Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 86

Estimation in Semiparametric Transition Measurement Error Models for Longitudinal Data

Wenqin Pan* Donglin Zeng†
Xihong Lin‡

*Duke University, wendy.pan@duke.edu
†University of North Carolina, dzeng@bios.unc.edu
‡Harvard School of Public Health, xlin@hsph.harvard.edu

Estimation in Semiparametric Transition Measurement Error Models for Longitudinal Data

Wenqin Pan[1*], Donglin Zeng[2], and Xihong Lin[3**]**

[1]Department of Biostatistics and Bioinformatics, Duke University

Durham, NC, 27705.

[2]Department of Biostatistics, University of North Carolina

Chapel Hill, NC, 27599.

[3]Department of Biostatistics, Harvard School of Public Health

Boston, MA 02115.

[*]*email: wendy.pan@duke.edu*

[**]*email: dzeng@bios.unc.edu*

[***]*email: xlin@hsph.harvard.edu*

SUMMARY. We consider semiparametric transition measurement error models for longitudinal data, where one of the covariates is measured with error in transition models, and no distributional assumption is made for the underlying unobserved covariate. An estimating equation approach based on the pseudo conditional score method is proposed. We show the resulting estimators of the regression coefficients are consistent and asymptotically normal. We also discuss the issue of efficiency loss. Simulation studies are conducted to examine the finite-sample performance of our estimators. The longitudinal AIDS Costs and Services Utilization Survey data are analyzed for illustration.

KEY WORDS: Asymptotic efficiency; Conditional score method; Functional modeling; Measurement error; Longitudinal data; Transition models.

1. Introduction

Longitudinal data are common in health science research, where repeated measures are obtained for each subject over time. One class of longitudinal models is the transitional model, where the conditional mean of an outcome at the current time point is modeled as a function of the past outcomes and covariates (Diggle et al., 2002, Chapter 10). This class of models is particularly useful when one is interested in predicting the future response given the past history, or when past history contains important adjustor variables. The within-subject correlation is automatically accounted for by conditioning on the past responses, and the model can be easily fit within the generalized linear model framework. Transition models and their wide practical applications have been well demonstrated (e.g., Young et al. 1999, Have and Morabia 2002, Heagerty 2002, Roy and Lin 2005).

Measurement error in covariate is a common problem in longitudinal data, due to equipment limitation, longitudinal variation, or recall bias. In one study from the AIDS Costs and Services Utilization Survey (ACSUS) (Berk, Maffeo and Schur 1993), which consisted of subjects from 10 randomly selected U.S. cities with the highest AIDS rates, a series of quarterly interviews were conducted for each participant enrolled between 1991 and 1992. A question of interest was to study how CD4 count predicted the risk of future hospitalization given a subject's past history of hospitalizations. Thus, a natural model for analyzing this data set is to fit a prediction model with the outcome being whether a participant had a hospital admission (yes/no) in the past quarter. However, CD4 count is known to be subject to considerable measurement error due to its substantial variability, e.g., its coefficient of variation within the same subject was found to be 50% (Tsiatis et al. 1995). Another source of

measurement error in CD4 count in this study was due to the fact that CD4 count was not measured at the time of each interview but abstracted from each respondent's most recent medical record.

The methods for handling measurement error for independent outcomes are comprehensively reviewed in Fuller (1987) and Carroll, et al. (2006). For longitudinal data, Wang, et al. (1998) among others considered measurement error in mixed effects models. Schmid, Segal and Rosner (1994) and Schmid (1996) studied measurement error in first-order autoregressive models for continuous longitudinal outcomes. There is a vast amount of work in the econometrics literature on panel data with errors in variables. For example, Griliches and Hausman (1986) and Biorn and Klette (1998) proposed estimating the effect of the error-prone covariates using the generalized moment method but their method required that longitudinal outcomes be linearly related to covariates and the residue terms be non-autocorrelated. In the literature of structural equations models, longitudinal covariates subject to measurement errors are treated as latent variables and are modelled longitudinally and explicitly (c.f. Duncan, Duncan and Strycker, 2006). The maximum likelihood estimation is used for inference. Additionally, using the same idea of latent modelling, Pan, Lin and Zeng (2006) considered estimation in generalized transitional measurement error models for general outcomes. However, these approaches require that the normality assumption and the correlation structure of the unobserved covariate be correctly specified. The normality assumption is often too strong in reality, and the correlation structure of the unobserved covariate may be difficult to be specified correctly. One can show that when a first-order autoregressive structure for the unobserved covariate is misspecified as an independent structure, the effect of this covariate in transition model is at-

tenuated and the effect of the past outcome is the same as the one ignoring the measurement error (Pan, 2002). Therefore, it is necessary to develop a method which leaves the distribution of the unobserved covariate fully unspecified. On the other hand, since the repeated measures of the unobserved covariate are usually correlated and have at least three waves, the attempt to estimate their joint distribution nonparametrically, for example, using the kernel method in Carroll and Wand (1991), breaks down due to the curse of dimensionality.

