Statistics and Its Interface Volume 6 (2013) 303–313

Estimation in longitudinal studies with nonignorable dropout

Jun Shao*,† and Jiwei Zhao

A sampled subject with repeated measurements often drops out prior to the study end. Data observed from such a subject is longitudinal with monotone missing. If dropout at a time point t is only related to past observed data from the response variable, then it is ignorable and statistical methods are well developed. When dropout is related to the possibly missing response at t even after conditioning on all past observed data, it is nonignorable and statistical analysis is difficult. Without any further assumption, unknown parameters may not be identifiable when dropout is nonignorable. We develop a semiparametric pseudo likelihood method that produces consistent and asymptotically normal estimators under nonignorable dropout with the assumption that there exists a dropout instrument, a covariate related to the response variable but not related to the dropout conditioned on the response and other covariates. Although consistency and asymptotic normality for the proposed estimators can be established using a standard argument, their asymptotic covariance matrices are very complicated because the estimation at t uses estimators from all time prior to t. Our main effort is to derive easy-to-compute consistent estimators of the asymptotic covariance matrices for assessing variability or inference. For illustration, we present an example using the HIV-CD4 data and some simulation results.

AMS 2000 subject classifications: Primary 62H12; secondary 62G20.
Keywords and phrases: Asymptotic covariance matrix, Dropout instrument, Pseudo likelihood, Repeated measurements, Semiparametric model.

1. INTRODUCTION

Longitudinal data or repeated measurements are often encountered in medical, health, economical, and social studies. For a sampled subject, let $Y = (Y_1, \ldots, Y_T)$ be a T-dimensional longitudinal response vector and X be a cross-sectional or longitudinal covariate vector associated with Y. We focus on the estimation or inference on some unknown parameters in $p(Y|X)$ or $p(Y)$, where $p(\cdot|\cdot)$ is a generic notation for the conditional density and $p(\cdot)$ is for the unconditional density. In many studies, values of X are completely observed but a subject may dropout at $t \le T$ so that $(Y_1, \ldots, Y_{t-1})$ is observed and $(Y_t, \ldots, Y_T)$ is missing. This is referred to as monotone missing, but we use the term dropout for monotone missing throughout this paper. Let $R = (R_1, \ldots, R_T)$, where $R_t = 1$ if Y_t is observed and $R_t = 0$ if Y_t is missing, $t = 1, \ldots, T$. If the subject drops out at time t, then $R_1 = \cdots = R_{t-1} = 1$ and $R_t = \cdots = R_T = 0$. Also, if $R_{t-1} = 0$, then $R_t = 0$ with certainty, $t = 1, \ldots, T$.

For longitudinal data, it is natural that dropout at time t is not related to future values $Y_{t+1}, \ldots, Y_T$ (e.g., Diggle and Kenward, 1994) and, thus,

$$P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_t, X, R_{t-1} = 1), \quad t = 1, \ldots, T. \tag{0}$$

If dropout is not related to the current value Y_t, i.e.,

$$P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_{t-1}, X, R_{t-1} = 1), \quad t = 1, \ldots, T,$$

then the dropout is ignorable (e.g., Little, 1995; Little and Rubin, 2002), which is a much stronger assumption than (0) because the dropout propensity only depends on $(Y_1, \ldots, Y_{t-1}, X)$ that is observed. Estimation methods under ignorable dropout are well developed (e.g., Little and Rubin, 2002; Paik, 1997). However, in many longitudinal studies dropout depends on not only $(Y_1, \ldots, Y_{t-1}, X)$ but also Y_t that may be missing and, hence, is nonignorable. Nonignorable dropout presents a great challenge in the estimation of unknown parameters in $p(Y|X)$ or $p(Y)$ (see, e.g., Robins, Rotnitzky and Zhao, 1995; Troxel, Harrington and Lipsitz, 1998; Troxel, Lipsitz and Harrington, 1998).

The purpose of our study is to develop an estimation method under nonignorable dropout. Without any further assumption, however, some unknown parameters in $p(Y|X)$ or $p(Y)$ are not identifiable. To identify the unknown parameters, we need to assume that some component of $V_t = (Y_1, \ldots, Y_t, X)$ is not related to dropout, conditioned on the other components. Note that the ignorable dropout assumption assumes that Y_t is not related to dropout, conditional on the rest of the components of V_t. For nonignorable dropout, we assume that $X = (U, Z)$ and the component Z is not related to dropout conditional on other components of V_t, i.e.,

$$P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_t, U, R_{t-1} = 1), \quad t = 1, \ldots, T. \tag{1}$$

*Corresponding author.
†Partially supported by the NSF Grant DMS-1007454.

The difference between (0) and (1) is that the covariate Z is not present on the right-hand side of (1), which makes it possible for us to identify and estimate unknown parameters, provided that Y and Z are dependent conditioned on U, i.e., Z is a useful covariate. Such a covariate Z is referred to as an instrument for dropout. Furthermore, we need to assume that at least one of $p(Y|X)$ and $P(R_t = 1|Y_1, \ldots, Y_t, U, R_{t-1} = 1)$ is parametric. Otherwise some unknown parameters are not identifiable (Robins and Ritov, 1997). In this paper, we follow Tang, Little, and Raghunathan (2003) and assume a parametric model on $p(Y|X)$:

$$p(Y|X) = \prod_{t=1}^{T} f_t(Y_t|V_{t-1}, \theta_t), \tag{2}$$

where $f_t(Y_t|V_{t-1}, \theta_t)$ is the probability density of Y_t given $V_{t-1} = (Y_1, \ldots, Y_{t-1}, X), t > 1$, or $V_0 = X$, f_t's are known functions, and θ_t's are distinct unknown parameter vectors.

The approach in Tang et al. (2003) is for a general multivariate Y under the assumption $p(R|Y, X) = p(R|Y)$ that allows us to estimate $p(X|Y)$ using the observed (X, Y), as well as the parameters in $p(Y|X)$ through the Bayes formula $p(X|Y) = p(Y|X)p(X)/\int p(Y|x)p(x)dx$. Tang et al. (2003) proposed both parametric and nonparametric methods to estimate $p(X)$. However, this approach has the following two problems. First, it discards observed but incomplete data from dropped out subjects. Second, the dimension of X is required to be as large as the dimension of Y, which limits the application scope. For longitudinal Y, Tang et al. (2003) actually improved their approach regarding the previously discussed problems, but under the following assumption much stronger than (1):

$$P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_t, R_{t-1} = 1), \quad t = 1, \ldots, T,$$

that is, conditioned on Y_t, the dropout propensity depends on neither past responses $Y_1, \ldots, Y_{t-1}$ nor the entire covariate vector X.

Assuming (1) with no model on $P(R_t = 1|Y_1, \ldots, Y_t, U, R_{t-1} = 1)$ and assuming (2), we derive a semiparametric pseudo likelihood for estimating parameters in $p(Y|X)$ or $p(Y)$. We are able to utilize all observed data. Since our method is based on pseudo likelihoods constructed sequentially as $t = 1, \ldots, T$, we do not require a high-dimensional covariate X to identify parameters. Also, at each step the maximization in our method is carried out with a low dimensional vector of parameters and, hence, the computation is sensible.

The methodology is developed in Section 2. Consistency and asymptotic normality of the proposed estimators are shown in Section 3. Although the asymptotic normality follows from a standard argument, the asymptotic covariance matrices of the proposed estimators are very complicated, because of the use of previously estimated parameters in the pseudo likelihoods. We establish an asymptotic representation that allows us to obtain easy-to-compute consistent estimators of the asymptotic covariance matrices. Section 4 contains some empirical results. A discussion on assumptions is given in Section 5. The Appendix contains technical details.

2. ESTIMATION BASED ON PSEUDO LIKELIHOODS

Under assumptions (1) and (2), we consider the estimation of $\theta = (\theta_1, \ldots, \theta_T)$ based on an independent and identically distributed sample $(Y_1^{(i)}, \ldots, Y_T^{(i)}, R_1^{(i)}, \ldots, R_T^{(i)}, X^{(i)})$, $i = 1, \ldots, n$, from the population $p(Y, R, X)$, where $Y_t^{(i)}$ is observed if and only if $R_t^{(i)} = 1$.

