\text{AR1\text{Joint}}. In the text, to avoid redundant notation, we omit the superscript if the correlation structure is correctly specified.

In addition, we also compare the proposed method with the random-effects model (Sun, Rosen, and Sampson 2007). In the random-effects model, we assume $E[Z_{it}] = \pi_{it}$ and
Table 1. The parameter estimators and their empirical standard errors (provided in the subscripts) for a well-separated two-component univariate normal mixture model, based on 1000 replicates.

|       | True  | UEE\textsuperscript{Ind} | UEE\textsuperscript{Ex} \textsuperscript{Mar} | UEE\textsuperscript{Ex} \textsuperscript{Joint} | UEE\textsuperscript{AR1} \textsuperscript{Mar} | UEE\textsuperscript{AR1} \textsuperscript{Joint} | UEE\textsuperscript{Uns} \textsuperscript{Mar} | UEE\textsuperscript{Uns} \textsuperscript{Joint} | RE | Oracle |
|-------|-------|---------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|-----|---------|
| $\mu_1$ | $-1.50_{0.09}$ | $-1.50_{0.09}$ | $-1.50_{0.05}$ | $-1.48_{0.09}$ | $-1.50_{0.06}$ | $-1.50_{0.10}$ | $-1.49_{0.07}$ | $-1.52_{0.05}$ | $-1.50_{0.05}$ |
| $\sigma_1$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.04}$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.01_{0.06}$ | $0.98_{0.04}$ | $1.00_{0.03}$ |
| Exch $\mu_2$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.05}$ | $1.48_{0.10}$ | $1.49_{0.06}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.05}$ | $1.50_{0.05}$ |
| $\sigma_2$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.04}$ | $1.01_{0.06}$ | $1.01_{0.06}$ | $1.01_{0.06}$ | $0.99_{0.06}$ | $0.98_{0.04}$ | $1.00_{0.03}$ |
| $\eta_0$ | $-0.31_{0.22}$ | $-0.30_{0.19}$ | $-0.30_{0.23}$ | $-0.32_{0.23}$ | $-0.32_{0.20}$ | $-0.31_{0.22}$ | $-0.31_{0.21}$ | $-0.88_{0.22}$ | $-0.30_{0.19}$ |
| $\eta_1$ | $0.52_{0.16}$ | $0.52_{0.16}$ | $0.50_{0.13}$ | $0.51_{0.20}$ | $0.51_{0.17}$ | $0.50_{0.20}$ | $0.50_{0.18}$ | $1.50_{0.48}$ | $0.50_{0.12}$ |
| $\mu_1$ | $-1.50_{0.10}$ | $-1.50_{0.10}$ | $-1.50_{0.09}$ | $1.50_{0.05}$ | $-1.50_{0.09}$ | $-1.50_{0.09}$ | $-1.50_{0.07}$ | $-1.51_{0.05}$ | $-1.50_{0.05}$ |
| $\sigma_1$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.04}$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.05}$ | $1.00_{0.06}$ | $1.00_{0.03}$ |
| AR1 $\mu_2$ | $1.50_{0.11}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.50_{0.09}$ | $1.52_{0.05}$ | $1.50_{0.05}$ |
| $\sigma_2$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.04}$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.06}$ | $1.00_{0.05}$ | $1.00_{0.06}$ | $1.00_{0.04}$ |
| $\eta_0$ | $-0.31_{0.26}$ | $-0.33_{0.27}$ | $-0.33_{0.27}$ | $-0.31_{0.26}$ | $-0.31_{0.26}$ | $-0.31_{0.26}$ | $-0.30_{0.25}$ | $-0.30_{0.23}$ | $-0.41_{0.32}$ | $-0.31_{0.21}$ |
| $\eta_1$ | $0.51_{0.36}$ | $0.53_{0.38}$ | $0.53_{0.38}$ | $0.51_{0.36}$ | $0.51_{0.36}$ | $0.51_{0.36}$ | $0.52_{0.36}$ | $0.51_{0.36}$ | $0.83_{0.49}$ | $0.51_{0.34}$ |

NOTE: The latent variable $z_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_0 + \eta_1 \mathbf{T} + \gamma_i,$$  
where $\gamma_i \sim N(0, \sigma_\gamma^2)$ is the random intercept accounting for the subject effect. Given $\gamma_i$, the latent variables within the $i$th subject are independent. The random-effects estimator (RE) are obtained by the EM algorithm.

