EMPIRICAL LIKELIHOOD APPROACH FOR LONGITUDINAL DATA WITH MISSING VALUES AND TIME-DEPENDENT COVARIATES∗†

Yan Zhang, Weiping Zhang‡, Xiao Guo
(Dept. of Statistics and Finance, University of Science and Technology of China, Anhui 230026, PR China)

Abstract

Missing data and time-dependent covariates often arise simultaneously in longitudinal studies, and directly applying classical approaches may result in a loss of efficiency and biased estimates. To deal with this problem, we propose weighted corrected estimating equations under the missing at random mechanism, followed by developing a shrinkage empirical likelihood estimation approach for the parameters of interest when time-dependent covariates are present. Such procedure improves efficiency over generalized estimation equations approach with working independent assumption, via combining the independent estimating equations and the extracted additional information from the estimating equations that are excluded by the independence assumption. The contribution from the remaining estimating equations is weighted according to the likelihood of each equation being a consistent estimating equation and the information it carries. We show that the estimators are asymptotically normally distributed and the empirical likelihood ratio statistic and its profile counterpart follow central chi-square distributions asymptotically when evaluated at the true parameter. The practical performance of our approach is demonstrated through numerical simulations and data analysis.

Keywords empirical likelihood; estimating equations; longitudinal data; missing at random

2000 Mathematics Subject Classification 62G05

1 Introduction

Longitudinal data frequently occur in many studies such as medical follow-up studies. A key characteristic of longitudinal data is that outcomes measured repeatedly on the same subject are typically correlated. Regression methods for such

∗This work was supported by the NNSF of China (No.11271347) and the Fundamental Research Funds for the Central Universities.
†Manuscript received September 25, 2015; Revised March 31, 2016
‡Corresponding author. E-mail: zwp@ustc.edu.cn

32:2(2016), 200-220
datasets accounting for within-subject correlation is abundant in the literatures [3,4]. Among which, the generalized estimating equations (GEEs) method by [12] has been widely used since it considers the mean structure and the correlation structure separately. When the marginal mean outcome given the covariates at current time is the same as that on all the past, present and future covariate values, the GEEs method assures consistency of the mean estimates even if the correlation is misspecified, and achieves efficiency if the correlation is correctly specified.

In practice, however, it is common that some covariates may vary over time in a longitudinal study, that is, some of the covariates may be time-dependent. For example, in the Mother’s Stress and Children’s Morbidity Study (MSCM, [1]), the daily ratings of child illness $Y_{it}$ and maternal stress $X_{it}$ are measured during a 28-day follow-up period. Obviously, $Y_{it}$ and $X_{it}$ are time-dependent, both of which vary over time and may correlate with the other measurements. It has been noted that the consistency of GEEs is not assured with arbitrary working correlation structures when there are time-dependent covariates [17]. The reason is that the estimating functions generated by the longitudinal data are no longer unbiased under an arbitrary correlation structure unless the marginal mean outcome given the covariates at the current time is the same as that on all the past, present and future covariate values. Clearly, the use of an independence assumption guarantees the consistency of GEEs but can result in a substantial loss in efficiency due to the fact that, only a subset of all the unbiased estimating functions is used. Some methods have been developed for such situations, for example, [10] classified the time-dependent covariates into three different types based on the moment conditions that are valid to the covariates. They then introduced a generalized method of moments (GMM) to combine all available valid estimating equations optimally. In general, the basic idea is to select estimating functions by minimizing some criteria, see [9,15,24] among others. Recently, [10] proposed a shrinkage empirical likelihood (EL, [16]) approach by including all the estimating functions under the independence correlation assumption and/or those are known to be unbiased a priori, and shrinking all other estimating functions according to the likelihood of each being a biased, uninformative or informative estimating equation. Their approach avoids identifying the uninformative and biased estimating functions and allows different shrinkage parameters for different estimating functions.

Another common problem in the longitudinal data analysis is the missing data problem. For example, there were approximately 4% of dropout in the illness record during the 4 weeks in MSCM study. Statistical methods are available to handle such issue by incorporating missing data mechanisms to provide valid statistical inference, including complete-case analysis method, imputation method, inverse
probability weighted method and likelihood based method [13]. Among which, the complete-case method is a straightforward way to handle missing data and has been widely used. It excludes all units in the dataset with one or more unobserved values. This can dramatically decrease the amount of data to be analyzed since a small percentage of missing values can lead to a considerable proportion of excluded units. Additionally, to be valid, it also requires data to be strictly MCAR (missing completely at random), that is, the probability of missingness is independent of all the covariates and observations at other measurement occasions. This is a rather strong assumption and is seldom met, especially for longitudinal data. Therefore, it often leads to biased or misleading results when directly applying the complete-case method in longitudinal data analysis. The GEEs method also encounters such embarrassment for missing data. [20] proposed a weighting method for rendering GEEs analysis correct under missing-at-random (MAR) mechanisms, that is, the probability of missingness does not depend on the unobserved data given the observed data.

Since it is very common in longitudinal study that some measurements may be missing and some of the covariates are time-dependent simultaneously, and available statistical methods are insufficient to deal with such situation, it is of great interest to jointly study the missing data and time-dependent covariates issues. In this paper, we propose a shrinkage empirical likelihood (EL) approach to estimate the regression parameters under the MAR mechanism and the existence of time-dependent covariates at the same time. The empirical likelihood approaches have been proved to be advantageous in combing multiple sources of information and flexibly formulating models when dealing with missing data, see, for example, [2,19,22,26]. Recently, [14] proposed an EL approach to construct the confidence intervals for the parameters of interest in linear regression models under nonignorable missing mechanism. When some covariates are missing at random, [7] introduced an augmented inverse probability weighted-type empirical likelihood ratio for the parameters of interest for single index model. To deal with missing data, we first multiply impute the missing values by the inverse probability weighting method similar to [22], which was shown to be a key to achieve semiparametric efficiency in their case. Then the estimating equations are partitioned into two groups. The first group contains those unbiased estimating functions and are always used, the second group consists of all other estimating functions which need to be appropriately shrank by the shrinkage parameters. The shrinkage parameters are given according to the likelihood of each estimating equation being unbiased. Finally the EL approach is employed by combining these two groups of estimating functions to gain efficiency.

The rest of this paper is organized as follows. Section 2 describes the models and
the missing mechanism. We describe the details of the shrinkage empirical likelihood estimation approach in Section 3. Numerical studies and real data analysis are presented in Sections 4 and 5 respectively. All technical proofs are relegated to the appendix.

2 Models and Missing Data

Consider a longitudinal study with data set

\[(y_i, X_i, R_i), \quad i = 1, \ldots, n,\]

where \(y_i = (y_{i,1}, \ldots, y_{iT})^T\) denotes the outcome for the \(i\)th subject which is measured at \(T\) time points and may be missing because of skipped visits or dropouts. Let \(X_i = (x_{i1}, \ldots, x_{iT})^T\) be an associated \(p \times T\) matrix of covariates values, where \(x_{it} = (x_{1it}, \ldots, x_{p_{it}})^T\) is a \(p\)-vector of covariates observed at time \(t\). Denote \(R_i = (R_{i1}, \ldots, R_{iT})^T\) the missingness indicator with \(R_{it} = 1\) or \(0\) corresponding to \(y_{it}\) being observed or missing, respectively. We assume that \(X_i\) is always observable. Consider the missing at random mechanism [13], that is,

\[P(R_{it} = 1 | y_i, X_i) = P(R_{it} = 1 | X_i) = P(R_{it} = 1 | x_{it}) = p(x_{it}). \tag{1}\]

The probability \(p(x_{it})\) is called the propensity score or selection probability function.