2.1 The case where X is a dropout instrument

We first consider the case of $X = Z$ and $U = 0$ in (1), i.e., conditioned on $(Y_1, \ldots, Y_t)$, the dropout propensity does not depend on the entire covariate vector X so that X is a dropout instrument. When $t = 1$, we consider the likelihood

$$\begin{aligned}
&\prod_{R_1^{(i)}=1} p(X^{(i)}|Y_1^{(i)}, R_1^{(i)} = 1) \\
&= \prod_{R_1^{(i)}=1} p(X^{(i)}|Y_1^{(i)}) \\
&= \prod_{R_1^{(i)}=1} \frac{p(Y_1^{(i)}|X^{(i)})p(X^{(i)})}{\int p(Y_1^{(i)}|x)p(x)dx} \\
&= \prod_{R_1^{(i)}=1} \frac{f_1(Y_1^{(i)}|X^{(i)})p(X^{(i)})}{\int f_1(Y_1^{(i)}|x)p(x)dx},
\end{aligned}$$

where the first equality follows from assumption (1) with $U = 0$, the second equality follows from the Bayes formula, and the last equality follows from assumption (2). Substituting $p(X)$ by the nonparametric empirical distribution of X putting mass n^{-1} to each $X^{(j)}$, we obtain an estimator $\hat{\theta}_1$ by maximizing the pseudo likelihood

$$\prod_{R_1^{(i)}=1} \frac{f_1(Y_1^{(i)}|X^{(i)}, \theta_1)}{\sum_{j=1}^{n} f_1(Y_1^{(i)}|X^{(j)}, \theta_1)}.$$

Note that we can also assume a parametric model $p(X) = h_\varphi(X)$ and replace the previous expression by

$$\prod_{R_1^{(i)}=1} \frac{f_1(Y_1^{(i)}|X^{(i)}, \theta_1)h_{\hat{\varphi}}(X^{(i)})}{\int f_1(Y_1^{(i)}|x, \theta_1)h_{\hat{\varphi}}(x)dx},$$

where $\hat{\varphi}$ is an estimator of ϕ using X-data. However, the integral may not have an explicit form and using the empirical distribution of X is more robust.

For $t = 2, \ldots, T$, suppose that $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$ have been obtained. Consider the likelihood

$$\prod_{R_t^{(i)}=1} p(X^{(i)}|Y_1^{(i)}, \ldots, Y_t^{(i)}, R_t^{(i)} = 1)$$
$$= \prod_{R_t^{(i)}=1} \frac{p(Y_1^{(i)}, \ldots, Y_t^{(i)}|X^{(i)})p(X^{(i)})}{\int p(Y_1^{(i)}, \ldots, Y_t^{(i)}|x)p(x)dx}.$$

Under (2),

$$p(Y_1^{(i)}, \ldots, Y_t^{(i)}|X^{(i)}) = f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \theta_s),$$

where $V_s^{(i)} = (Y_1^{(i)}, \ldots, Y_s^{(i)}, X^{(i)})$. Replacing each θ_s by the previously obtained $\hat{\theta}_s$ and $p(X^{(i)})$ by the nonparametric empirical distribution of X, we estimate θ_t by maximizing the pseudo likelihood

(3)
$$\prod_{R_t^{(i)}=1} \frac{f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \hat{\theta}_s)}{\sum_{j=1}^{n} f_t(Y_t^{(i)}|X^{(j)}, \vec{Y}_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|X^{(j)}, \vec{Y}_{s-1}^{(i)}, \hat{\theta}_s)},$$

where $\vec{Y}_{t-1}^{(i)} = (Y_1^{(i)}, \ldots, Y_{t-1}^{(i)})$. Note that all observed values up to time t are included in this likelihood.

It is implicitly assumed that $f_t(Y_t|V_{t-1}, \theta_t)$ depends on X, i.e., X is a useful covariate. Otherwise, $f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t)$ can be canceled in (3) and the likelihood does not contain θ_t.

Let the number of covariates be $K \geq 1$. If X is cross-sectional, then X is K-dimensional. If X is longitudinal, then $X = (X_1, \ldots, X_T)$ and each X_t is K-dimensional so that the dimension of X is KT. For many longitudinal studies, Y_t is statistically related to $X_1, \ldots, X_t$ only, $t = 1, \ldots, T$. In such cases,

$$p(Y_t|V_{t-1}) = p(Y_t|X_1, \ldots, X_t, Y_1, \ldots, Y_{t-1})$$
$$= f_t(Y_t|X_1, \ldots, X_t, Y_1, \ldots, Y_{t-1}, \theta_t)$$

and the proposed pseudo likelihood (3) can be used with $f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t)$ replaced by $f_t(Y_t^{(i)}|X_1^{(i)}, \ldots, X_t^{(i)}, Y_1^{(i)}, \ldots, Y_{t-1}^{(i)}, \theta_t)$, $t = 1, \ldots, T$.

Maximizing (3) can be done using an efficient algorithm (see our discussion in Section 4). If we do not substitute $\theta_1, \ldots, \theta_{t-1}$ by their estimates, in theory we can estimate $(\theta_1, \ldots, \theta_t)$ by maximizing (3) with $\hat{\theta}_s$ replaced by θ_s, $s = 1, \ldots, t-1$. However, the computation may not be feasible because the dimension of $(\theta_1, \ldots, \theta_t)$ is much higher than that of θ_t.

The original approach in Tang et al. (2003) requires a check on whether we can identify θ from the parameters in $p(X|Y)$ because we estimate parameters in $p(Y|X)$ through estimating parameters in $p(X|Y)$ and the Bayes formula. When (Y, X) is multivariate normal, the requirement is that the dimension of X has to be at least T. This restrictive requirement is not needed in our proposed approach under assumption (1), because we estimate θ_t's one at a time. In fact, when $p(Y|X)$ is normal, a one-dimensional continuous X or discrete X taking at least 3 values is sufficient for estimating θ. This can be shown using induction. For $t = 1$, Y_1 is one-dimensional and the result in Tang et al. (2003) showed that θ_1 in $p(Y_1|X) = f_1(Y_1|X, \theta_1)$ can be estimated with a one-dimensional continuous X or discrete X taking at least 3 values when X is related to Y_1. Assuming that $\theta_1, \ldots, \theta_{t-1}$ are estimated, we want to show that θ_t in $p(Y_t|V_{t-1}) = f_t(Y_t|V_{t-1}, \theta_t)$ can be estimated. Since parameters in $p(Y_1, \ldots, Y_{t-1}|X)$ have been estimated, we can treat $(Y_1, \ldots, Y_{t-1}, X)$ as a covariate vector and, thus, θ_t can be estimated based on the result in Tang et al. (2003).

2.2 The case where a sub-vector of X is a dropout instrument

Let $X = (U, Z)$ as in (1). Note that

$$p(Z|Y_1, \ldots, Y_t, U, R_t = 1)$$
$$= p(Z|Y_1, \ldots, Y_t, U)$$
$$= \frac{p(Y_1, \ldots, Y_t, U|Z)p(Z)}{\int p(Y_1, \ldots, Y_t, U|z)p(z)dz}$$
$$= \frac{p(Y_1, \ldots, Y_t|U, Z)p(U|Z)p(Z)}{\int p(Y_1, \ldots, Y_t|U, z)p(U|z)p(z)dz}$$
$$= \frac{p(Y_1, \ldots, Y_t|U, Z)p(Z|U)}{\int p(Y_1, \ldots, Y_t|U, z)p(z|U)dz}.$$

First, if U is a discrete covariate, then we can substitute $p(Z|U = u)$ by the empirical distribution of Z conditioned on $U = u$, which results in the following likelihood for the estimation of θ_t:

$$\prod_{\substack{u,\, U^{(i)}=u \\ R_t^{(i)}=1}} \frac{f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \hat{\theta}_s)}{\sum_{U^{(j)}=u} f_t(Y_t^{(i)}|X^{(j)}, \vec{Y}_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|X^{(j)}, \vec{Y}_{s-1}^{(i)}, \hat{\theta}_s)},$$

where $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$ are estimators from the previous steps. Next, consider the case where U is continuous and a parametric model on $p(Z|U) = g_\xi(Z|U)$ is assumed, where ξ is an unknown parameter vector. Since U and Z have no missing data, ξ can be estimated by $\hat{\xi}$ using the likelihood based on $X^{(1)}, \ldots, X^{(n)}$, which leads to the following likelihood for the estimation of θ_t:

$$\prod_{R_t^{(i)}=1} \frac{f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \hat{\theta}_s) g_{\hat{\xi}}(Z^{(i)}|U^{(i)})}{\int f_t(Y_t^{(i)}|C_{t-1}^{(i)}, z, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|C_{s-1}^{(i)}, z, \hat{\theta}_s) g_{\hat{\xi}}(z|U^{(i)})dz},$$

where $C_t^{(i)} = (U^{(i)}, Y_1^{(i)}, \ldots, Y_t^{(i)})$. Finally, consider the case where U is continuous, a parametric model on $p(U|Z) = h_\zeta(U|Z)$ is assumed, where ζ is an unknown parameter vector, and ζ is estimated by $\hat{\zeta}$ using the likelihood based on