In Tables 1 and 2, we observe that all the estimating equation estimators are consistent as discussed in Section 3, including the estimators using misspecified correlation structures. However, we can improve the estimating efficiency if the correct correlation information is incorporated especially for the joint-imputed model. This is reflected in that the UEE\textsuperscript{Joint} estimators have much smaller standard errors compared with the other estimators in both well-separated and poorly separated cases. Indeed, the UEE\textsuperscript{Joint} performs almost the same as the Oracle estimator. For a well-separated case or if an exchangeable correlation structure is assumed, the performance of the UEE\textsuperscript{Mar} is quite similar to that of the UEE\textsuperscript{Ind} since the longitudinal data generated here are balanced (Liang and Zeger 1986). But when two subgroups are poorly separated and the latent variable has an AR-1 correlation structure, then the UEE\textsuperscript{Mar} has smaller standard errors than UEE\textsuperscript{Ind}.

In addition, if we compare Table 1 and Table 2, we notice that the standard errors of the independent estimators increase significantly from a well-separated case to a poorly separated case, while the proposed UEE\textsuperscript{Joint} is much more stable and performs similarly to the Oracle estimator. Table 3 provides the classification error rates from the model-based clustering. The proposed UEE\textsuperscript{Joint} model has sufficient power in predicting the subgroup membership since we incorporate the information from other time points for the same subject.

In general, it is difficult to know the true correlation structure for latent variables. The unstructured working correlation is always a possible choice because of its flexibility, as it does not assume any pattern for the correlation structure. However, the unstructured correlation leads to additional computational cost, as it has more correlation parameters $T^2(T-1)/2$ to estimate compared with the AR-1 and exchangeable working correlations. In addition, the variation introduced by unstructured correlation leads to less efficient estimations for regression parameters and increases the chance of the convergence problem.
Table 4 indicates that although the standard errors of the estimating equation estimators are relatively small in the well-separated case. In Tables 1 and 2, UEE\textsubscript{Uns\_Mar} and UEE\textsubscript{Uns\_Joint} do not show significant improvement in estimations, but they have more power in prediction compared with the independent and misspecified models in Table 3.

In Tables 1 and 2, the random-effect estimators perform poorly with large bias and standard errors. This is because Sun, Rosen, and Sampson (2007) approach can only incorporate the random intercept which might not be sufficient to handle the within-subject dependence if group membership changes over time. In addition, we also conduct a simulation study where the latent variable $z_i$ is generated from the random-effects model only (Table 4). The random-effects model generates correlations which do not have an obvious pattern. Therefore, the proposed estimating equation assuming a certain pattern of working structure for serial correlations might not be the most efficient in estimation. Nevertheless, Table 4 indicates that although the standard errors of the estimating equation estimators are slightly underestimated, they are still acceptable. In all, regardless of which source the dependence within subjects is induced from, the proposed UEE estimators are robust and efficient in general.