Let the marginal mean outcome at the \(t\)th time point for the \(i\)th observation be

\[E[y_{it} | x_{it}] \equiv \mu_{it}(\beta) = g(\beta^T x_{it}) \tag{2}\]

where \(g\) is a known monotone link function and \(\beta = (\beta_1, \ldots, \beta_p)^T\) is a \(p\)-vector of unknown parameters. Let \(V_i\) be an estimate of \(\text{var}(y_i)\) or a “working” covariance matrix and \(\bar{v}_{st} = \{y_{it} - \mu_{it}\}\) be the \((s, t)\)th entry of \(V_i^{-1}\). Then the GEEs estimate of \(\beta\) is obtained by solving the following equations

\[
\sum_{i=1}^{n} \frac{\partial \mu_{i}}{\partial \beta_j} V_i^{-1} \{y_i - \mu_i\} = \sum_{i=1}^{n} \sum_{s=1}^{T} \sum_{t=1}^{T} \frac{\partial \mu_{is}}{\partial \beta_j} \bar{v}_{st} \{y_{it} - \mu_{it}\} = 0, \quad j = 1, \ldots, p, \tag{3}\]

where \(\mu_i = (\mu_{i1}, \ldots, \mu_{iT})^T\). Under the assumption of (asymptotic) unbiasedness of the estimating equations, [3] showed that the estimate of \(\beta\) by GEEs is nearly efficient relative to the maximum likelihood estimates of \(\beta\) in many practical situations, provided that \(\text{var}(y_i)\) has been reasonably approximated. However, when time-dependent covariates are present, the unbiasedness assumption might not hold for arbitrary working correlation structure and consequently the GEEs estimate of \(\beta\) is not necessarily consistent [17]. Note that we do not assume that
Following [22], we estimate \( F \) by using the inverse probability weighted estimating equations approach in case of missing data occurring in some estimating equations under working independence, we use the inverse probability weighted estimating equations approach similar to [23] and [22] to get a weighted estimating function with multiple imputations:

\[
\tilde{S}_{it}(y_{it}, x_{it}; \beta_j) = \frac{R_{it}}{p(x_{it})} S_{it}(y_{it}, x_{it}; \beta_j) + \left\{ 1 - \frac{R_{it}}{p(x_{it})} \right\} n_{it}^{-1} \sum_{l=1}^{n_{it}} S_{il}(\tilde{y}_{il}, x_{it}; \beta_j), \quad (4)
\]

where \( S_{it}(y_{it}, x_{it}; \beta_j) = \frac{\partial \mu_{it}}{\partial \beta_j} \{y_{it} - \mu_{it}\} \), \( \tilde{y}_{il} \) is generated from the conditional distribution \( F(y|x_{it}) \) and \( n_{it} \) is the number of imputations. Under the MAR assumption, the above weighted estimating function still satisfies the unbiasedness condition. For other possible biased estimating functions, the augmented inverse probability weighted (AIPW, [5]) estimating equation approach can be used:

\[
\tilde{S}_{ist}(y_{it}, x_{it}, x_{is}; \beta_j) = \frac{R_{it}}{p(x_{it})} S_{ist}(y_{it}, x_{it}, x_{is}; \beta_j) + \left\{ 1 - \frac{R_{it}}{p(x_{it})} \right\} E[S_{ist}(y_{it}, x_{it}, x_{is}; \beta_j)|x_{it}, x_{is}], \quad (5)
\]

where \( S_{ist}(y_{it}, x_{it}, x_{is}; \beta_j) = \frac{\partial \mu_{ist}}{\partial \beta_j} \{y_{it} - \mu_{it}\}, s \neq t \). The unknown conditional expectation \( E[S_{ist}|x_{it}, x_{is}] \) can be naturally estimated by \( n_{it}^{-1} \sum_{l=1}^{n_{it}} S_{ist}(\tilde{y}_{il}, x_{it}, x_{is}; \beta_j) \), here \( \tilde{y}_{il} \) is generated from the conditional cumulative probability function \( F(y|x_{it}, x_{is}) \). Following [22], we estimate \( F(y|x_{it}) \) and \( F(y|x_{it}, x_{is}) \) by

\[
\hat{F}(y|x_{it}) = \frac{\sum_{l=1}^{n} K_h(x_{it} - x_{lt}) R_{lt}I(y_{lt} \leq y)}{\sum_{l=1}^{n} K_h(x_{it} - x_{lt}) R_{lt}}, \quad (6)
\]

\[
\hat{F}(y|x_{it}, x_{is}) = \frac{\sum_{l=1}^{n} K_{ho}(x_{it}^\tau - x_{is}^\tau) R_{lt}I(y_{lt} \leq y)}{\sum_{l=1}^{n} K_{ho}(x_{it}^\tau - x_{is}^\tau) R_{lt}}, \quad (7)
\]
respectively, where \( I(y_{lt} \leq y) = 1 \) if all components of \( y_{lt} \) are less than or equal to the corresponding ones in \( y \) and \( I(y_{lt} \leq y) = 0 \) otherwise. \( K_h(x) \) and \( K_{ho}(x) \) are respectively normalized multivariate kernel functions in \( \mathbb{R}^p \) and \( \mathbb{R}^{2p} \) where \( h \) and \( h^o \) are the smoothing bandwidths. Similarly, the propensity function \( p(x_{it}) \) can also be estimated nonparametrically by

\[
\hat{p}(x_{it}) = \frac{\sum_{l=1}^{n} W_{hw}(x_{lt} - x_{it}) R_{lt}}{\sum_{l=1}^{n} W_{hw}(x_{lt} - x_{it})}, \tag{8}
\]

where \( W_{hw}(x) \) is a normalized multivariate kernel with smoothing bandwidth \( h_w \). One of the commonly used multivariate kernels is the product of univariate kernels with possible different bandwidths. When there are too many covariates, the estimates (6), (7) and (8) can suffer the curse of dimensionality, and dimension reduction is necessary to efficiently estimate these quantities, for example, [6] proposed two alternative approaches for such situation within sufficient dimension reduction framework.

Plugging \( \hat{p}(x) \) for \( p(x) \) in (4) and (5) and generating \( \hat{y}_{lt}^d \) from \( \hat{F}(y|x_{lt}, x_{is}) \) and \( \hat{y}_{lt}^l \) from \( \hat{F}(y|x_{lt}, x_{is}) \), we obtain

\[
\hat{S}_{it}(y_{lt}, x_{lt}; \beta_j) = \frac{R_{lt}}{\hat{p}(x_{lt})} S_{it}(y_{lt}, x_{lt}; \beta_j) + \left( 1 - \frac{R_{lt}}{\hat{p}(x_{lt})} \right) \frac{1}{n_l} \sum_{l=1}^{n_l} S_{it}(\hat{y}_{lt}^d, x_{lt}; \beta_j) \tag{9}
\]

and

\[
\hat{S}_{ist}(y_{lt}, x_{lt}, x_{is}; \beta_j) = \frac{R_{lt}}{\hat{p}(x_{lt})} S_{ist}(y_{lt}, x_{lt}, x_{is}; \beta_j) \\
+ \left( 1 - \frac{R_{lt}}{\hat{p}(x_{lt})} \right) \frac{1}{n_l} \sum_{l=1}^{n_l} S_{ist}(\hat{y}_{lt}^l, x_{lt}, x_{is}; \beta_j), \quad s \neq t. \tag{10}
\]

Since we have a total of \( pT^2 \) estimating functions available, selecting the appropriate estimating functions for parameter estimation becomes a key to produce consistent and more efficient estimates than the class of usual GEEs estimators. To improve efficiency over a GEEs estimate with an independence correlation assumption, we introduce an EL with moment shrinkage approach in next section, as EL allows the easy incorporation of additional information and the number of estimating equations can be greater than the number of parameters.