$X^{(1)},\ldots,X^{(n)}$. Then, the following likelihood can be used for the estimation of θ_t:

$$\prod_{R_t^{(i)}=1} \frac{f_t(Y_t^{(i)}|V_{t-1}^{(i)},\theta_t)\prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)},\hat{\theta}_s)h_{\hat{\zeta}}(U^{(i)}|Z^{(i)})}{\sum_{j=1}^{n} f_t(Y_t^{(i)}|W_{t-1}^{(i,j)},\theta_t)\prod_{s=1}^{t-1} f_s(Y_s^{(i)}|W_{s-1}^{(i,j)},\hat{\theta}_s)h_{\hat{\zeta}}(U^{(i)}|Z^{(j)})},$$

where $W_t^{(i,j)} = (U^{(i)}, Z^{(j)}, Y_1^{(i)},\ldots,Y_t^{(i)})$. In any case it is assumed that $f_t(Y_t|V_{t-1},\theta_t)$ depends on Z, i.e., Z is a useful covariate, although $f_t(Y_t|V_{t-1},\theta_t)$ may not depend on U.

3. ASYMPTOTIC PROPERTIES

Under some regularity conditions, we now show that $\hat{\theta}_t$, $t = 1,\ldots,T$, are consistent and asymptotically normal as $n \to \infty$. For simplicity, we focus on the situation where $X = Z$ (Section 2.1). Results for the situations described in Section 2.2 can be similarly derived. In addition to (1) and (2), the following are two key conditions for the consistency of $\hat{\theta}_t$:

$$\pi_t = P(R_t = 1) > 0, \quad t = 1,\ldots,T, \tag{4}$$

and, for any θ_t in the parameter space that is not the same as the true parameter value θ_t^0 and any function ϕ of $(Y_1,\ldots,Y_t,\theta_t)$,

$$P\left(\vec{Y}_t : \frac{f_t(Y_t|V_{t-1},\theta_t)}{f_t(Y_t|V_{t-1},\theta_t^0)} = \phi(\vec{Y}_t,\theta_t) \text{ for any } X\right) < 1, \tag{5}$$

where $\vec{Y}_t = (Y_1,\ldots,Y_t)$. We now explain why $\hat{\theta}_t$ is consistent under (4)–(5). Let F denote the distribution of X and, for any t, $\varphi_t = (\theta_1,\ldots,\theta_t,F)$ $(\varphi_0 = F)$,

$$G_t(\varphi_t) = \frac{f_t(Y_t|V_{t-1},\theta_t)\prod_{s=1}^{t-1} f_s(Y_s|V_{s-1},\theta_s)dF(X)}{\int f_t(Y_t|x,\vec{Y}_{t-1},\theta_t)\prod_{s=1}^{t-1} f_s(Y_s|x,\vec{Y}_{s-1},\theta_s)dF(x)},$$

and $H_t(\varphi_t) = R_t \log G_t(\varphi_t)$. Let $\varphi_t^0 = (\theta_1^0,\ldots,\theta_t^0,F^0)$ be the true value of φ_t. Then

$$\begin{aligned} &E[H_t(\theta_t,\varphi_{t-1}^0)] - E[H_t(\varphi_t^0)] \\ &\quad = E\left\{R_t \log\frac{G_t(\theta_t,\varphi_{t-1}^0)}{G_t(\varphi_t^0)}\right\} \\ &\quad = \pi_t E\left\{\log\frac{G_t(\theta_t,\varphi_{t-1}^0)}{G_t(\varphi_t^0)}\right\} \\ &\quad \le \pi_t \log E\left\{\frac{G_t(\theta_t,\varphi_{t-1}^0)}{G_t(\varphi_t^0)}\right\} \\ &\quad = 0 \end{aligned}$$

with the equality holds if and only if (since $\pi_t > 0$)

$$\frac{G_t(\theta_t,\varphi_{t-1}^0)}{G_t(\varphi_t^0)} = 1 \text{ a.s.},$$

which is, by the definition of G_t function, equivalent to that, almost surely,

$$\begin{aligned} &\frac{f_t(Y_t|V_{t-1},\theta_t)}{f_t(Y_t|V_{t-1},\theta_t^0)} \\ &\quad = \frac{\int f_t(Y_t|x,\vec{Y}_{t-1},\theta_t)\prod_{s=1}^{t-1} f_s(Y_s|x,\vec{Y}_{s-1},\theta_s^0)dF^0(x)}{\int f_t(Y_t|x,\vec{Y}_{t-1},\theta_t^0)\prod_{s=1}^{t-1} f_s(Y_s|x,\vec{Y}_{s-1},\theta_s^0)dF^0(x)}, \end{aligned}$$

a function of $(Y_1,\ldots,Y_t,\theta_t)$. Therefore, conditions (4) and (5) ensure that

$$E[H_t(\theta_t,\varphi_{t-1}^0)] < E[H_t(\varphi_t^0)].$$

This means that the expectation of the log of the likelihood function in (3) has a unique maximum at $\theta_t = \theta_t^0$. Let $H_t^{(i)}(\varphi_t)$ be defined as $H_t(\varphi_t)$ but based on data from the ith subject. Then, the estimator $\hat{\theta}_t$ obtained by maximizing (3) satisfies

$$\sum_{i=1}^{n} H_t^{(i)}(\hat{\theta}_t,\hat{\varphi}_{t-1}) = \max_{\theta_t}\sum_{i=1}^{n} H_t^{(i)}(\theta_t,\hat{\varphi}_{t-1}),$$

where $\hat{\varphi}_{t-1} = (\hat{\theta}_1,\ldots,\hat{\theta}_{t-1},\hat{F})$, $\hat{\theta}_1,\ldots,\hat{\theta}_{t-1}$ are estimators from the previous steps, and $\hat{F}$ is the empirical distribution based on $X^{(j)}$, $j = 1,\ldots,n$. Under some regularity conditions (such as those given in Theorem 1 of Tang et al., 2003), $\hat{\theta}_t$ converges in probability to the unique maximum point θ_t^0.

Asymptotic normality of $\hat{\theta}_t$, which is crucial for large sample inference, can be established using a standard argument. Our contribution is to derive an asymptotic representation of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$, which allows us to obtain an easy-to-compute consistent estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$ without knowing its actual form. The asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$ is very complicated because of the fact that $\hat{\theta}_t$ is defined in terms of previous estimators $\hat{\theta}_1,\ldots,\hat{\theta}_{t-1}$ and $\hat{F}$. As we discussed in Section 2.1, without using $\hat{\theta}_1,\ldots,\hat{\theta}_{t-1}$, the estimation of θ_t may not be computationally feasible.

Theorem 3.1. *Assume (1), (2), (4), (5), and the following two conditions.*

C1 The functions f_t's in (2) are continuously twice differentiable with respect to θ_t and $E[\frac{\partial^2 H_t(\varphi_t^0)}{\partial\theta_t\partial\theta_t'}]$ is positive definite.

C2 There exists an open subset Ω_t containing θ_t^0 such that

$$\sup_{\theta_t\in\Omega_t}\left\|\frac{\partial^2 H_t(\theta_t,\varphi_{t-1}^0)}{\partial\theta_t\partial\theta_j'}\right\| < M_{tj}, \quad j = 1,\ldots,t,$$

where M_{tj} are integrable functions and $\|A\|^2 = \text{trace}(A'A)$ for a matrix A.

Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_t - \theta_t^0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_t(W_t^{(i)},A_t,\varphi_t^0) + o_p(1) \to_d N(0,\Sigma_t), \tag{6}$$

where $\to_d$ *denotes convergence in distribution,* $o_p(1)$ *denotes a quantity converging to 0 in probability,* Σ_t *is the covariance matrix of* $\psi_t(W_t^{(i)}, A_t, \varphi_t^0)$, $W_t^{(i)} = (V_t^{(i)}, R_t^{(i)})$, $i = 1, \ldots, n$, $A_1 = A_{11}$, $A_t = (A_{t-1}, A_{t1}, \ldots, A_{tt}), t \geq 2$,

$$A_{tj} = E\left[\frac{\partial^2 H_t(\varphi_t^0)}{\partial\theta_t \partial\theta_j'}\right], \quad j = 1, \ldots, t,$$

and ψ_t *is a known function defined in (9)–(10) of the Appendix,* $t = 1, \ldots, T$.

The functions ψ_t, $t = 1, \ldots, T$, are defined iteratively according to (9)–(10) and, hence, their covariance matrices are very complicated. The explicit forms of ψ_t, when $t = 1, 2, 3, 4$, are given in the Appendix. One may apply the bootstrap method to obtain estimators of Σ_t's, but in each bootstrap replication, maximizing a bootstrap analog of (3) is required, which results in a very large amount of computation. Instead, we propose the following estimator of Σ_t, utilizing the representation in (6). Let $D_t^{(i)} = \psi_t(W_t^{(i)}, A_t, \varphi_t^0)$. Since $\Sigma_t = \mathrm{Var}(D_t^{(i)})$, the sample covariance matrix based on $D_t^{(1)}, \ldots, D_t^{(n)}$ is a consistent estimator of Σ_t. However, $D_t^{(i)}$ contains the unknown φ_t^0 and A_t. Substituting $D_t^{(i)}$ by $\hat{D}_t^{(i)} = \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t)$, $i = 1, \ldots, n$, where $\hat{A}_t = (\hat{A}_{t-1}, \hat{A}_{t1}, \ldots, \hat{A}_{tt})$ and

$$\hat{A}_{tj} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 H_t^{(i)}(\varphi_t)}{\partial\theta_t \partial\theta_j'}\Big|_{\varphi_t = \hat{\varphi}_t}, \quad j = 1, \ldots, t,$$

we define the sample covariance matrix based on $\hat{D}_t^{(1)}, \ldots, \hat{D}_t^{(n)}$ as our estimator $\hat{\Sigma}_t$. This estimator is easy to compute, using (9)–(10) in the Appendix. Under some conditions, $\hat{\Sigma}_t$ is consistent, which is proved in the Appendix.

Theorem 3.2. *Assume that the conditions in Theorem 3.1 hold and that*

C3 $\sup_{\|w\| \leq c} \|\psi_t(w, \hat{A}_t, \hat{\varphi}_t) - \psi_t(w, A_t, \varphi_t^0)\| = o_p(1)$ *for any* $c > 0$.

C4 There exist a constant $c_0 > 0$ *and a function* $h(w) \geq 0$ *such that* $E[h(W_t^{(1)})] < \infty$ *and* $P(\|\psi_t(w, \hat{A}_t, \hat{\varphi}_t)\|^2 \leq h(w)$ *for all* $\|w\| \geq c_0) \to 1$.

Then, as $n \to \infty$, $\|\hat{\Sigma}_t - \Sigma_t\| = o_p(1)$.

4. SOME EMPIRICAL RESULTS

In this section, we present some results based on a real data set and a simulation study.

4.1 Estimation based on HIV-CD4 data

We applied the proposed method to a longitudinal data set from the study of HIV-AIDS patients with advanced immune suppression, conducted by the AIDS Clinical Trial Group 193A. Patients were randomized to one of the four daily regimens containing 600mg of zidovudine: zidovudine alternating monthly with 400mg didanosine (Treatment 1), zidovudine plus 2.25mg of zalcitabine (Treatment 2), zidovudine plus 400mg of didanosine (Treatment 3), and zidovudine plus 400mg of didanosine and 400mg of nevirapine (Treatment 4). The data set can be accessed at the following website: "http://biosun1.harvard.edu/˜fitzmaur/ala/cd4.txt".

For the HIV study, the CD4 cell count, which decreases as HIV progresses, is of prime interest. CD4 counts were collected from patients before the treatments were applied (baseline measurements). After the treatments were applied, CD4 counts were collected from each patient every 8 weeks. In this dataset, there were originally 1,309 patients, but 10 of them did not have baseline measurements and were ignored from our analysis. Also, we ignored measurements from 18 patients in the week interval $(0, 4]$. We considered the first $T = 4$ follow-up time points. For each patient, the tth observation is the one closest to week $8t$ in the interval $(8t - 4, 8t + 4]$, $t = 1, 2, 3, 4$. Following the approach in Robins, Rotnitzky and Zhao (1995), we ignored subsequent data from any patient after the first missed clinic visit to obtain a data set with monotone missing (dropout). The following is a summary of the number of observed values by time points and treatment.

Treatment	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
1	320	223	174	127	110
2	322	218	184	143	110
3	327	221	184	135	108
4	330	235	187	136	116
Total	1299	897	729	541	444

The average dropout proportion for 4 time points $t = 1, 2, 3, 4$ are 31.9%, 43.9%, 58.4%, and 66.8%, respectively.

To apply the proposed method, we considered log(CD4+1) at time point t as Y_t and log(baseline measurement +1) as the dropout instrument Z. Because the baseline measurements were taken before the treatments were applied, it is reasonable to assume that the dropout propensity at time t does not depend on Z given the CD4 counts at time $1, \ldots, t$. We assumed that

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11} Z + \varepsilon_1 \\ Y_2 &= \beta_{20} + \beta_{21} Z + \beta_{22} Y_1 + \varepsilon_2 \\ Y_3 &= \beta_{30} + \beta_{31} Z + \beta_{32} Y_1 + \beta_{33} Y_2 + \varepsilon_3 \\ Y_4 &= \beta_{40} + \beta_{41} Z + \beta_{42} Y_1 + \beta_{43} Y_2 + \beta_{44} Y_3 + \varepsilon_4, \end{aligned} \tag{7}$$

where $\varepsilon_t \sim N(0, \sigma_t^2)$, $t = 1, \ldots, 4$, ε_t's are independent, and β_{tj}'s and σ_t's are unknown parameters.

Tables 1–4 display estimates of parameters based on the HIV-CD4 data under treatments 1–4, respectively, and their standard errors (SE). For each parameter, we computed two estimates, the proposed estimate and the estimate obtained by regression (Paik, 1997) under the ignorable dropout assumption (which is denoted as the MAR estimate since ignorable missing is also called missing at random). In maximizing (3), we applied the Nelder-Mead simplex algorithm

Table 1. Estimates and SE's under treatment 1 of the HIV-AIDS study

Parameter	β_{10}	β_{11}	σ_1			
MAR	0.127	0.913	0.696			
SE	0.203	0.062	0.050			
Proposed	−0.109	0.994	0.741			
SE	0.248	0.079	0.059			
Difference	0.235	−0.081	−0.045			
SE	0.132	0.048	0.023			
p-value	0.075	0.090	0.049			
Parameter	β_{20}	β_{21}	β_{22}	σ_2		
MAR	0.171	0.512	0.353	0.677		
SE	0.243	0.101	0.089	0.048		
Proposed	0.001	0.761	0.130	0.798		
SE	0.305	0.153	0.132	0.094		
Difference	0.170	−0.241	0.222	−0.121		
SE	0.186	0.086	0.074	0.039		
p-value	0.361	0.005	0.003	0.002		
Parameter	β_{30}	β_{31}	β_{32}	β_{33}	σ_3	
MAR	0.254	0.203	0.290	0.324	0.570	
SE	0.223	0.093	0.111	0.086	0.050	
Proposed	0.164	0.041	0.487	0.364	0.199	
SE	0.306	0.103	0.188	0.133	0.101	
Difference	0.090	0.162	−0.197	−0.041	0.371	
SE	0.165	0.071	0.107	0.094	0.062	
p-value	0.585	0.022	0.067	0.663	0.000	
Parameter	β_{40}	β_{41}	β_{42}	β_{43}	β_{44}	σ_4
MAR	0.482	−0.084	−0.048	0.489	0.380	0.648
SE	0.256	0.136	0.103	0.157	0.172	0.067
Proposed	0.504	−0.109	−0.052	0.541	0.320	0.633
SE	0.312	0.124	0.115	0.176	0.154	0.106
Difference	−0.022	0.025	0.005	−0.052	0.060	0.015
SE	0.172	0.095	0.074	0.118	0.119	0.047
p-value	0.900	0.791	0.952	0.661	0.616	0.749