This paper aims to develop a semiparametric method for transition measurement error models without specifying the distribution of the unobserved covariate. Our approach is to construct an estimating equation based on the pseudo conditional score method, originally proposed for independent data by Stefanski and Carroll (1987). However, its generalization to transition models is not trivial in presence of repeatedly measured unobserved covariates. In the second part of this paper, we further discuss the efficiency issue in the proposed method.

The rest of the paper is structured as follows. In §2, we present the general form of the semiparametric transition measurement error model for longitudinal data. In §3, we derive the pseudo conditional score estimating equation and study the theoretical properties of the resulting estimator. In §4, we illustrate the method using simulation studies and apply the proposed method to analyze the ACSUS data. The issue of efficiency loss is also studied. Discussions are given in §5.

2. Semiparametric Transition Measurement Error Model

Suppose each of the n subjects has m repeated measures over time. Let Y_{ij} be the outcome at time j $(j = 1, \cdots, m)$ of subject i $(i = 1, \cdots, n)$. Let W_{ij} be a scalar observed error-prone covariate, which measures the unobserved covariate X_{ij} with error. Let $\mathbf{Z}_{ij}$ be a vector of covariates that are accurately measured. A transition model assumes the conditional distribution of Y_{ij} given the history of the outcome and the history of the covariates satisfies the (q, r)-order Markov property (Ch 10, Diggle et al., 2002) and belongs to the exponential family.

Specifically, for $j > s$, where $s = (r-1) \vee q = \max(r-1, q)$, the conditional distribution of Y_{ij} is

$$f(Y_{ij}|H_{ij}) = \exp\left\{(Y_{ij}\eta_{ij} - b(\eta_{ij}))/a\phi + c(H_{ij}, \phi)\right\}, \tag{1}$$

where $H_{ij} = \{Y_{i,j-1}, \cdots, Y_{i,j-q}, X_{ij}, ..., X_{i,j-r+1}, \mathbf{Z}_{ij}, ..., \mathbf{Z}_{i,j-r+1}\}$, $f(\cdot)$ denotes a density function, a is a prespecified weight, ϕ is a scale parameter, and $b(\cdot)$ and $c(\cdot)$ are specific functions associated with the exponential family. We assume a canonical generalized linear model (McCullagh and Nelder, 1989) for $\mu_{ij} = E(Y_{ij}|H_{ij}) = b'(\eta_{ij})$ as

$$h(\mu_{ij}) = \eta_{ij} = \beta_0 + \sum_{k=1}^{q} \alpha_k Y_{i,j-k} + \sum_{l=1}^{r} \{\beta_{xl} X_{i,j-l+1} + \boldsymbol{\beta}_{zl}^T \mathbf{Z}_{i,j-l+1}\}, \tag{2}$$

where $h(\cdot)$ is the canonical link function satisfying $h^{-1}(\cdot) = b'(\cdot)$, β_0, α_k $(k = 1, ..., q)$, $\boldsymbol{\beta}_l = (\beta_{xl}, \boldsymbol{\beta}_{zl}^T)^T$ $(l = 1, ..., r)$ are regression coefficients. In addition, we treat $Y_{i1}, ..., Y_{i,s}$ as initial states which the subsequent inference will be conditioned on. One note is that when $\mathbf{Z}$ covariates do not change with time, we tacitly keep only one $\mathbf{Z}$ term in equation (2).

We assume that X_{ij} is subject to measurement error and the measurement error is additive, i.e.,

$$W_{ij} = X_{ij} + U_{ij}, \tag{3}$$

where the measurement errors U_{ij} are independent of the X_{ij} and are independently and identically distributed from a normal distribution with a known variance σ_u^2. The variance σ_u^2 usually needs to be estimated beforehand, either from replications or from validation data (Carroll, et al, 2006). We assume that the joint distribution of $\{X_{i1}, \cdots, X_{im}\}$ is fully unspecified.

We suppose that measurement error is non-differential, i.e.,

$$f(Y_{ij}, W_{ij}|H_{ij}) = f(Y_{ij}|H_{ij})f(W_{ij}|X_{ij}),$$

where H_{ij} was defined in (1). This means that conditional on the true covariates, the observed error-prone covariate does not contain additional information about Y_{ij}.