3 Empirical Likelihood with Moment Shrinkage

Under the independence correlation assumption, a subject-wise empirical likelihood for the regression parameters \( \beta \) is
\[ L_1(\beta_j) = \sup \left\{ \prod_{i=1}^{n} q_i : q_i \geq 0, \sum_{i=1}^{n} q_i = 1, \sum_{i=1}^{n} q_i \sum_{t=1}^{T} \hat{S}_{it}(y_{it}, x_{it}; \beta_j) = 0 \right\}, \]

\[ j = 1, \cdots, p, \tag{11} \]

where \( \hat{S}_{it}(y_{it}, x_{it}; \beta_j) \) is defined in (9).

Similar to \cite{[25]} and \cite{[22]}, we can easily have the following result.

**Theorem 1** Under Conditions 1-5 in the appendix, as \( n \to \infty \) and \( n_1 \to \infty \), \( n^{1/2}(\hat{\beta} - \beta_0) \) converges in distribution to a normal random vector with mean zero and variance \( V = \{T^2E(\partial S/\partial \beta)^T\Gamma_1^{-1}E(\partial S/\partial \beta)\}^{-1} \) evaluated at \( \beta_0 = (\beta_{01}, \cdots, \beta_{0p}) \)

where

\[ \Gamma_1 = TE[p^{-1}(X)\text{cov}\{S(Y, X; \beta_0)|X\} + E\{S(Y, X; \beta_0)|X\}E\{S^T(Y, X; \beta_0)|X\}]. \tag{12} \]

Let \( S(\beta_j) = S(y_{it}, x_{it}; \beta_j) = S_{st}(y_{st}, x_{st}; \beta_j) \) and therefore \( \hat{S}(\beta_j) = \hat{S}(y_{it}, x_{it}; \beta_j) = \hat{S}_{st}(y_{st}, x_{st}; \beta_j) \).

To extract additional information from the estimating equations that are excluded by the independence assumption, and to improve efficiency over a GEEs estimate and subject-wise EL estimate with independence correlation assumption, denote by \( \hat{S}^M(\beta) = \{\hat{S}_{st}(y_{st}, x_{st}; \beta_j), j = 1, \cdots, p, t = 1, \cdots, T\} \) the main estimating functions which are always selected and are known a priori to be unbiased, and other estimating functions are denoted as auxiliary estimating functions, \( \hat{S}^A(\beta) = \{\hat{S}_{st}(y_{st}, x_{st}, x_{st}; \beta_j), t, s = 1, \cdots, T, t \neq s, j = 1, \cdots, p\} \). We introduce a vector \( \gamma \) of shrinkage parameters with \( p(T^2 - T) \) dimensions; each element of \( \gamma \) is a real number in \([0, 1]\) that depends on the data. Let \( \hat{S}^A \gamma(\beta) = \gamma^T \hat{S}^A(\beta) \). Ideally, the shrinkage parameters should down weight those biased or un-informative estimating functions in \( S^A \).

For a particular choice of \( \gamma \), we have estimating functions \( \{\hat{S}^M_i(\beta), \hat{S}^A_i \gamma(\beta)\} \), where \( \hat{S}^M_i(\beta) \) is \( \hat{S}^M(\beta) \) applied to \( (y_{it}, X_i) \) and \( \hat{S}^A_i \gamma(\beta) \) is similarly defined. Thus, consider the maximizer \( \beta^* \) of the following empirical likelihood ratio function

\[ L_2(\beta) = \sup \left\{ \prod_{i=1}^{n} q_i : q_i \geq 0, \sum_{i=1}^{n} q_i = 1, \sum_{i=1}^{n} q_i \{\hat{S}^M_i(\beta), \hat{S}^A_i \gamma(\beta)\} = 0 \right\}. \tag{13} \]

In application, we need to choose appropriate \( \gamma \). Without loss of generality, let the estimating functions in \( S^A \) be \( \{S^A_k = S_{st}(\beta_j) : k = (j - 1)T^2 + (s - 1)T + t, k = 1, \cdots, p(T^2 - T)\} \). Let each element of \( p(T^2 - T) \) vector \( c, c^k \), take 1 if the corresponding estimating function \( S^A_k \) is deemed unbiased by the testing procedure of [10] and be 0 otherwise. Then the shrinkage parameter \( \gamma = \{\gamma^k, k = 1, \cdots, p(T^2 - T)\} \) is defined as
\[
\gamma^k = c^k \exp \left\{-a_0 \sum_{i=1}^{n} S^A_k(\hat{\beta}) \right\}, \quad k = 1, \cdots, p(T^2 - T),
\] (14)

where \(\hat{\beta}\) is a solution of (3) using an independence working correlation and \(a_0\) is a positive constant that depends on the data and satisfies the following conditions: \(\sqrt{n}a_0 \to 0\) and \(na_0 \to \infty\) as \(n \to \infty\), that is, \(a_0 = O(n^{-\delta})\), \(1/2 < \delta < 1\). In practice, \(a_0\) can be determined by cross-validation. Details for choosing \(\gamma\) and solving the EL ratio (13) can be found in [11].

Let \(c_0\) and \(\gamma_0\) be the true parameters of \(c\) and \(\gamma\), respectively. If \(S^A_k\) is truly unbiased, we have \(c_0^k = \gamma_0^k = 1\) and \(c_0^k = \gamma_0^k = 0\) otherwise. Define \(\Lambda = \{k : \hat{c}^k = 1, c_0^k = 1\}\) and let \(\hat{\gamma}\) be the estimator of \(\gamma\) by (14). We can obtain the following result similar to Theorem 1.

**Theorem 2** Under Conditions 1–5 in the appendix, as \(n \to \infty\) and \(n_1 \to \infty\), \(n^{1/2}(\hat{\beta}^* - \beta_0)\) converges in distribution to a normal random vector with mean zero and variance \(V = (\Omega^T \Gamma_{-1}^2 \Omega)^{-1}\) evaluated at \(\beta_0\), where \(\Omega = \left\{E\left(\frac{\partial S^M}{\partial \beta}\right)^T, \gamma_0^0 E\left(\frac{\partial S^A}{\partial \beta}\right)\right\}^T\), \(\Gamma^M = E\left[\frac{1}{p(X)} \text{cov}\{S^M(Y, X; \beta_0)|X\} + E\{S^M(Y, X; \beta_0)|X\} E\{(S^M)^\gamma(Y, X; \beta_0)|X\}\right]\), \(\Gamma^A_k = E\left[\frac{1}{p(X)} \text{cov}\{S^A_k(y_t, x_t, x_s; \beta_{0j})(x_t, x_s)\} + E^2\{S^A_k(y_t, x_t, x_s; \beta_{0j})(x_t, x_s)\}\right]\), and

\[
\Gamma_2 = \begin{pmatrix}
\Gamma^M & 0 \\
0 & \sum_{k \in \Lambda} \Gamma^A_k
\end{pmatrix}.
\]

Because of unknown nuisance parameters in the limiting variances, it is difficult to make statistical inference using the above theorem. For this purpose, we consider the properties of the empirical likelihood ratios induced from \(L_2(\beta)\). Let \(R(\beta_0) = 2l_2(\beta_0) - 2l_2(\hat{\beta}^*)\) where \(l_2(\beta) = -\log\{L_2(\beta)\}\). We have the following theorem.