Table 2. Estimates and SE's under treatment 2 of the HIV-AIDS study

Parameter	β_{10}	β_{11}	σ_1			
MAR	0.586	0.809	0.812			
SE	0.237	0.076	0.062			
Proposed	0.278	0.928	0.874			
SE	0.285	0.089	0.064			
Difference	0.308	−0.119	−0.062			
SE	0.1778	0.066	0.039			
p-value	0.083	0.071	0.115			
Parameter	β_{20}	β_{21}	β_{22}	σ_2		
MAR	0.175	0.409	0.461	0.703		
SE	0.215	0.094	0.083	0.052		
Proposed	0.521	0.321	0.428	0.578		
SE	0.255	0.165	0.118	0.052		
Difference	−0.346	0.088	0.033	0.124		
SE	0.152	0.072	0.069	0.034		
p-value	0.022	0.225	0.636	0.000		
Parameter	β_{30}	β_{31}	β_{32}	β_{33}	σ_3	
MAR	−0.083	0.088	0.344	0.482	0.743	
SE	0.239	0.133	0.123	0.110	0.060	
Proposed	−0.097	−0.107	0.294	0.686	0.152	
SE	0.319	0.175	0.091	0.115	0.074	
Difference	0.014	0.194	0.049	−0.204	0.591	
SE	0.168	0.107	0.082	0.093	0.057	
p-value	0.933	0.068	0.545	0.029	0.000	
Parameter	β_{40}	β_{41}	β_{42}	β_{43}	β_{44}	σ_4
MAR	−0.086	0.279	−0.146	0.319	0.501	0.664
SE	0.285	0.121	0.089	0.113	0.141	0.073
Proposed	−0.094	0.315	−0.234	0.288	0.652	0.779
SE	0.311	0.114	0.105	0.111	0.138	0.084
Difference	0.008	−0.036	0.089	0.031	−0.151	−0.115
SE	0.122	0.085	0.068	0.096	0.131	0.061
p-value	0.946	0.676	0.195	0.744	0.250	0.059

via MATLAB under the UNIX environment. This is a direct search algorithm that does not use numerical or analytic gradients. We found that it was more stable than gradient-based methods. The MAR estimates were used as initial values in maximizing (3). The SE's of the proposed estimates were calculated using the results in Theorem 2 in Section 3. To compare, we also computed the difference of the MAR estimate and the proposed estimate, its SE, and the two-sided p-value of testing whether two estimates are the same. The SE's of the MAR estimates and the differences were computed by bootstrapping.

It can be seen from Tables 1–4 that the differences between the MAR and proposed estimates are not negligible in some cases (p-values less or nearly equal to 5%) while in some cases the two estimates are about the same.

4.2 A simulation study

A simulation study was conducted under model (7) with $n = 300$ and parameters equal to the estimated values under Treatment 3 in the HIV-CD4 example. These values are shown in Table 5. The covariate Z was generated from $N(2.9065, 0.9544^2)$, where the parameters are the estimates of the baseline CD4. The dropout indicators at time points $t = 1, 2, 3, 4$ were generated from the following logistic model:

$$
\begin{aligned}
&P(R_1 = 1|Z, Y_1) \\
&\quad = 1 - [1 + \exp(-9 + 4Y_1)]^{-1}, \\
&P(R_2 = 1|Z, Y_1, Y_2, R_1 = 1) \\
&\quad = 1 - [1 + \exp(-14 + Y_1 + 5Y_2)]^{-1}, \\
&P(R_3 = 1|Z, Y_1, Y_2, Y_3, R_2 = 1) \\
&\quad = 1 - [1 + \exp(-18 + 2Y_2 + 4Y_3)]^{-1}, \\
&P(R_4 = 1|Z, Y_1, Y_2, Y_3, Y_4, R_3 = 1) \\
&\quad = 1 - [1 + \exp(-15 + 2Y_3 + 3Y_4)]^{-1}.
\end{aligned} \tag{8}
$$

The parameters in (8) were chosen so that the unconditional dropout rates are similar to the observed dropout proportions under Treatment 3 of the HIV-CD4 data, approximately 30%, 40%, 60%, and 70% for $t = 1, 2, 3$, and 4, respectively.

Table 3. Estimates and SE's under treatment 3 of the HIV-AIDS study

Parameter	β_{10}	β_{11}	σ_1			
MAR	0.836	0.771	0.940			
SE	0.305	0.095	0.058			
Proposed	0.555	0.859	1.015			
SE	0.401	0.126	0.083			
Difference	0.281	−0.088	−0.075			
SE	0.232	0.081	0.054			
p-value	0.227	0.276	0.162			
Parameter	β_{20}	β_{21}	β_{22}	σ_2		
MAR	0.024	0.347	0.609	0.750		
SE	0.284	0.107	0.079	0.057		
Proposed	0.007	0.421	0.547	0.802		
SE	0.326	0.155	0.150	0.182		
Difference	0.018	−0.074	0.062	−0.052		
SE	0.152	0.096	0.058	0.056		
p-value	0.907	0.441	0.285	0.356		
Parameter	β_{30}	β_{31}	β_{32}	β_{33}	σ_3	
MAR	0.044	0.096	0.150	0.687	0.632	
SE	0.205	0.092	0.087	0.143	0.067	
Proposed	0.387	0.172	0.079	0.588	0.941	
SE	0.358	0.134	0.118	0.133	0.073	
Difference	−0.343	−0.077	0.072	0.098	−0.310	
SE	0.162	0.087	0.068	0.124	0.032	
p-value	0.034	0.377	0.295	0.427	0.000	
Parameter	β_{40}	β_{41}	β_{42}	β_{43}	β_{44}	σ_4
MAR	−0.331	0.097	0.095	0.347	0.461	0.624
SE	0.224	0.129	0.129	0.120	0.113	0.068
Proposed	−0.349	0.110	0.089	0.355	0.451	0.601
SE	0.249	0.128	0.130	0.092	0.128	0.064
Difference	0.018	−0.014	0.006	−0.008	0.011	0.022
SE	0.169	0.109	0.114	0.068	0.109	0.058
p-value	0.915	0.900	0.956	0.908	0.921	0.694

Table 4. Estimates and SE's under treatment 4 of the HIV-AIDS study

Parameter	β_{10}	β_{11}	σ_1			
MAR	0.594	0.916	0.839			
SE	0.175	0.057	0.055			
Proposed	0.298	1.046	0.897			
SE	0.251	0.074	0.073			
Difference	0.296	−0.130	−0.058			
SE	0.174	0.067	0.036			
p-value	0.088	0.051	0.106			
Parameter	β_{20}	β_{21}	β_{22}	σ_2		
MAR	0.433	0.113	0.728	0.697		
SE	0.202	0.101	0.085	0.045		
Proposed	0.178	0.131	0.677	1.624		
SE	0.164	0.119	0.127	0.083		
Difference	0.256	−0.019	0.051	−0.927		
SE	0.170	0.116	0.075	0.086		
p-value	0.132	0.873	0.492	0.000		
Parameter	β_{30}	β_{31}	β_{32}	β_{33}	σ_3	
MAR	0.126	0.110	0.281	0.511	0.663	
SE	0.185	0.097	0.149	0.125	0.055	
Proposed	0.136	0.382	0.045	0.524	1.278	
SE	0.246	0.114	0.126	0.138	0.074	
Difference	−0.010	−0.272	0.236	−0.013	−0.615	
SE	0.169	0.086	0.106	0.120	0.075	
p-value	0.952	0.002	0.026	0.914	0.000	
Parameter	β_{40}	β_{41}	β_{42}	β_{43}	β_{44}	σ_4
MAR	−0.149	0.265	0.031	0.114	0.589	0.606
SE	0.257	0.120	0.136	0.114	0.131	0.049
Proposed	−0.141	0.263	0.012	0.117	0.622	0.586
SE	0.237	0.112	0.131	0.120	0.129	0.052
Difference	−0.008	0.002	0.019	−0.004	−0.033	0.020
SE	0.182	0.114	0.127	0.105	0.123	0.047
p-value	0.965	0.986	0.881	0.971	0.789	0.670

We studied the MAR estimates (Paik, 1997) based on the ignorable assumption, the method discussed in Tang et al. (2003), and the proposed method. As a standard, we also included the standard regression method when there is no dropout.