3. Inference Procedures

3.1 Pseudo conditional score equation

Let $\boldsymbol{\theta}$ denote $(\beta_0, \alpha_1, \cdots, \alpha_q, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_r^T, \phi)^T$. In this section, we propose a pseudo conditional score method to estimate $\boldsymbol{\theta}$. The idea is to pretend $\boldsymbol{\theta}$ to be known but treat the X_{ij} as fixed parameters by writing X_{ij} as x_{ij}.

In a classical conditional score approach (Stefanski and Carroll, 1987), one would aim to derive simple sufficient summary statistics for $(x_{i1}, ..., x_{im})$ and construct an estimating equation based on the conditional likelihood function of the observed data given the sufficient statistics. Unfortunately, due to the transition structure and the possibly nonlinear link function in (2), obtaining

the summary sufficient statistics for x_{ij} based on the distribution of the observed data is usually difficult. For example, the likelihood function for a first-order transition model for dichotomous Y_2 and Y_3 with $X_1 = X_2 = X_3 = X$ given initial state Y_1 is

$$\exp\left\{Y_2(\beta_0 + \alpha_1 Y_1 + \beta_{x1} X) - \log(1 + e^{\beta_0 + \alpha_1 Y_1 + \beta_{x1} X}) \right.$$

$$\left. + Y_3(\beta_0 + \alpha_1 Y_2 + \beta_{x1} X) - \log(1 + e^{\beta_0 + \alpha_1 Y_2 + \beta_{x1} X})\right\}$$

and it does not belong to any exponential family.

Instead, we note that for each $j = s+1, ..., m$, the conditional density of $(Y_{ij}, W_{ij}, ..., W_{i,j-r+1})$ given $(Y_{i,j-1}, \cdots, Y_{i,j-q}, \mathbf{Z}_{ij}, ..., \mathbf{Z}_{i,j-r+1})$ and $(x_{ij}, x_{i,j-1}, ..., x_{i,j-r+1})$ is given by

$$\exp\left[Y_{ij}(\beta_0 + \sum_{k=1}^{q} \alpha_k Y_{i,j-k} + \sum_{l=1}^{r} \{\beta_{xl} x_{i,j-l+1} + \boldsymbol{\beta}_{zl}^T \mathbf{Z}_{i,j-l+1}\})/a\phi \right.$$
$$- b(\beta_0 + \sum_{k=1}^{q} \alpha_k Y_{i,j-k} + \sum_{l=1}^{r} \{\beta_{xl} x_{i,j-l+1} + \boldsymbol{\beta}_{zl}^T \mathbf{Z}_{i,j-l+1}\})/a\phi$$
$$+ c(Y_{i,j-1}, \cdots, Y_{i,j-q}, x_{ij}, ..., x_{i,j-r+1}, \mathbf{Z}_{ij}, ..., \mathbf{Z}_{i,j-r+1}, \phi)$$
$$\left. - \sum_{l=1}^{r} (W_{i,j-l+1} - x_{i,j-l+1})^2/2\sigma_u^2 - r \log\sqrt{2\pi\sigma_u^2} \right].$$

We recognize that this conditional density still belongs to an exponential family. The sufficient statistics for $x_{i,j-l+1}, l = 1, ..., r$, are

$$T_{i1}^{(j)} = \frac{\beta_{x1}}{a\phi} Y_{ij} + \frac{1}{\sigma_u^2} W_{ij}, \;\; T_{i2}^{(j)} = \frac{\beta_{x2}}{a\phi} Y_{ij} + \frac{1}{\sigma_u^2} W_{i,j-1}, \;\; ..., \;\; T_{ir}^{(j)} = \frac{\beta_{xr}}{a\phi} Y_{ij} + \frac{1}{\sigma_u^2} W_{i,j-r+1}. \tag{4}$$

Therefore, the distribution of Y_{ij} given $(Y_{i,j-1}, ..., Y_{i,j-q}, \mathbf{Z}_{ij}, ..., \mathbf{Z}_{i,j-r+1})$ and $(T_{i1}^{(j)}, ..., T_{ir}^{(j)})$ only depends on $\boldsymbol{\theta}$ but not $(x_{ij}, ..., x_{i,j-r+1})$. For convenience, we abbreviate this distribution as $\widetilde{f}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}); \boldsymbol{\theta})$, where $\mathbf{V}_{ij}(\boldsymbol{\theta})$ denotes the statistics that Y_{ij} are conditioned on. Clearly,

$$E_{\boldsymbol{\theta}_0}\left[\nabla_{\boldsymbol{\theta}} \log \widetilde{f}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0); \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right]$$