**Theorem 3** Under Conditions 1–5, as \(n \to \infty\) and \(n_1 \to \infty\), \(R(\beta_0)\) converges in distribution to a chi-square random variable with \(p\) degrees of freedom.

The profile empirical likelihood ratio can be used to test a subvector of parameters when existing nuisance parameters. Let \(\beta = (\beta_1^T, \beta_2^T)^T\), where \(\beta_1\) is \(q < p\) dimensional subvector of \(\beta\). To test \(H_0 : \beta_1 = \beta_{01}\), denote the profile empirical likelihood ratio \(R(\beta_{01}, \hat{\beta}_2^* (\beta_{01})) = 2l_2(\beta_{01}, \hat{\beta}_2^* (\beta_{01})) - 2l_2(\hat{\beta}^*)\), where \(\hat{\beta}_2^* (\beta_{01})\) is the maximizer under the null hypothesis. We have the following theorem for the profile empirical likelihood ratio statistic.

**Theorem 4** Under Conditions 1–5, as \(n \to \infty\) and \(n_1 \to \infty\), \(R(\beta_{01}, \hat{\beta}_2^* (\beta_{01}))\) converges in distribution to a chi-square random variable with \(q\) degrees of freedom.
The central chi-square distributions in Theorems 3 and 4 are convenient for hypotheses testing and constructing confidence regions for \( \beta \) and its components.

4 Simulations

In this section, we present simulation studies to compare the finite sample properties of our proposed method with the other two alternative methods in terms of the bias and the root mean squared error (RMSE). The three methods are:

1. GEEs using an independence working correlation (GEE1), where the missing data \( y_{it} \) is generated from the conditional distribution \( F(y|x_{it}) \);

2. Subject-wise empirical likelihood method in (11) using an independence working correlation (EL1, EL with no shrinkage);

3. Using the empirical likelihood approach in (13), with the vector of shrinkage parameters (EL2, EL with shrinkage).

The data are generated by two different models under four different percentages of missing. We consider sample sizes \( n = 150 \) and 300, and \( T = 6 \) throughout the simulation. The number of replications is 1000 for each situation. The following two models are used by many other authors, see [10,27,11] for example. We are mainly interested in the regression coefficients of \( E[y_{it}|x_{it}] = \zeta_0 + \beta^T x_{it} \) in each case.

**Model 1** A Type II time-dependent covariate. The data generating process is

\[
y_{it} = \zeta_0 + \zeta_1 x_{it} + \zeta_2 x_{i,t-1} + b_i + \epsilon_{it},
\]

\[
x_{it} = \rho x_{i,t-1} + \epsilon_{it}, \quad t = 1, \ldots, 6,
\]

where \( b_i \sim N(0, \sigma_b^2 = 4) \), \( \epsilon_{it} \sim N(0, \sigma^2_{\epsilon} = 1) \) and \( \epsilon_{it} \sim N(0, \sigma^2_{\epsilon} = 1) \) are mutually independent and \( x_{i0} \sim N(0, \sigma^2_{x}/(1-\rho^2)) \) with \( \rho = 0.5, \zeta_0 = 0, \zeta_1 = \zeta_2 = 1 \). For this model, the outcome variable depends on the most recent two values of the covariate and the covariate itself depends on its previous measurement. In this model, the covariate is Type II because \( x_{it} \) is only dependent on \( x_{is}, s < t \). It follows that

\[
E[y_{it}|x_{it}] = \zeta_0 + (\zeta_1 + \zeta_2 \rho) x_{it} := \beta_1 x_{it}.
\]

**Model 2** Three Type III time-dependent covariates. The data generating process is

\[
y_{it} = \zeta_0 + \zeta_1^T x_{it} + \zeta_2^T x_{i,t-1} + b_i + \epsilon_{it},
\]

\[
x_{it} = (x_{it}^1, x_{it}^2, x_{it}^3)^T; \quad x_{it}^j = \rho x_{i,t-1}^j + \kappa b_i + \epsilon_{it}^j, \quad t = 1, \ldots, 6,
\]
where $b_i \sim N(0, \sigma^2_b = 4)$, $\epsilon_{it} \sim N(0, \sigma^2_\epsilon = 1)$, $e_{ij} \sim N(0, \sigma^2_e = 1)$, $j = 1, 2, 3$ are mutually independent and $x_{i0} \sim \text{MV N}(0, \text{diag}[(\sigma^2_e/(1-\rho^2))]_{3 \times 3})$, with $\rho = 0.5$, $\kappa = 0.05$, $\zeta_0 = 0$, $\zeta_1 = \zeta_2 = (1, 1, 1)^T$. Since $x_{it}$ and $y_{it}$ share a common factor $b_i$, there is a feedback effect of past outcome values on future covariate values and hence, the covariates are three Type III covariates. And it follows that

$$E[y_{it}|x_{it}] = \zeta_0 + \sum_{j=1}^3 \left[ \zeta_j^1 + \zeta_j^2 \rho + \frac{(\kappa^2 \sigma^2_b)(1 + \rho)}{\kappa^2 \sigma^2_b + \sigma^2_\epsilon} \right] x_{it}^j := \beta_1 x_{it}^1 + \beta_2 x_{it}^2 + \beta_3 x_{it}^3.$$

We use product Gaussian kernels with possible bandwidth when computing (6), (7) and (8). It is well known that selecting a suitable bandwidth is a critical issue in nonparametric or semiparametric inferences. The classical optimal rate for the bandwidth is $h = O(n^{-1/5})$, see [21]. However, as [27] pointed out, the optimal rate $h = O(n^{-1/5})$ is not allowed here since we require $nh^{2m} \to 0$ for the $m$th order kernel in Condition 5. Hence, we use a simple bandwidth $h = \tilde{\sigma}_X n^{-1/3}$ in model 1 and $h_1 = \tilde{\sigma}_{x_1} n^{-1/3}$, $h_2 = \tilde{\sigma}_{x_2} n^{-1/3}$ and $h_3 = \tilde{\sigma}_{x_3} n^{-1/3}$ in model 2 respectively. Set $n_l = 20$ in (9) and (10).

The simulation results are tabulated in the following two tables. We only give results for the time-dependent slope parameter in each model as the intercept is a time-independent covariate and is seldom of interest. Table 1 corresponds to the situation where there is a single Type II time-dependent covariate, and shows that all estimators are approximately unbiased for $n = 150$ and 300, and our approach outperforms the other two methods especially in the case with higher missing percentage in the response.