| True | UEE\textsubscript{Ind} | UEE\textsubscript{Ex\_Mar} | UEE\textsubscript{Ex\_Joint} | UEE\textsubscript{AR1\_Mar} | UEE\textsubscript{AR1\_Joint} | UEE\textsubscript{Uns\_Mar} | UEE\textsubscript{Uns\_Joint} | RE | Oracle |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----|--------|
| $\mu_1$ | $-1.21_{0.14}$ | $-1.21_{0.14}$ | $-1.20_{0.05}$ | $-1.21_{0.16}$ | $-1.21_{0.09}$ | $-1.21_{0.16}$ | $-1.21_{0.07}$ | $-1.22_{0.05}$ | $-1.20_{0.05}$ |
| $\sigma_1$ | 1.00_{0.07} | 1.00_{0.07} | 0.99_{0.04} | 1.00_{0.08} | 1.00_{0.05} | 0.99_{0.08} | 1.00_{0.05} | 0.98_{0.04} | 1.00_{0.03} |
| Exch $\mu_2$ | 1.21_{0.14} | 1.21_{0.14} | 1.20_{0.05} | 1.21_{0.16} | 1.20_{0.09} | 1.20_{0.16} | 1.20_{0.07} | 1.22_{0.06} | 1.20_{0.05} |
| $\sigma_2$ | 1.00_{0.08} | 1.00_{0.08} | 1.00_{0.04} | 1.01_{0.08} | 1.01_{0.05} | 1.00_{0.08} | 1.00_{0.05} | 0.98_{0.05} | 1.00_{0.03} |
| $\eta_0$ | -0.33_{0.32} | -0.33_{0.31} | -0.30_{0.19} | -0.32_{0.34} | -0.32_{0.31} | -0.31_{0.35} | -0.31_{0.23} | -0.31_{0.23} | -0.31_{0.18} |
| $\eta_1$ | 0.52_{0.20} | 0.52_{0.20} | 0.50_{0.14} | 0.52_{0.20} | 0.51_{0.19} | 0.49_{0.23} | 0.50_{0.20} | 1.11_{0.50} | 0.50_{0.12} |
| AR1 $\mu_1$ | 1.22_{0.19} | 1.22_{0.15} | 1.21_{0.10} | 1.21_{0.14} | 1.22_{0.06} | 1.22_{0.14} | 1.22_{0.09} | 1.23_{0.05} | 1.20_{0.04} |
| $\sigma_1$ | 1.00_{0.09} | 1.00_{0.07} | 1.00_{0.07} | 1.00_{0.07} | 1.00_{0.04} | 1.00_{0.07} | 1.00_{0.05} | 0.97_{0.06} | 1.00_{0.03} |
| $\eta_0$ | -0.31_{0.39} | -0.32_{0.32} | -0.31_{0.28} | -0.31_{0.33} | -0.31_{0.23} | -0.31_{0.34} | -0.31_{0.25} | -0.51_{0.43} | -0.31_{0.21} |
| $\eta_1$ | 0.51_{0.38} | 0.53_{0.37} | 0.48_{0.39} | 0.51_{0.36} | 0.51_{0.33} | 0.52_{0.36} | 0.53_{0.36} | 0.83_{0.62} | 0.51_{0.34} |

NOTE: The latent variable $z_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

Table 3. Classification error rate based on a two-component univariate normal mixture model

| Separation | Correlation | UEE\textsubscript{Ind} | UEE\textsubscript{Ex\_Mar} | UEE\textsubscript{Ex\_Joint} | UEE\textsubscript{AR1\_Mar} | UEE\textsubscript{AR1\_Joint} | UEE\textsubscript{Uns\_Mar} | UEE\textsubscript{Uns\_Joint} | RE | Oracle |
|------------|-------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----|--------|
| Well       | AR-1        | 0.070 | 0.069 | 0.055 | 0.069 | 0.038 | 0.069 | 0.051 | 0.055 |
|            | Exch        | 0.069 | 0.069 | 0.030 | 0.069 | 0.045 | 0.069 | 0.042 | 0.041 |
| Poorly     | AR-1        | 0.127 | 0.122 | 0.095 | 0.121 | 0.068 | 0.121 | 0.089 | 0.097 |
|            | Exch        | 0.121 | 0.121 | 0.051 | 0.123 | 0.092 | 0.122 | 0.088 | 0.093 |

NOTE: The latent variable $z_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$. 

MIXTURE MODELING FOR LONGITUDINAL DATA

1129
Table 4. The parameter estimators and their empirical standard errors (provided in the subscripts) for a two-component univariate normal mixture model, based on 1000 replicates.