Based on 1,000 simulation runs, Table 5 reports the bias for parameter estimation, standard deviation (SD) of the parameter estimate, standard error (SE), which is an estimate of SD, and the coverage probability (CP) of the approximate 95% confidence intervals of the parameter, using estimate ± 1.96SE. The SE's for the method in Tang et al. (2003) and the proposed method are obtained based on Theorem 2 and the SE's for the MAR estimates are computed by bootstrapping. The results in Table 5 show that the proposed estimators and their SE's work well, and the MAR estimators are biased and the biases are large enough to result in poor CP. The SD's of the MAR estimators, however, may be smaller than those of the proposed estimators. Hence, the MAR estimators may be more efficient when they are nearly unbiased, e.g., when the ignorable dropout assumption holds. When $t = 1$, the method in Tang et al. (2003) and the proposed method are the same. When $t > 1$, however, the method in Tang et al. (2003) may also produce biased estimators since it is based on a stronger assumption than (8). In addition, when the estimators in Tang et al. (2003) are approximately unbiased, the corresponding SD's are larger than those of the proposed method, which illustrates that our proposed method is more efficient since we used all observed data in the estimation procedure.

5. DISCUSSION ON ASSUMPTIONS

The key assumptions for our approach are (1) and (2). As we discussed in Section 1, to identify the unknown parameters, it is necessary that at least one component of $V_t = (Y_1, \ldots, Y_t, X)$ is not related to dropout at time point t, conditioned on the other components. This component is Y_t under the ignorable dropout assumption, whereas it is a component Z of X under our assumption (1). Unfortunately, none of these assumptions on the dropout mechanism can be checked using data due to the presence of missing values. We have to carefully study each particular problem and decide

Table 5. Simulation results under dropout mechanism (8), $n = 300$

	No dropout				MAR				Tang et al. (2003)				Proposed			
True value	bias	SD	SE	CP	bias	SD	SE	CP	bias	SD	SE	CP	bias	SD	SE	CP
β_{10}=0.555	−0.006	0.186	0.187	94.7	1.222	0.216	0.209	0.0	−0.029	0.362	0.353	95.6	−0.029	0.362	0.353	95.6
β_{11}=0.859	0.002	0.061	0.061	94.2	−0.275	0.069	0.067	2.0	0.008	0.101	0.101	96.2	0.008	0.101	0.101	96.2
σ_1=1.015	−0.001	0.041	0.041	94.0	−0.181	0.041	0.041	1.3	−0.000	0.098	0.092	93.7	−0.000	0.098	0.092	93.7
β_{20}=0.007	0.003	0.150	0.150	94.4	0.880	0.249	0.253	7.9	−0.174	0.353	0.358	93.0	−0.003	0.350	0.354	96.5
β_{21}=0.421	−0.001	0.064	0.062	94.4	−0.081	0.075	0.074	78.3	−0.098	0.158	0.166	84.7	0.003	0.138	0.154	97.0
β_{22}=0.547	0.001	0.048	0.046	93.9	−0.118	0.068	0.066	55.7	0.102	0.126	0.139	79.2	−0.002	0.113	0.123	97.1
σ_2=0.802	−0.002	0.033	0.032	93.6	−0.083	0.039	0.038	41.3	−0.083	0.153	0.142	91.8	−0.013	0.147	0.144	95.9
β_{30}=0.387	0.000	0.179	0.177	94.2	1.641	0.411	0.410	3.5	−0.392	0.298	0.307	70.5	0.003	0.323	0.337	95.4
β_{31}=0.172	−0.001	0.077	0.079	94.8	−0.048	0.108	0.108	91.3	−0.110	0.137	0.129	79.5	−0.010	0.101	0.107	94.9
β_{32}=0.079	0.001	0.064	0.065	94.9	−0.007	0.102	0.098	93.5	−0.009	0.117	0.138	94.7	−0.007	0.118	0.116	93.8
β_{33}=0.588	0.000	0.066	0.068	95.3	−0.281	0.114	0.109	29.5	0.062	0.138	0.135	92.1	−0.006	0.131	0.133	96.2
σ_3=0.941	0.000	0.038	0.038	94.6	−0.142	0.054	0.052	26.4	0.088	0.095	0.113	78.3	0.012	0.077	0.081	94.6
β_{40}=−0.349	0.002	0.111	0.114	95.7	0.696	0.389	0.387	55.5	−0.305	0.251	0.268	73.2	−0.012	0.243	0.253	95.7
β_{41}=0.110	0.000	0.052	0.051	94.4	−0.015	0.088	0.087	93.0	−0.024	0.138	0.156	95.6	−0.008	0.131	0.134	94.5
β_{42}=0.089	−0.000	0.043	0.042	94.0	−0.010	0.074	0.077	95.4	0.020	0.133	0.137	94.4	0.008	0.121	0.131	94.8
β_{43}=0.355	−0.000	0.049	0.049	94.8	−0.022	0.083	0.087	95.0	0.088	0.103	0.108	79.1	−0.008	0.079	0.081	95.2
β_{44}=0.451	−0.001	0.037	0.037	95.0	−0.099	0.078	0.078	74.4	0.053	0.117	0.108	93.6	0.008	0.118	0.119	93.7
σ_4=0.601	−0.002	0.024	0.024	93.8	−0.037	0.042	0.040	81.9	−0.032	0.080	0.092	92.7	−0.013	0.071	0.068	94.8

which assumption is reasonable or approximately holds. In the HIV-CD4 problem, for example, the difference between the MAR estimate and our proposed estimate is whether the current response Y_t or the baseline response Z is related to dropout at time point t, given the other values. Since Y_t is a more recent value for each patient at time point t, we think that our assumption is more reasonable.

It is important to develop estimation methods under various assumptions on the dropout mechanism. The results will be useful as different tools for application and/or for a sensitivity analysis under different assumptions.

We also need to assume at least one of $p(Y|X)$ and $p(R|Y, X)$ is parametric to be able to identify parameters. Again, with missing data, we are not able to verify a parametric model such as (2) using observed data. This is because the parametric model is imposed on the density $p(Y_t|X, Y_1, \ldots, Y_{t-1})$, which is a mixture of $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 1)$ and $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 0)$, and we are not able to check a parametric model assumption on $p(Y_t|X, Y_1, \ldots, Y_{t-1}, R_t = 0)$ since no Y_t-observation comes from it. Parametric models may be sensitive to model violations. The same issue exists for the likelihood approach in Little and Rubin (2002) under ignorable nonresponse. The robustness of the proposed method against violation of assumption (2) is under further investigation.

APPENDIX A. PROOFS

Proof of Theorem 3.1. Let

$$l_t(\theta_t, \hat{\varphi}_{t-1}) = \frac{1}{n}\sum_{i=1}^{n} H_t^{(i)}(\theta_t, \hat{\varphi}_{t-1})$$

and $\nabla l_t(\theta_t, \hat{\varphi}_{t-1})$ be its derivative with respect to θ_t. We first prove the case of $t = 1$. By Taylor's expansion and the fact that $\hat{\varphi}_0 = \hat{F}$ and $\nabla l_1(\hat{\theta}_1, \hat{F}) = 0$, we have

$$\begin{aligned}
&-\nabla l_1(\theta_1^0, F^0)\\
&\quad = \nabla l_1(\theta_1^0, \hat{F}) - \nabla l_1(\theta_1^0, F^0) + \nabla l_1(\hat{\theta}_1, \hat{F}) - \nabla l_1(\theta_1^0, \hat{F})\\
&\quad = \nabla l_1(\theta_1^0, \hat{F}) - \nabla l_1(\theta_1^0, F^0) + \nabla^2 l_1(\theta_1^0, \hat{F})(\hat{\theta}_1 - \theta_1^0)\\
&\qquad + o_p(n^{-1/2}),
\end{aligned}$$