Table 1: Bias and RMSE for estimates of $\beta_1 = \zeta_1 + \zeta_2 \rho$ using three methods under Model 1

<table>
<thead>
<tr>
<th>Missing percentage</th>
<th>50% Bias</th>
<th>RMSE</th>
<th>30% Bias</th>
<th>RMSE</th>
<th>17% Bias</th>
<th>RMSE</th>
<th>8% Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>GEE1</td>
<td>-0.1768</td>
<td>0.0895</td>
<td>-0.1193</td>
<td>0.0635</td>
<td>-0.0670</td>
<td>0.0494</td>
<td>-0.0375</td>
</tr>
<tr>
<td></td>
<td>EL1</td>
<td>-0.1299</td>
<td>0.0834</td>
<td>-0.0916</td>
<td>0.0603</td>
<td>-0.0556</td>
<td>0.0479</td>
<td>-0.0340</td>
</tr>
<tr>
<td></td>
<td>EL2</td>
<td>-0.1189</td>
<td>0.0872</td>
<td>-0.0734</td>
<td>0.0697</td>
<td>-0.0290</td>
<td>0.0605</td>
<td>-0.0006</td>
</tr>
<tr>
<td>300</td>
<td>GEE1</td>
<td>-0.1348</td>
<td>0.0645</td>
<td>-0.0954</td>
<td>0.0464</td>
<td>-0.0536</td>
<td>0.0350</td>
<td>-0.0278</td>
</tr>
<tr>
<td></td>
<td>EL1</td>
<td>-0.0945</td>
<td>0.0595</td>
<td>-0.0713</td>
<td>0.0444</td>
<td>-0.0436</td>
<td>0.0336</td>
<td>-0.0245</td>
</tr>
<tr>
<td></td>
<td>EL2</td>
<td>-0.0801</td>
<td>0.0678</td>
<td>-0.0517</td>
<td>0.0552</td>
<td>-0.0150</td>
<td>0.0499</td>
<td>0.0161</td>
</tr>
</tbody>
</table>

Table 2 corresponds to the situation with three Type III time-dependent covariates. Obviously, our approach performs better than the other two methods especially in cases with relatively higher missing percentages. Both simulation results under these two models show that our method works much better than GEE1 and EL1 which simply use an independence working correlation, especially under the situations with relatively higher missing percentage.
Table 2: Bias and RMSE for estimates $\hat{\beta}=(\beta_1, \beta_2, \beta_3)^\top$ using three methods under Model 2

<table>
<thead>
<tr>
<th>Missing percentage</th>
<th>50%</th>
<th>35%</th>
<th>23%</th>
<th>13%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>150</td>
<td>GEE1</td>
<td>-0.3115</td>
<td>0.1068</td>
<td>-0.3165</td>
</tr>
<tr>
<td></td>
<td>Bias</td>
<td>RMSE</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td>GEE1</td>
<td>-0.2476</td>
<td>0.0734</td>
<td>-0.2478</td>
</tr>
<tr>
<td></td>
<td>EL1</td>
<td>-0.3047</td>
<td>0.1031</td>
<td>-0.3074</td>
</tr>
<tr>
<td></td>
<td>EL2</td>
<td>-0.2301</td>
<td>0.0694</td>
<td>-0.2301</td>
</tr>
</tbody>
</table>

Figure 1 shows the plots of the empirical quantiles of empirical likelihood ratios for EL2 evaluated at the true parameters versus the theoretical quantiles of $\chi^2_1$ for model 1 and $\chi^2_3$ for model 2, respectively. Both plots indicate the validity of Theorems 3 and 4.

Figure 1: Quantile-Quantile plot of $R(\beta_{01}, \hat{\beta}_2(\beta_{01}))$ relative to $\chi^2$ distribution with $\beta_{01}$ being the true parameters of the time-dependent covariates, where the missing percentage is about 8% and $n=500$. (a) the empirical quantiles of EL2 relative to $\chi^2_1$ in Model 1; (b) the empirical quantiles of EL2 relative to $\chi^2_3$ in Model 2.
5 Analysis of Morbidity Study

We apply the methods in the previous section to the data from the Mothers’ Stress and Children’s Morbidity Study (MSCM, [1]). This study enrolled \( n = 167 \) preschool children between the ages of 18 months and 5 years that attended an inner-city paediatric clinic. During 4 weeks of follow-up daily measures of maternal stress and child illness were recorded. Additional baseline covariates including child race, maternal educational attainment, marital status, and household size were also available (see [3] for example). Time-dependent measures for household \( i \) included daily ratings of child illness \( Y_{it} \) and maternal stress \( X_{it} \) during a 28-day follow-up period \( t = 1, 2, \ldots, 28 \). Here we only focus on maternal educational attainment among all the other time-independent covariates to examine the effect of mothers’ stress on paediatric care utilization.

In this study, the child illness is missing at random during the 4 weeks of follow-up daily record for each child. We first divide the 28-day follow-up period into 4 weeks. Hence \( Y_{it} \) denotes the number of days that the \( i \)th child was ill in the \( t \)th week, and is missing if some daily records in that week are missing. Data with missing covariates are excluded. There were approximately 4\% of dropout in the illness record during the 4 weeks. Similarly, \( X_{it} \) denotes the number of days the \( i \)th child’s mother felt stressful in the \( t \)th week and there is no missingness in \( X_{it}, t = 1, \ldots, 4 \). Since we are interested in a time-dependent covariate, stress and a time-independent covariate, education, that is, \( \beta = (\beta_1, \beta_2, \beta_3)^T \), where \( \beta_1 \) is the intercept, \( \beta_2 \) and \( \beta_3 \) are the coefficients for stress and education, respectively. We apply the three methods in the previous section to this dataset. The estimates \( \hat{\beta} \) are \((0.7860, 0.2765, -0.0390)^T\), and \((0.7813, 0.3195, -0.0423)^T\) and \((0.7960, 0.3180, -0.0477)^T\) by EL2, EL1 and GEE1 respectively. These estimates are very close as they are all consistent estimates.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEE1</td>
<td>-0.00271</td>
<td>0.02473</td>
</tr>
<tr>
<td>EL1</td>
<td>-0.00270</td>
<td>0.02479</td>
</tr>
<tr>
<td>EL2</td>
<td>0.00214</td>
<td>0.02299</td>
</tr>
</tbody>
</table>

We use the bootstrap method to obtain the bias of the estimator. Table 3 shows the bias and the root mean squared error (RMSE) of estimator for each method based on 500 bootstrap sampling. It can be seen that EL2 outperforms GEE and EL1. For model testing \( H_0 : \beta = 0 \), the \( p \)-value is smaller than 1\% by Theorem 3, that is, the model is significant. Using Theorem 4, we can test \( H_0 : \beta_i = 0 \) for \( i = 1, 2, 3 \), the \( p \)-values are \( 4.51 \times 10^{-5} \), \( 1.96 \times 10^{-6} \) and 0.614 for \( i = 1, 2, 3 \).
respectively. While the $p$-value for testing $H_0: \beta_2 = \beta_3 = 0$ is $9.89 \times 10^{-6}$, we can conclude that the covariate stress is significant while education is nonsignificant.

6 Conclusion

In this paper, we have proposed a shrinkage empirical likelihood estimate for the parameters of interests when the response might be missing at random and the covariates may vary at time. After imputing the missing values, the estimating equations that are known a priori to be unbiased are always selected in the estimation procedure, and the contribution of other estimating functions is controlled by the shrinkage parameters. The shrinkage parameters put the appropriate weights on the additional information from the estimating equations that are excluded by the independence assumption, and therefore we obtain guaranteed improved efficiency over classical generalized estimating equations approach and subject-wise empirical likelihood approach.

Acknowledgements We would like to thank the editor and two referees for all the constructive comments and detailed suggestions, which lead to a substantially improvement of our paper.

References


Appendix

This supplement contains necessary regularity conditions and proofs of the lemmas and Theorems 2-4. The proof of Theorem 1 is similar to that of Theorem 1 in [22]. Therefore we focus on the proofs of Theorems 2-4.