| Separation | UEE_{Ind} | UEE_{Mar} | UEE_{Joint} | UEE_{X} | UEE_{Joint} | UEE_{Uns} | UEE_{Joint} | RE |
|------------|-----------|-----------|-------------|---------|-------------|-----------|-------------|-----|
|            | \mu_1     | \sigma_1  | \mu_2       | \sigma_2 | \eta_0       | \eta_1    | \sigma_2    |     |
| Well       | -1.510.09  | 1.000.06  | 1.510.09    | 1.000.06 | -0.310.18    | 0.510.25  | 0.069       |     |
|            | -1.510.09  | 1.000.06  | 1.510.09    | 1.000.06 | -0.310.18    | 0.510.25  | 0.069       |     |
| Classification errors | 0.069 | 0.068 | 0.067 | 0.069 | 0.069 | 0.069 | 0.067 |     |
| Poorly     | -1.210.13  | 1.000.07  | 1.220.14    | 1.000.06 | -0.330.26    | 0.540.28  | 0.116       |     |
|            | -1.200.14  | 1.000.07  | 1.210.14    | 1.000.06 | -0.330.26    | 0.540.28  | 0.116       |     |
| Classification errors | 0.116 | 0.118 | 0.116 | 0.120 | 0.116 | 0.130 | 0.128 | 0.114 |

NOTE: The latent variable $z_j$ is generated by random-effects model.
4.2 Study 2: Two-Component Mixture of Linear Regression Models

In simulation study 2, we consider a mixture of two linear regression models. In this case, not only the mixing proportion, but also the component densities follow a mean regression model with time covariates.

Similar to the simulation study 1, we generate component indicator variables $Z_i$’s from a logistic model, and for each $Z_i$, we assume there are $T$ repeated measurements over time with $Z_{it} = (Z_{i1}, \ldots, Z_{iT})$. The mixing proportion $E[Z_{it}] = \pi_{it}$ is generated from the logistic model:

$$\log \left( \frac{\pi_{it}}{1 - \pi_{it}} \right) = \eta_0 + \eta_1 \frac{t}{T},$$

with either an AR-1 or exchangeable correlation structure. Conditional on $Z_{it}$, we generate the outcome response $Y_{it}$ from two normal distributions:

$$Y_{it}|(Z_{it} = 1) \sim N(\beta_0^{(1)} + \beta_1^{(1)} \frac{t}{T}, \sigma_1^2),$$
$$Y_{it}|(Z_{it} = 0) \sim N(\beta_0^{(2)} + \beta_1^{(2)} \frac{t}{T}, \sigma_2^2).$$

In this study, we let the sample size $n = 100$ and time points $T = 5$, the mixing proportion parameters in the logistic model are set to be $(\eta_0, \eta_1) = (-0.3, 0.5)$, and the component’s regression parameters are set as $(\beta_0^{(1)}, \beta_1^{(1)}) = (-3, 2)$ and $(\beta_0^{(2)}, \beta_1^{(2)}) = (3, -2)$, the variance parameters are set as $(\sigma_1^2, \sigma_2^2) = (1, 1)$. In contrast to the first simulation setting, the mixture components are regression functions of time covariates, where the component means are changing over time, leading to different separation levels at different times.

Figure 1 illustrates four different separations of two components at different time points in this simulation study. This motivates us to take advantage of serial correlations within subjects to improve accuracy in predicting class memberships and thus to improve the estimations of the regression parameters. In addition, we allow subjects to change group memberships over time (see Figure 2 as an illustration), which makes our approach more flexible compared to growth-curve mixture modeling.

The results in Table 5 show that we gain extra efficiency with smaller standard errors in estimating both the mixing proportion parameters and the component regression parameters for the proposed UEEjoint estimators especially on slope estimators. Table 6 indicates that our approach using the within-subject correlation information provides more predictive power, especially at poorly separated times. In general, when within-subject correlation is large, our approach is able to “borrow” information from well-separated observations to enhance membership prediction for poorly separated observations by incorporating the correlations within each individual subject. Consequently, improvement of the prediction of the subgroup’s membership leads to better estimation of the slope parameter associated with time effect.
Figure 1. Plots of the mixture density of two univariate normal components with mixing proportion (corresponding to the negative-mean component): 0.4, 0.45, 0.55 and 0.6 for (a) to (d), a common variance \( \sigma^2 = 1 \), and the mean parameters are: (a) \((\mu_1, \mu_2) = (-3, 3)\), (b) \((\mu_1, \mu_2) = (-2, 2)\), (c) \((\mu_1, \mu_2) = (-1.5, 1.5)\), (d) \((\mu_1, \mu_2) = (-1.1, 1.1)\).