where ∇^2 is the second order derivative with respect to θ_1 and $o_p(n^{-1/2}) = n^{-1/2}o_p(1)$. Note that

$$\begin{aligned}
&\nabla l_1(\theta_1^0, \hat{F}) - \nabla l_1(\theta_1^0, F^0)\\
&\quad = \frac{1}{n}\sum_{i=1}^{n} R_1^{(i)} \left\{ \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d\hat{F} \int f_1(Y_1^{(i)}|x, \theta_1^0)dG}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0 \int f_1(Y_1^{(i)}|x, \theta_1^0)d\hat{F}} \right.\\
&\qquad \left. - \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)dG}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0} \right\}\\
&\quad = \frac{1}{n}\sum_{i=1}^{n} R_1^{(i)} \left\{ \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d\hat{F} \int f_1(Y_1^{(i)}|x, \theta_1^0)dG}{[\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0]^2} \right.\\
&\qquad \left. - \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)dG}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0} \right\} + o_p(n^{-1/2})
\end{aligned}$$

where $G = \hat{F} - F^0$. Let

$$\begin{aligned}
g_1(Y_1^{(i)}, \varphi_1^0) &= \int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0,\\
\nabla g_1(Y_1^{(i)}, \varphi_1^0) &= \int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)dF^0.
\end{aligned}$$

Then

$$\nabla l_1(\theta_1^0, \hat{F}) - \nabla l_1(\theta_1^0, F^0) = B_{n1} + o_p(n^{-1/2}),$$

where

$$\begin{aligned} B_{n1} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{R_1^{(i)}\nabla g_1(Y_1^{(i)},\varphi_1^0) f_1(Y_1^{(i)}|X^{(j)},\theta_1^0)}{[g_1(Y_1^{(i)},\varphi_1^0)]^2}\right. \\ &\quad \left. - \frac{R_1^{(i)}\nabla f_1(Y_1^{(i)}|X^{(j)},\theta_1^0)}{g_1(Y_1^{(i)},\varphi_1^0)}\right\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h_1(W_1^{(i)}, W_1^{(j)}) \end{aligned}$$

is a V-statistic with the following kernel:

$$\begin{aligned} &h_1(W_1^{(i)}, W_1^{(j)}) \\ &\quad = \frac{1}{2}\left\{\frac{R_1^{(i)}\nabla g_1(Y_1^{(i)},\varphi_1^0) f_1(Y_1^{(i)}|X^{(j)},\theta_1^0)}{[g_1(Y_1^{(i)},\varphi_1^0)]^2}\right. \\ &\qquad - \frac{R_1^{(i)}\nabla f_1(Y_1^{(i)}|X^{(j)},\theta_1^0)}{g_1(Y_1^{(i)},\varphi_1^0)} \\ &\qquad + \frac{R_1^{(j)}\nabla g_1(Y_1^{(j)},\varphi_1^0) f_1(Y_1^{(j)}|X^{(i)},\theta_1^0)}{[g_1(Y_1^{(j)},\varphi_1^0)]^2} \\ &\qquad \left. - \frac{R_1^{(j)}\nabla f_1(Y_1^{(j)}|X^{(i)},\theta_1^0)}{g_1(Y_1^{(j)},\varphi_1^0)}\right\}. \end{aligned}$$

Let

$$\begin{aligned} &h_{11}(W_1^{(i)}) \\ &\quad = E[h_1(W_1^{(i)}, W_1^{(j)}) \mid W_1^{(i)}] \\ &\quad = \frac{\pi_1}{2} E\left\{\frac{\nabla g_1(Y_1^{(j)},\varphi_1^0) f_1(Y_1^{(j)}|X^{(i)},\theta_1^0)}{[g_1(Y_1^{(j)},\varphi_1^0)]^2}\right. \\ &\qquad \left. - \frac{\nabla f_1(Y_1^{(j)}|X^{(i)},\theta_1^0)}{g_1(Y_1^{(j)},\varphi_1^0)}\,\middle|\, R_1^{(i)} = 1, X^{(i)}\right\}, \end{aligned}$$

which is a function of φ_1^0 and $X^{(i)}$, not depending on $Y_1^{(i)}$ or $R_1^{(i)}$. We denote this function as $h_{11}(X^{(i)}, \varphi_1^0)$. From the theory of V-statistics,

$$B_{n1} = \frac{1}{n}\sum_{i=1}^{n} 2h_{11}(X^{(i)}, \theta_1^0) + o_p(n^{-1/2}).$$

Under the given regularity conditions, $\nabla^2 l_1(\theta_1^0, \hat{F}) - A_{11} = o_p(1)$. Therefore,

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_1 - \theta_1^0) \\ &\quad = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} A_{11}^{-1}\left\{\frac{\partial H_1^{(i)}(\theta_1^0, F^0)}{\partial\theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0)\right\} \\ &\quad + o_p(1). \end{aligned}$$

Thus, result (6) with $t = 1$ follows by letting

(9)

$$\psi_1(W_1^{(i)}, A_1, \varphi_1^0) = -A_{11}^{-1}\left\{\frac{\partial H_1^{(i)}(\theta_1^0, F^0)}{\partial\theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0)\right\}.$$

Now, suppose that we have obtained result (6) for $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$. Let's prove (6) for $\hat{\theta}_t$. Note that

$$\begin{aligned} &-\nabla l_t(\varphi_t^0) \\ &\quad = \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \theta_1^0, \hat{F}) - \nabla l_t(\varphi_t^0) \\ &\qquad + \nabla l_t(\theta_t^0, \hat{\varphi}_{t-1}) - \nabla l_t(\theta_t^0, \theta_{t-1}^0, \hat{\varphi}_{t-2}) \\ &\qquad + \cdots \\ &\qquad + \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \hat{\theta}_1, \hat{F}) - \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \theta_1^0, \hat{F}) \\ &\qquad + \nabla l_t(\hat{\varphi}_t) - \nabla l_t(\theta_t^0, \hat{\varphi}_{t-1}) \\ &\quad = B_{nt} + \nabla_{tt}^2 l_t(\theta_t^0, \hat{\varphi}_{t-1})(\hat{\theta}_t - \theta_t^0) \\ &\qquad + \nabla_{t(t-1)}^2 l_t(\theta_t^0, \theta_{t-1}^0, \hat{\varphi}_{t-2})(\hat{\theta}_{t-1} - \theta_{t-1}^0) \\ &\qquad + \cdots \\ &\qquad + \nabla_{t1}^2 l_t(\theta_t^0, \ldots, \theta_1^0, \hat{F})(\hat{\theta}_1 - \theta_1^0) + o_p(n^{-1/2}), \end{aligned}$$

where ∇_{tj}^2 is the second order derivative with respect to θ_t and θ_j, $j = 1, \ldots, t$. Similar to the case of $t = 1$, we can show that

$$B_{nt} = \frac{1}{n}\sum_{i=1}^{n} 2h_{1t}(X^{(i)}, \varphi_t^0) + o_p(n^{-1/2}),$$

where

$$\begin{aligned} &h_{1t}(X^{(i)}, \varphi_t^0) \\ &\quad = \frac{\pi_t}{2} E\left\{\frac{\nabla g_t(\vec{Y}_t^{(j)}, \varphi_t^0)\prod_{s=1}^{t} f_s(Y_s^{(j)}|V_{s-1}^{(j)}, \theta_s^0)}{[g_t(\vec{Y}_t^{(j)}, \varphi_t^0)]^2}\right. \\ &\qquad \left. - \frac{\nabla f_t(Y_t^{(j)}|V_{t-1}^{(j)}, \theta_t^0)\prod_{s=1}^{t-1} f_s(Y_s^{(j)}|V_{s-1}^{(j)}, \theta_s^0)}{g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0)}\right. \\ &\qquad \left.\,\middle|\, R_1^{(i)} = 1, X^{(i)}\right\}, \\ &g_t(\vec{Y}_t^{(j)}, \varphi_t^0) = \int\prod_{s=1}^{t} f_s(Y_s^{(j)}|x, \vec{Y}_{s-1}^{(j)}, \theta_s^0) dF^0, \end{aligned}$$

and

$$\begin{aligned} &\nabla g_t(\vec{Y}_t^{(j)}, \varphi_t^0) \\ &\quad = \int \nabla f_t(Y_t^{(j)}|x, \vec{Y}_{t-1}^{(j)}, \theta_t^0)\prod_{s=1}^{t-1} f_s(Y_s^{(j)}|x, \vec{Y}_{s-1}^{(j)}, \theta_s^0) dF^0. \end{aligned}$$