In the following conditions, lemmas and theorems, we are interested in these estimating equations:

\[
\hat{S}_{it}(y_{it}, x_{it}; \beta_j) = \frac{R_{it}}{\hat{p}(x_{it})} S_{it}(y_{it}, x_{it}; \beta_j) + \left\{ 1 - \frac{R_{it}}{\hat{p}(x_{it})} \right\} \frac{1}{n_t} \sum_{l=1}^{n_t} S_{il}(y_{il}, x_{il}; \beta_j),
\]

\[
\hat{S}^A_{ik} = \hat{S}_{ist}(y_{it}, x_{it}, x_{is}; \beta_j) = \frac{R_{it}}{\hat{p}(x_{it})} S_{ist}(y_{it}, x_{it}, x_{is}; \beta_j) + \left\{ 1 - \frac{R_{it}}{\hat{p}(x_{it})} \right\} \frac{1}{n_t} \sum_{l=1}^{n_t} S_{isl}(y_{il}, x_{il}, x_{is}; \beta_j), \quad s \neq t,
\]

where \( k = (j - 1)T^2 + (s - 1)T + t \). Denote \( \hat{S}(\beta) = \hat{S}(y_{it}, x_{it}; \beta) = \hat{S}^i_t(y_{it}, x_{it}; \beta), \) \( S^M(\beta) = \{ \hat{S}(y_{it}, x_{it}; \beta), j = 1, \ldots, p, t = 1, \ldots, T \} \) and \( S^A(\beta) = \{ \hat{S}_{ist}(y_{it}, x_{it}, x_{is}; \beta), j = 1, \ldots, p, t, s = 1, \ldots, T, t \neq s \} \), respectively. Then define

\[
\Gamma_1 = TE[p^{-1}(X)\text{cov}\{S(Y, X; \beta_0)|X\} + E\{S(Y, X; \beta_0)|X\} E\{S^T(Y, X; \beta_0)|X\}]_{A.1}
\]

and

\[
\Gamma_2 = \begin{pmatrix}
\Gamma^M & 0 \\
0 & \sum_{k \in \Lambda} \Gamma^A_k
\end{pmatrix}, \quad (A.2)
\]

respectively, where

\[
\Gamma^M = E\left[ \frac{1}{\hat{p}(X)} \text{cov}\{S^M(Y, X; \beta_0)|X\} + \text{cov}\{S^M(Y, X; \beta_0)|X\} \text{cov}\{S^M(Y, X; \beta_0)|X\}^T \right], \quad (A.3)
\]

\[
\Gamma^A_k = E\left[ \frac{1}{\hat{p}(X)} \text{cov}\{S^A_k(y_{it}, x_{it}, x_{is}; \beta_0)|(x_{it}, x_{is})\} + \text{cov}\{S^A_k(y_{it}, x_{it}, x_{is}; \beta_0)|(x_{it}, x_{is})\} \right]. \quad (A.4)
\]

**Conditions**

Before stating the lemmas, we impose the following regularity conditions. For simplicity, we denote all the estimating functions as a generic term \( g(Y, X; \beta) \), which includes \( S(Y, X; \beta), S^M(Y, X; \beta) \) and \( S^A(Y, X; \beta) \). Let \( f(x) \) be the probability
density function of $X$ and $m_y(x) = \mathbb{E}[g(Y, X; \beta_0)|X = x]$, which therefore includes $m_S(x) = \mathbb{E}[S(Y, X; \beta_0)|X = x]$, $m_{SM}(x) = \mathbb{E}[S^M(Y, X; \beta_0)|X = x]$ and $m_{SA}(x) = \mathbb{E}[S^A_k(y_t, x_t, x_s; \beta_0)](x_t, x_s) = x]$. 

**Condition 1** The functions $p(x)$, $f(x)$, $m_y(x)$ have bounded partial derivatives with respect to $x$ up to an order $q$ with $q \geq 2$, $2q > p$ and $\inf_x p(x) \geq c_0$ for some $c_0 > 0$.

**Condition 2** The true parameter value $\beta_0$ is the unique solution to $\mathbb{E}[g(Y, X; \beta_0)] = 0$, and all the estimating functions denoted by $g(y, x; \beta_0)$ have bounded $q$th order partial derivatives with respect to $x$, $\mathbb{E}[g(y, x; \beta_0)]^4 < \infty$.

**Condition 3** Both $\partial g(y, x; \beta)/\partial \beta$ and $\partial^2 g(y, x; \beta)/\partial \beta \partial \beta^r$ are continuous in a neighbourhood of the true value $\beta_0$. The functions $\|\partial g(y, x; \beta)/\partial \beta\|$, $\|\partial^2 g(y, x; \beta)/\partial \beta \partial \beta^r\|$ and $\|g(y, x; \beta)\|^3$ are uniformly bounded by some integrable function in the neighborhood of $\beta_0$.

**Condition 4** The matrices $\Gamma_1$ and $\Gamma_2$ are positive definite and $\mathbb{E}[\partial g(y, x; \beta)/\partial \beta]$ has full column rank.

**Condition 5** The functions $W$, $K$ and $K^o$ are $m$th order kernels when $p > 4$. As $n \to \infty$, $nh^p/\log n \to \infty$, $nh^p/\log n \to \infty$, $n(h^o)^{2p}/\log n \to \infty$, $nh^{2m} \to 0$, $nh_{w^m} \to 0$ and $n(h^o)^{2m} \to 0$, where $h, h^o, h_w$ are bandwidths for (6), (7) and (8) respectively.

**Lemmas**

For simplicity, we use $C$ to denote a generic positive constant independent of $n$ in the proofs. We present the proofs of Theorems 2-4 as consequences of the following lemmas.

To get the estimate $\gamma$, following the approach in [11], we first get the $k$th element of $c$, $c^k$, which takes a value of 1 if $S^A_k$ is deemed unbiased by the testing procedure of [10] and 0 otherwise. Then we obtain a proper $a_0$ by cross-validation depending on the data. Let

$$\hat{\gamma}^k = \hat{c}^k \exp \left\{ -a_0 \sum_{i=1}^n S^A_{ik}(\hat{\gamma}) \right\}, \quad k = 1, \cdots, p(T^2 - T),$$

hence, we have the estimator of $\gamma$. Let $\gamma_0$ and $\gamma_0$ be the true parameters of $c$ and $\gamma$, respectively. If $S^A_k$ is truly unbiased, we have $c^k_0 = \gamma^k_0 = 1$ and $c^k_0 = \gamma^k_0 = 0$ otherwise. Let $\Lambda = \{k : c^k = 1, c^k_0 = 1, k = 1, \cdots, p(T^2 - T)\}$.

**Lemma 1** Suppose Conditions 1-5 are satisfied, then as $n \to \infty$ and $n_t \to \infty$, we have for $k \in \{k : c^k = 1\}$,

$$|\hat{\gamma}^k - \gamma^k_0| = O_p(1). \quad (A.5)$$
Proof We partition \( \{ k : \tilde{c}^k = 1, k = 1, \cdots, p(T^2 - T) \} \) into two subsets: 
\( \Lambda \) and \( \tilde{\Lambda} \). For \( k \in \Lambda \), by applying Taylor’s expansion,

\[
|\tilde{\gamma}^k - \gamma_0^k| = \left| \exp \left\{ - a_0 \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right\} - 1 \right| = a_0 \left| \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right| + o_p(1). \tag{A.6}
\]

Following Lemma 2 in [22], we have

\[
n^{-1/2} \sum_{i=1}^{n} \tilde{S}_{ik}^A(\beta_0) \rightarrow N(0, \Gamma_k^A) \tag{A.7}
\]

in distribution, where \( \Gamma_k^A \) is defined in (A.4). Noting that \( \|\tilde{\beta} - \beta_0\| = o(n^{-1/2}) \), by using Chebyshev’s inequality, we obtain

\[
\mathbb{P} \left( \left| \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right| \geq \varepsilon \right) \leq \frac{\text{var} \left\{ \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right\}}{\varepsilon^2}.
\]

Therefore (A.5) is proved by noting that \( a_0 = O(n^{-\delta}), 1/2 < \delta < 1 \).