5. REAL-DATA EXAMPLE: 2008 ELECTION DATA

In this section, we apply the proposed method to the 2007–2008 AP-YAHOO NEWS election panel study (http://www.knowledgenetworks.com/GANP/election2008/index.html). The study was conducted by Knowledge Networks on behalf of the Associated Press and Yahoo! News (APYN) which intends to measure opinion changes starting with the primary elections through the presidential election in November 2008. The data consist of 4965 participants over 11 waves from November 2007 to December 2008, with nine waves before the election, one wave on election day, and the last wave after the election. The primary goal of the study is to investigate the change of participants’ interest in the election over time and important factors associated with their interest. One important factor associated with the interest in the election is interest in election news.

Therefore, we choose one of the survey questions: “Question LV3: How much interest do you have in the following news about the campaign for president, a great deal, quite a bit, only some, very little, or no interest at all.” The five categories of opinions are recorded as an ordinal response variable: 1, 2, 3, 4, and 5, where a smaller value corresponds to a high level of interest in the election news. To measure the opinion change toward the election, we focus on the first nine waves occurring before the election date. There are

Figure 2. Plots of longitudinal responses from a subset of subjects in simulation study 2.
Table 5. The parameter estimators and their empirical standard errors (provided in the subscripts) for a mixture of two regression models, based on 1000 replicates

| True | UEE\textsubscript{Ind} | UEE\textsubscript{Ex} \textsubscript{Mar} | UEE\textsubscript{Ex} \textsubscript{Joint} | UEE\textsubscript{AR1} \textsubscript{Mar} | UEE\textsubscript{AR1} \textsubscript{Joint} | UEE\textsubscript{Uns} \textsubscript{Mar} | UEE\textsubscript{Uns} \textsubscript{Joint} | Oracle |
|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|---------|
| \( \beta^{(1)} \) | \(-3.00_{0.17} \) | \(-3.00_{0.17} \) | \(-3.00_{0.16} \) | \(-2.99_{0.17} \) | \(-3.00_{0.16} \) | \(-3.00_{0.17} \) | \(-3.00_{0.16} \) | \(-3.00_{0.14} \) |
| \( \beta^{(2)} \) | \(2.00_{0.30} \) | \(2.00_{0.30} \) | \( \mathbf{2.00}_{0.24} \) | \(1.95_{0.30} \) | \(1.98_{0.29} \) | \(1.98_{0.30} \) | \(1.97_{0.25} \) | \(2.00_{0.21} \) |
| \( \sigma^{(1)} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.05} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) |
| Exch | \( \beta^{(1)} \) | \(-1.97_{0.29} \) | \(-1.97_{0.29} \) | \( \mathbf{-2.00}_{0.23} \) | \(-1.98_{0.32} \) | \(-1.99_{0.24} \) | \(-1.99_{0.29} \) | \(-1.99_{0.24} \) |
| \( \sigma^{(2)} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.04} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(1.00_{0.06} \) | \(1.00_{0.04} \) |
| \( \eta_0 \) | \(-0.31_{0.22} \) | \(-0.31_{0.22} \) | \( \mathbf{-0.30}_{0.19} \) | \(-0.31_{0.26} \) | \(-0.32_{0.23} \) | \(-0.31_{0.24} \) | \(-0.30_{0.22} \) | \(-0.30_{0.19} \) |
| \( \eta_1 \) | \(0.52_{0.36} \) | \(0.52_{0.36} \) | \( \mathbf{0.49}_{0.19} \) | \(0.47_{0.40} \) | \(0.52_{0.28} \) | \(0.47_{0.36} \) | \(0.46_{0.25} \) | \(0.51_{0.18} \) |
| AR-1 | \( \beta^{(1)} \) | \(1.98_{0.29} \) | \(1.98_{0.32} \) | \(2.04_{0.29} \) | \(1.98_{0.30} \) | \( \mathbf{2.00}_{0.24} \) | \(1.99_{0.34} \) | \(2.00_{0.28} \) | \(2.00_{0.21} \) |
| \( \sigma^{(2)} \) | \(0.99_{0.06} \) | \(0.99_{0.06} \) | \(1.01_{0.05} \) | \(0.99_{0.05} \) | \(0.99_{0.04} \) | \(0.98_{0.06} \) | \(0.98_{0.05} \) | \(1.00_{0.04} \) |
| \( \eta_0 \) | \(-0.31_{0.26} \) | \(-0.31_{0.27} \) | \(-0.33_{0.26} \) | \(-0.30_{0.26} \) | \(-0.31_{0.23} \) | \(-0.28_{0.28} \) | \(-0.29_{0.26} \) | \(-0.31_{0.23} \) |
| \( \eta_1 \) | \(0.51_{0.45} \) | \(0.50_{0.48} \) | \(0.54_{0.44} \) | \(0.51_{0.45} \) | \( \mathbf{0.50}_{0.37} \) | \(0.46_{0.47} \) | \(0.48_{0.40} \) | \(0.51_{0.33} \) |