Under the given regularity conditions, $\nabla_{tj}^2 l_t(\theta_t^0, \ldots, \theta_j^0, \hat{\varphi}_{j-1}) - A_{tj} = o_p(1)$. Then

$$\sqrt{n}(\hat{\theta}_t - \theta_t^0)$$

$$= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_{tt}^{-1} \left\{ \frac{\partial H_t^{(i)}(\varphi_t^0)}{\partial \theta_t} + 2h_{1t}(X^{(i)}, \varphi_t^0) + \sum_{j=1}^{t-1} A_{tj} \psi_j(W_j^{(i)}, A_j, \varphi_j^0) \right\} + o_p(1)$$

and result (6) holds with the following iteratively defined ψ_t: ψ_1 is given by (9); having $\psi_1, \ldots, \psi_{t-1}$, ψ_t is defined as

(10)
$$\psi_t(W_t^{(i)}, A_t, \varphi_t^0) = -A_{tt}^{-1} \left\{ \frac{\partial H_t^{(i)}(\varphi_t^0)}{\partial \theta_t} + 2h_{1t}(X^{(i)}, \varphi_t^0) + \sum_{j=1}^{t-1} A_{tj} \psi_j(W_j^{(i)}, A_j, \varphi_j^0) \right\}.$$

The explicit forms of ψ_t, $t = 1, 2, 3, 4$, are shown as follows:

$$\psi_1(W_1^{(i)}, A_1, \varphi_1^0) = -A_{11}^{-1} \left\{ \frac{\partial H_1^{(i)}(\varphi_1^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\},$$

$$\psi_2(W_2^{(i)}, A_2, \varphi_2^0) = -A_{22}^{-1} \left\{ \frac{\partial H_2^{(i)}(\varphi_2^0)}{\partial \theta_2} + 2h_{12}(X^{(i)}, \varphi_2^0) \right\} + A_{22}^{-1} A_{21} A_{11}^{-1} \left\{ \frac{\partial H_1^{(i)}(\varphi_1^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\},$$

$$\psi_3(W_3^{(i)}, A_3, \varphi_3^0) = -A_{33}^{-1} \left\{ \frac{\partial H_3^{(i)}(\varphi_3^0)}{\partial \theta_3} + 2h_{13}(X^{(i)}, \varphi_3^0) \right\} + A_{33}^{-1} A_{32} A_{22}^{-1} \left\{ \frac{\partial H_2^{(i)}(\varphi_2^0)}{\partial \theta_2} + 2h_{12}(X^{(i)}, \varphi_2^0) \right\} + (-A_{33}^{-1} A_{32} A_{22}^{-1} A_{21} A_{11}^{-1} + A_{33}^{-1} A_{31} A_{11}^{-1}) \left\{ \frac{\partial H_1^{(i)}(\varphi_1^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\},$$

$$\psi_4(W_4^{(i)}, A_4, \varphi_4^0) = -A_{44}^{-1} \left\{ \frac{\partial H_4^{(i)}(\varphi_4^0)}{\partial \theta_4} + 2h_{14}(X^{(i)}, \varphi_4^0) \right\} + A_{44}^{-1} A_{43} A_{33}^{-1} \left\{ \frac{\partial H_3^{(i)}(\varphi_3^0)}{\partial \theta_3} + 2h_{13}(X^{(i)}, \varphi_3^0) \right\} + (-A_{44}^{-1} A_{43} A_{33}^{-1} A_{32} A_{22}^{-1} + A_{44}^{-1} A_{42} A_{22}^{-1}) \left\{ \frac{\partial H_2^{(i)}(\varphi_2^0)}{\partial \theta_2} + 2h_{12}(X^{(i)}, \varphi_2^0) \right\} + (-A_{44}^{-1} A_{43} A_{33}^{-1} A_{31} A_{11}^{-1} + A_{44}^{-1} A_{43} A_{33}^{-1} A_{32} A_{22}^{-1} A_{21} A_{11}^{-1} - A_{44}^{-1} A_{42} A_{22}^{-1} A_{21} A_{11}^{-1} + A_{44}^{-1} A_{41} A_{11}^{-1}) \left\{ \frac{\partial H_1^{(i)}(\varphi_1^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\}.$$

□

Proof of Theorem 3.2. Note that

$$\Sigma_t = \mathrm{Var}(D_t^{(i)}) = \int \psi_t(w, A_t, \varphi_t^0) \psi_t(w, A_t, \varphi_t^0)^\tau dP(w)$$

and

$$\hat{\Sigma}_t = \int \psi_t(w, \hat{A}_t, \hat{\varphi}_t) \psi_t(w, \hat{A}_t, \hat{\varphi}_t)^\tau dP_n(w),$$

where $P(w)$ denotes the underlying true distribution of W_t and $P_n(w)$ denotes its empirical distribution based on data $W_t^{(j)}, j = 1, \ldots, n$. Note that $\|\hat{\Sigma}_t - \Sigma_t\|$ is bounded by

$$\begin{aligned} &\left\| \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\psi}_t) \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\psi}_t)^\tau \right. \\ &\quad - \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, A_t, \psi_t^0) \psi_t(W_t^{(i)}, A_t, \psi_t^0)^\tau \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, A_t, \psi_t^0) \psi_t(W_t^{(i)}, A_t, \psi_t^0)^\tau \\ &\quad \left. - \int \psi_t(w, A_t, \psi_t^0) \psi_t(w, A_t, \psi_t^0)^\tau dP(w) \right\| \\ &\le \frac{1}{n} \sum_{i=1}^{n} \|Q_{n,t}\| \, I_{[0,c]}(\|W_t^{(i)}\|) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \|Q_{n,t}\| \, I_{(c,\infty)}(\|W_t^{(i)}\|) + o_p(1), \end{aligned}$$

where the inequality follows from the triangular inequality and law of large numbers, and

$$\begin{aligned} Q_{n,t} = {} & \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\psi}_t) \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\psi}_t)^\tau \\ & - \psi_t(W_t^{(i)}, A_t, \psi_t^0) \psi_t(W_t^{(i)}, A_t, \psi_t^0)^\tau. \end{aligned}$$

By condition (C3), for any $\epsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^{n} \|Q_{n,t}\| \, I_{[0,c]}(\|W_t^{(i)}\|) < \epsilon/2$$

when n is sufficiently large. For any $\tilde{\epsilon} > 0$, we can choose c such that $E[h(w) I_{(c,\infty)}(\|w\|)] < \epsilon\tilde{\epsilon}/4$. By Chebyshev's inequality and condition (C4), we have

$$P\left(\frac{1}{n} \sum_{i=1}^{n} \|Q_{n,t}\| \, I_{(c,\infty)}(\|W_t^{(i)}\|) > \epsilon/2 \right) < \tilde{\epsilon}.$$

Then,

$$P\left(\frac{1}{n} \sum_{i=1}^{n} \|Q_{n,t}\| > \epsilon \right) \to 0.$$

This proves that $\|\hat{\Sigma}_t - \Sigma_t\| = o_p(1)$. □

ACKNOWLEDGEMENTS

We would like to thank a referee and an associate editor for providing helpful comments and suggestions. The research was partially supported by the NSF Grant DMS-1007454.

Received 26 October 2011

REFERENCES

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43** 49–93.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90** 1112–1121. MR1354029

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York. MR1925014

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92** 1320–1329.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16** 285–319.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90** 106–121. MR1325118

Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90** 747–764. MR2024755

Troxel, A. B., Harrington, D. P. and Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics* **47** 425–438.

Troxel, A. B., Lipsitz, S. R. and Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotone missing data. *Biometrika* **85** 661–672. MR1665826

Jun Shao
School of Finance and Statistics
East China Normal University
Shanghai 200241
China

Department of Statistics
University of Wisconsin
Madison, WI 53706
USA
E-mail address: shao@stat.wisc.edu

Jiwei Zhao
Department of Statistics
University of Wisconsin
Madison, WI 53706
USA
E-mail address: zhaoj@stat.wisc.edu