For \( k \in \tilde{\Lambda} \),

\[
|\tilde{\gamma}^k - \gamma_0^k| = |c^k - c_0^k| \exp \left\{ - a_0 \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right\} = \exp \left\{ - na_0 \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right\}. \tag{A.8}
\]

Noting that \( \tilde{\beta} \) is a \( \sqrt{n} \)-consistent estimator for \( \beta_0 \), \( n^{-1} \left| \sum_{i=1}^{n} \tilde{S}_{ik}(\tilde{\beta}) \right| = O_p(1) \), and the fact that \( a_0 = O(n^{-\delta}), 1/2 < \delta < 1, na_0 \rightarrow \infty \), (A.5) follows for \( k \in \tilde{\Lambda} \). This completes the proof of Lemma 1.

**Lemma 2** Under Conditions 1-5, as \( n \rightarrow \infty \) and \( n_1 \rightarrow \infty \),

\[
n^{-1/2} \sum_{i=1}^{n} \tilde{\gamma}^T \tilde{S}_{i}^A(\beta_0) \rightarrow N \left( 0, \sum_{k \in \Lambda} \Gamma_k^A \right) \tag{A.9}
\]

in distribution and

\[
\frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{\gamma}^T \tilde{S}_{i}^A(\beta_0) \right\}^2 = \sum_{k \in \Lambda} \Gamma_k^A + o_p(1), \tag{A.10}
\]

where \( \Gamma_k^A \) is defined in (A.4).
Proof Since
\[ n^{-1/2} \sum_{i=1}^{n} \hat{\gamma}^T \hat{S}_i^A(\beta_0) = n^{-1/2} \sum_{i=1}^{n} \sum_{k \in \Lambda} \hat{\gamma}^k \hat{S}_{ik}^A(\beta_0) + n^{-1/2} \sum_{i=1}^{n} \sum_{k \in \Lambda} (\hat{\gamma}^k - 1) \hat{S}_{ik}^A(\beta_0) \]
\[ + n^{-1/2} \sum_{i=1}^{n} \sum_{k \in \Lambda} \hat{\gamma}^k \hat{S}_{ik}^A(\beta_0) \]
\[ = I_1 + I_2 + I_3. \]

For \( k \in \Lambda \), noting \( \max_k \hat{S}_{ik}^A(\beta_0) = O_p(1) \) and using Lemma 1 again, we have \( I_2 = o_p(1) \). Notice that for \( k \in \tilde{\Lambda} \), \( \gamma^k = o_p(1) \) and \( n^{-1/2} \sum_{i=1}^{n} \hat{S}_{ik}^A(\beta_0) = O_p(1) \), we have \( I_3 = o_p(1) \). Hence, (A.9) follows from (A.7).

Now we prove (A.10),
\[ n^{-1} \sum_{i=1}^{n} [\hat{\gamma}^T \hat{S}_i^A(\beta_0)]^2 = n^{-1} \sum_{i=1}^{n} \left[ \sum_{k \in \Lambda} \hat{S}_k^A(\beta_0) \right]^2 + o_p(1) \]
\[ = n^{-1} \sum_{i=1}^{n} \left[ \sum_{k \in \Lambda} (\hat{S}_{ik}^A)^2(\beta_0) + 2 \sum_{k,l \in \Lambda} \hat{S}_{ik}^A(\beta_0) \hat{S}_{il}^A(\beta_0) \right] + o_p(1). \]
(A.11)

When \( k, l \in \Lambda \), \( c_0^k = c_0^l = 1 \). It means that the estimating functions \( S_{ik}^A \) and \( S_{il}^A \) are truly unbiased, thus \( \text{cov}[S_{ik}^A(\beta_0), S_{il}^A(\beta_0)] = 0 \). By the fact that \( \max_k \hat{S}_{ik}^A(\beta_0) = O_p(1) \), we have
\[ n^{-1} \sum_{i=1}^{n} \sum_{k,l \in \Lambda} \hat{S}_{ik}^A(\beta_0) \hat{S}_{il}^A(\beta_0) = o_p(1). \]
(A.12)

Following Lemma 2 in [22], we have
\[ n^{-1} \sum_{i=1}^{n} \{ \hat{S}_{ik}^A(\beta_0) \}^2 = \Gamma_k^A + o(1), \quad k \in \Lambda. \]
(A.13)

(A.10) can be concluded from (A.11) to (A.13). This completes the proof of Lemma 2.

Lemma 3 Under Conditions 1-5, as \( n \to \infty \) and \( n_l \to \infty \),
\[ n^{-1/2} \sum_{i=1}^{n} \{ \hat{S}_i^M(\beta), \hat{\gamma}^T \hat{S}_i^A(\beta_0) \} \to N(0, \Gamma_2). \]
(A.14)
in distribution and

\[
\frac{1}{n} \sum_{i=1}^{n} \{ \hat{S}_i^M(\beta), \gamma^\top \hat{S}_i^A(\beta_0) \} \{ \hat{S}_i^M(\beta), \gamma^\top \hat{S}_i^A(\beta_0) \}^\top = \Gamma_2 + o_p(1), \tag{A.15}
\]

where \( \Gamma_2 \) is given in (A.2).

**Proof** Following the approach of Lemma 1 in [22], we can easily show that

\[
n^{-1/2} \sum_{i=1}^{n} \hat{S}_i^M(\beta_0) \to N(0, \Gamma^M) \tag{A.16}
\]

in distribution and

\[
\frac{1}{n} \sum_{i=1}^{n} \hat{S}_i^M(\beta_0) \{ \hat{S}_i^M(\beta_0) \}^\top = \Gamma^M + o(1), \tag{A.17}
\]

where \( \Gamma^M \) is defined in (A.3). Note that

\[
\text{cov}(S_{it}(\beta_0), S_{it}^A(\beta_0)) = 0
\]

for all the equations \( S_{it} \in S_i^M \) and \( S_{ik}^A \in S_i^A \), and

\[
n^{-1/2} \sum_{i=1}^{n} \gamma^\top \hat{S}_i^A(\beta_0) \hat{S}_i^M(\beta_0) = n^{-1/2} \sum_{i=1}^{n} \sum_{k \in \Lambda} \hat{S}_{ik}^A(\beta_0) \hat{S}_i^M(\beta_0) + o_p(1) = o_p(1). \tag{A.18}
\]

The lemma can then be proved by directly following the approach of Lemma 1 in [22].