NOTE: Latent variable \( z_i \) is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter \( \rho = 0.7 \).

1200 participants who have completed the “Question LV3” over the first nine waves. In the following, we have \( n = 1200 \) and the repeated measurement size \( T = 9 \).

In this study, we intend to classify all the participants into two groups, whether they actively follow the election or not, based on their responses to the question “LV3” in the AP-Yahoo survey. Since the survey collects participants’ responses longitudinally, it is not surprising that their interest toward the election could be different at different time points. Consequently, this results in changes of group membership over time. The covariates include time, gender, and race, where gender is 1 for female and 0 for male, and “race” consists of “white,” “black,” and “the other” coded as dummy variables with “white” as the reference. In addition, we also include an interaction term between “time” and the “gender.”

Table 6. Classification error rate at each time point for a mixture of two regression models

| True Corr | model | Time Points | True Corr | model | Time Points |
|-----------|-------|-------------|-----------|-------|-------------|
| UEE\textsubscript{Ind} | 0.005 | 0.014 | 0.037 | 0.086 | 0.165 | UEE\textsubscript{Ind} | 0.004 | 0.014 | 0.036 | 0.084 | 0.164 |
| UEE\textsubscript{Ex} \textsubscript{Mar} | 0.006 | 0.013 | 0.035 | 0.081 | 0.165 | UEE\textsubscript{AR1} \textsubscript{Mar} | 0.005 | 0.017 | 0.034 | 0.076 | 0.168 |
| UEE\textsubscript{Ex} \textsubscript{Joint} | 0.005 | 0.009 | 0.022 | 0.057 | 0.140 | UEE\textsubscript{AR1} \textsubscript{Joint} | 0.003 | 0.011 | 0.028 | 0.060 | 0.098 |
| AR1 UEE\textsubscript{Uns} \textsubscript{Mar} | 0.005 | 0.009 | 0.037 | 0.083 | 0.162 | Exch UEE\textsubscript{Uns} \textsubscript{Mar} | 0.005 | 0.013 | 0.036 | 0.084 | 0.164 |
| UEE\textsubscript{Uns} \textsubscript{Joint} | 0.003 | 0.008 | 0.022 | 0.058 | 0.124 | UEE\textsubscript{Joint} | 0.002 | 0.008 | 0.019 | 0.040 | 0.069 |

NOTE: Latent variable \( z_i \) is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter \( \rho = 0.7 \).
We formulate a two-component mixture model as follows. Let the latent variable \( Z_{it} \) indicate whether the participant \( i \) at the time point \( t \) is interested in election news (\( Z_{it} = 1 \)) or not (\( Z_{it} = 0 \)). We model the mixing proportion using a logistic regression to capture the change in group membership, and the univariate Gaussian for the component distribution. We compare estimators via the proposed unbiased estimating equations using different working correlation structures: independent (UEEInd), exchangeable (UEEExch), AR-1 (UEEAR1) and unstructured (UEEUns), in addition to the random-effects model (RE). The joint imputation is applied in all cases. We focus on modeling the mixing proportion since we are interested in opinion changes over time. The estimates of mixing proportion parameters with corresponding \( p \)-values are summarized in Table 7. The \( p \)-values are calculated based on the asymptotic normal distribution since the sample size \( n = 1200 \) is quite large.

Table 7 indicates that the participants become more and more interested in the election campaign as the time gets closer to election day. Male participants are more interested in election news than females on average; however, females become increasingly more interested in election news than their male counterparts as the election gets closer. In addition, the “black” group is more interested in election news compared to the “white” group, while the “other” group is slightly less interested in election news than the “white” group. In total, about 43.5% of the participants changed their group memberships of showing interest in election news over time.

Furthermore, Table 7 shows that the proposed estimating equation estimators using different working correlations are similar in identifying the effects for “gender” and the interaction of “gender” and “time,” which are highly significant. However, the estimators for the gender and the interaction term are not significant using the random-effects model with much larger \( p \)-values. In addition, the Pew Research Center (2010) reports that the “race” factor played a significant role in the 2008 presidential election, where “black” voters showed more interest in the election than other races in general. For this survey study, the “black” group is only about 7.5% of the participant population (90 out of 1200) which makes it more difficult to detect the “race” factor. Neither the random-effects model nor the independent estimating equations are able to detect a significant difference between the “black” group and the “white” group. However, the proposed method using the unstructured correlation is capable of identifying that the “black” group is significantly different from the “white” group with a \( p \)-value of 0.04. This implies that the proposed method accounting for the serial correlation can improve the estimation efficiency and increase testing power to detect a factor which might not be picked by other approaches.
6. DISCUSSION

In this article, we propose an unbiased estimating equation approach for mixture modeling in longitudinal data, and illustrate how to induce correlation at the level of latent subgroup indicator variables. The proposed method does not require that each subject belong to the same group at different time. To circumvent the complicated form of joint likelihood of the multivariate Bernoulli variables, we propose an unbiased estimating equation approach using the first two moment approximations of the full likelihood, where the serial correlation is incorporated through a working correlation structure. The proposed estimating equations can be regarded as a projection of optimal estimating equations obtained from the complete data via taking the conditional expectation for the missing latent variables.

Our numerical studies show that incorporating correlation information allows one to gain estimation efficiency for the mean regression parameters and the mixing proportion parameters compared to the independent models. The efficiency improvement is significant if the correlation from the longitudinal data is strong and the working structure is correctly specified. In addition, we can improve the classification accuracy for the boundary observations through joint imputation for the missing latent variable.

We can further generalize the current method for more than two subgroups in the population through the extended generalized estimating equation, applying the cumulative logit model for a multinomial latent variable. In addition, we can extend the univariate outcome variable to the multivariate component distribution. If the dimension of the multivariate distribution is high, we can employ the variable selection method by incorporating some penalty terms (Xie, Pan, and Shen 2010). In this article, we mainly focus on modeling serial correlation arising from the latent variable, it would be worthwhile to consider a more complicated setting where within-subject dependence is induced by both latent variable and outcome variable.

SUPPLEMENTARY MATERIAL

Appendix: Appendices (A.1-A.4) include the estimation for dispersion parameter, proofs of Theorem 1, Proposition 1 and Lemma 1. (Appdendix.pdf)

R code: The code for mixture modeling of longitudinal data for a two-component regression model with time covariate. (Rcode.zip)

ACKNOWLEDGMENTS

The authors are very grateful to the Editor, an associate editor and two referees for their insightful comments and suggestions that have improved the article significantly. The research was supported by National Science Foundation Grants (DMS-1308227 and DMS1415308)
REFERENCES

Bahadur, R. R. (1961), “A Representation of the Joint Distribution of Responses to n Dichotomous Items,” in Studies on Item Analysis and Prediction, ed. H. Solomon, Stanford, CA: Stanford University Press, pp. 158–168. [1120]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society, Series B, 39, 1–47. [1120]

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994), Analysis of Longitudinal Data, New York: Oxford University Press. [1121]

Huang, M., Li, R., Wang, H., and Yao, W. (2014), “Estimating Mixture of Gaussian Processes by Kernel Smoothing,” Journal of Business and Economic Statistics, 32, 259–270. [1118]