Given a particular choice of \( \gamma \), we denote a profile log-EL of

\[
L_2(\beta) = \sup \left\{ \prod_{i=1}^{n} q_i : q_i \geq 0, \sum_{i=1}^{n} q_i = 1, \sum_{i=1}^{n} q_i \{ \hat{S}_i^M(\beta), \hat{S}_i^A(\gamma) \} = 0 \right\} \tag{A.19}
\]

as

\[
\ell(\beta, \lambda) = n^{-1} \sum_{i=1}^{n} \log \left\{ 1 + \lambda^\top \{ \hat{S}_i^M(\beta), \hat{S}_i^A(\gamma) \} \right\}. \tag{A.20}
\]

Differentiating (A.20) with respect to \((\beta, \lambda)\), we obtain the empirical likelihood equation

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \lambda^\top \{ \hat{S}_i^M(\beta), \hat{S}_i^A(\gamma) \}} \left\{ \frac{\partial \lambda^\top \{ \hat{S}_i^M(\beta), \hat{S}_i^A(\gamma) \}}{\partial \beta} \right\}^\top \lambda(\beta) = 0, \tag{A.21}
\]

which leads to the maximum EL estimates \((\hat{\beta}^\gamma, \hat{\lambda}^\gamma)\) of \( L_2(\beta) \) in (A.19).
Lemma 4 Suppose Conditions 1-5 are satisfied, then as \( n \to \infty \) and \( n_l \to \infty \), with the proper choice of \( \hat{\gamma} \) depending on the data by cross-validation, which therefore gives less weights to those estimating functions that are likely to be biased, and with probability tending to 1, the likelihood equation (A.20) has a solution \( \hat{\beta} \hat{\gamma} \) within the open ball \( \| \beta - \beta_0 \| < C n^{-1/3} \), and \( L_2(\beta) \) attains its local maximum at \( \hat{\beta} \hat{\gamma} \).

Proof As a consequence of Lemma 3 and the proof of Lemma 1 in [18], we can establish this lemma for the existence of a local maximizer of \( L_2(\beta) \).

Based on the above conditions and lemmas, we are dedicated to prove Theorems 2-4.

Proof of Theorem 2

Proof The results can follow directly from the proof of Theorem 1 in [22]. First, we differentiate \( l_\gamma(\beta, \lambda) \) with respect to \((\beta, \lambda)\) in (A.20), let

\[
Q_{n1}(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{S}_M(\beta), \tilde{S}_A^\Lambda(\beta) \right\},
\]

\[
Q_{n2}(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \lambda^T \{ \tilde{S}_M(\beta), \tilde{S}_A^\Lambda(\beta) \} \left\{ \frac{\partial \{ \tilde{S}_i^M(\beta), \tilde{S}_i^A(\beta) \} }{\partial \beta} \right\}^T} \lambda.
\]

With the proper \( \hat{\gamma} \), from Lemma 4, we can see that \( \hat{\beta} \hat{\gamma} \) and \( \hat{\lambda} \hat{\gamma} = \lambda(\hat{\beta} \hat{\gamma}) \) satisfy

\[
Q_{n1}(\hat{\beta} \hat{\gamma}, \hat{\lambda} \hat{\gamma}) = 0, \quad Q_{n2}(\hat{\beta} \hat{\gamma}, \hat{\lambda} \hat{\gamma}) = 0. \tag{A.22}
\]

Inverting a Taylor’s expansion on (A.22) near \( (\beta_0, 0)^T \), we show that

\[
n^{1/2}(\hat{\beta} \hat{\gamma} - \beta_0) = V \Omega^T \Gamma_2^{-1} n^{1/2} Q_{n1}(\beta_0, 0) + o_p(1), \tag{A.23}
\]

where \( V = \{ \Omega^T \Gamma_2^{-1} \Omega \}^{-1} \) evaluated at \( \beta_0 \), \( \Omega = \left\{ \mathbf{E} \left[ \frac{\partial S_M}{\partial \beta} \right]^T, \gamma_0 \mathbf{E} \left[ \frac{\partial S_A}{\partial \beta} \right] \right\}^T \) and \( \Gamma_2 \) is defined in (A.2). By applying Lemma 3, we have

\[
n^{1/2} Q_{n1}(\beta_0, 0) \to N(0, \Gamma_2) \tag{A.24}
\]

in distribution. This completes the proof of Theorem 2.

Proofs of Theorems 3 and 4

Proof By expanding the logarithm function in (A.20) at 1 for \( \beta \) and \( \beta_0 \) respectively, we establish the following expansion of the empirical likelihood ratio:

\[
R(\beta_0) = n^{1/2} Q_{n1}(\beta_0, 0) \Gamma_2^{-1} \Omega \Omega^T \Gamma_2^{-1} n^{1/2} Q_{n1}(\beta_0, 0) + o_p(1), \tag{A.25}
\]
and therefore
\[
2l_2(\hat{\beta}) = 2 \sum_{i=1}^{n} \log \{1 + (\hat{\lambda}_i \hat{\gamma})^\tau \{ \hat{S}_i^M(\hat{\beta}), \hat{S}_i^A(\hat{\beta}) \} \}
\]
\[
= -nQ_{n1}(\beta_0, 0)\Gamma_2^{-1}[I + \Omega \Omega^\tau \Gamma_2^{-1}]Q_{n1}(\beta_0, 0) + o_p(1). \tag{A.26}
\]

Theorem 3 follows from (A.25) by applying Lemma 3. Noting that \( \hat{\beta}_2(\beta_{01}) \) and \( \hat{\lambda}_1^\gamma = \hat{\lambda}_1 \{ \hat{\beta}_2(\beta_{01}) \} \) satisfy
\[
Q_{n3}\{\beta_{01}, \beta_2, \hat{\lambda}_1^\gamma\} = 0, \quad Q_{n4}\{\beta_{01}, \beta_2, \hat{\lambda}_1^\gamma\} = 0,
\]
where
\[
Q_{n3}(\beta_{01}, \beta_2, \lambda_1) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\hat{S}_i^M(\beta_{01}, \beta_2), \hat{S}_i^A(\beta_{01}, \beta_2)}{1 + \lambda_1^T \{ \hat{S}_i^M(\beta_{01}, \beta_2), \hat{S}_i^A(\beta_{01}, \beta_2) \}} \right\} \lambda_1
\]
\[
Q_{n4}(\beta_{01}, \beta_2, \lambda_1) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \{ \hat{S}_i^M(\beta_{01}, \beta_2), \hat{S}_i^A(\beta_{01}, \beta_2) \}/\partial \beta_2}{1 + \lambda_1^T \{ \hat{S}_i^M(\beta_{01}, \beta_2), \hat{S}_i^A(\beta_{01}, \beta_2) \}} \right\} \lambda_1.
\]

Similar to the proof of (A.26), we have
\[
2l_2(\beta_{01}, \beta_2(\beta_{01})) = -nQ_{n1}(\beta_0, 0)\Gamma_2^{-1}[I + \tilde{\Omega} \tilde{\Omega}^\tau \Gamma_2^{-1}]Q_{n1}(\beta_0, 0) + o_p(1),
\]
where \( \tilde{\Omega} = \left\{ \text{E} \left[ \frac{\partial S^M}{\partial \beta_2} \right] \right\}^\tau, \gamma_0 \text{E} \left[ \frac{\partial S^A}{\partial \beta_2} \right] \} \}^\tau \) and \( \tilde{V} = \{ \tilde{\Omega} \Gamma_2^{-1} \tilde{\Omega} \}^{-1} \) evaluated at \( \beta_0 \). Thus by Lemma 3 we can show that \( R(\beta_{01}, \beta_2(\beta_{01})) = 2l_2(\beta_{01}, \beta_2(\beta_{01})) - 2l_2(\beta_\gamma) \rightarrow \chi^2_2 \) in distribution and Theorem 4 follows.

(\text{edited by Mengxin He})