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), “Adaptive Mixtures of Local Experts,” Neural Computation, 3, 79–87. [1117,1127]

Jiang, W., and Tanner, M. A. (1999), “On the Identifiability of Mixtures-of-Experts,” Neural Networks, 12, 1253–1258. [1122]

Jordan, M. I., and Jacobs, R. A. (1994), “Hierarchical Mixtures of Experts and the EM Algorithm,” Neural Computation, 214, 181–214. [1117,1119]

Liang, K.-Y., and Zeger, S. L. (1986), “Longitudinal Data Analysis Using Generalized Linear Models,” Biometrika, 73, 13–22. [1118,1121,1128]

McCullagh, P. (1983), “Quasi-Likelihood Functions,” Annals of Statistics, 11, 59–67. [1121]

McLachlan, G. J., and Peel, D. (2000), Finite Mixture Models, New York: Wiley. [1127]

Muthén, B., and Muthén, L. K. (2000), “Integrating Person-Centered and Variable-Centered Analyses: Growth Mixture Modeling With Latent Trajectory Classes,” Alcoholism: Clinical and Experimental Research, 24, 882–891. [1118]

Ortega, J. M., and Rheinboldt, W. C. (1970), Iterative Solution of Nonlinear Equations in Several Variables, New York: Academic Press. [1125]

Pew Research Center (2010), “Four Years Later Republicans Faring Better With Men, Whites, Independents and Seniors” (Press Release), available at http://www.people-press.org/files/legacy-pdf/643.pdf. [1134]

Qaqish, B. F. (2003), “A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables With Specified Marginal Means and Correlations,” Biometrika, 90, 455–463. [1118,1126]

Qian, W., and Titterington, D. M. (1992), “Stochastic Relaxations and EM Algorithms for Markov Random Fields,” Journal of Statistical Computation and Simulation, 40, 55–69. [1118,1126]

Qu, A., Lindsay, G. B., and Lu, L. (2010), “Highly Efficient Aggregate Unbiased Estimating Functions Approach for Correlated Data With Missing at Random,” Journal of the American Statistical Association, 105, 194–204. [1118]

Robins, J. M., Rotnitzky, A., and Zhao, L. (1995), “Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data,” Journal of the American Statistical Association, 90, 106–121. [1121]

Rosen, O., Jiang, W., and Tanner, M. A. (2000), “Mixtures of Marginal Models,” Biometrika, 87, 391–404. [1117,1122]

Rotnitzky, A., Robins, J. M., and Zhao, L. (1998), “Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse,” Journal of the American Statistical Association, 93, 1321–1339. [1121]

Rubin, D. B. (1976), “Inference and Missing Data,” Biometrika, 63, 581–592. [1121]

Rubin, D. B., and Wu, Y. (1997), “Modeling Schizophrenic Behavior Using General Mixture Components,” Biometrics, 53, 243–261. [1117]
Seo, B., and Kim, D. (2012), “Root Selection in Normal Mixture Models,” *Computational Statistics & Data Analysis*, 56, 2454–2470. [1125]
Small, G. C., Wang, J., and Yang, Z. (2000), “Eliminating Multiple Root Problems,” *Statistical Science*, 15, 313–332. [1125]
Sun, Z., Rosen, O., and Sampson, A. R. (2007), “Multivariate Bernoulli Mixture Models With Application to Postmortem Tissue Studies in Schizophrenia,” *Biometrics*, 63, 901–909. [1118,1127,1129]
Wedderburn, R. W. M. (1974), “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method,” *Biometrika*, 61, 439–447. [1121]
Xie, B., Pan, W., and Shen, X. (2010), “Penalized Mixture of Factor Analyzers With Application to Variable Selection in Clustering High-Dimensional Data,” *Bioinformatics*, 26, 501–508. [1135]
Yao, W. (2012), “Model Based Labeling for Mixture Models,” *Statistics and Computing*, 22, 337–347. [1127]
Yao, W., and Lindsay, B. G. (2009), “Bayesian Mixture Labeling by Highest Posterior Density,” *Journal of American Statistical Association*, 104, 758–767. [1127]