$$= E_{\boldsymbol{\theta}_0}\left[E_{\boldsymbol{\theta}_0}\left\{\nabla_{\boldsymbol{\theta}}\log\widetilde{f}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0);\boldsymbol{\theta})|\mathbf{V}_{ij}(\boldsymbol{\theta}_0)\right\}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] = 0$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$, $E_{\boldsymbol{\theta}}$ denotes the expectation given the parameter $\boldsymbol{\theta}$, and $\nabla_{\boldsymbol{\theta}}$ denotes the gradient with respect to $\boldsymbol{\theta}$. We then construct the following estimating equation

$$\sum_{i=1}^{n}\sum_{j=s+1}^{m}\mathbf{g}\left(Y_{ij}|\mathbf{v}_{ij}=\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta}\right)=0, \tag{5}$$

where $\mathbf{g}(y_{ij}|\mathbf{v}_{ij};\boldsymbol{\theta})$ denotes the gradient of $\log\widetilde{f}(y_{ij}|\mathbf{v}_{ij};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Note that calculations of this gradient are done by viewing $\mathbf{v}_{ij}$ as fixed instead of a function of $\boldsymbol{\theta}$.

Essentially, our idea is to construct some conditional score functions based on the conditional density given the past history at each time then take the summation of all these scores as estimating function. Since the above construction is no based on the full likelihood function, we call our proposed estimating equation the *pseudo conditional score equation*. The Newton-Raphson iteration can be used to solve the equation; however, multiple solutions may exist. Thus, the following theorem gives the asymptotic property of a solution to (5) in a neighborhood of $\boldsymbol{\theta}_0$.

Theorem 1. *Assume that with probability one, in a neighborhood of* $\boldsymbol{\theta}_0$, $\nabla_{\boldsymbol{\theta}}\mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})$ *is Lipschitz continuous with respect to* $\boldsymbol{\theta}$ *and moreover,*

$$E_{\boldsymbol{\theta}_0}\left[\sum_{j=s+1}^{m}\nabla_{\boldsymbol{\theta}}\mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \text{ is non-singular.}$$

Then there exists a solution, $\widehat{\boldsymbol{\theta}}_n$, *to equation (5) such that* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)$ *converges in distribution to a normal distribution with mean zero and covariance*

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) &= \left\{E_{\boldsymbol{\theta}_0}\left[\sum_{j=s+1}^{m}\nabla_{\boldsymbol{\theta}}\mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]\right\}^{-1}\\ &\quad\times E_{\boldsymbol{\theta}_0}\left[\left\{\sum_{j=s+1}^{m}\mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)\right\}\left\{\sum_{j=s+1}^{m}\mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)\right\}^{T}\right]\end{aligned}$$

$$\times \left\{ E_{\boldsymbol{\theta}_0} \left[\sum_{j=s+1}^{m} \nabla_{\boldsymbol{\theta}} \mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})^T \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right] \right\}^{-1}.$$

The proof follows the usual argument for estimating equations. Clearly, a consistent estimator for $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}_n &= n \left[\sum_{i=1}^{n} \sum_{j=s+1}^{m} \nabla_{\boldsymbol{\theta}} \mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right]^{-1} \\
&\quad \times \left[\sum_{i=1}^{n} \left\{ \sum_{j=s+1}^{m} \mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\widehat{\boldsymbol{\theta}}_n);\widehat{\boldsymbol{\theta}}_n) \right\} \left\{ \sum_{j=s+1}^{m} \mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\widehat{\boldsymbol{\theta}}_n);\widehat{\boldsymbol{\theta}}_n) \right\}^T \right] \\
&\quad \times \left[\sum_{i=1}^{n} \sum_{j=s+1}^{m} \nabla_{\boldsymbol{\theta}} \mathbf{g}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})^T \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \right]^{-1}.
\end{aligned}$$

3.2 Examples

We illustrate our method using two examples.

Example 1. We consider a linear transition model with $r = 1$ and $q = 1$:

$$Y_{ij} = \beta_0 + \alpha Y_{i,j-1} + \beta_x X_{ij} + \boldsymbol{\beta}_z^T \mathbf{Z}_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_y^2), \quad j = 2, ..., m. \quad (6)$$

Then it is easy to calculate that the sufficient statistic for x_{ij} is $T_{i1}^{(j)} = \beta_x Y_{ij}/\sigma_y^2 + W_{ij}/\sigma_u^2$ and $\widetilde{f}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})$ is the conditional density of Y_{ij} given $T_{i1}^{(j)}$, $Y_{i,j-1}$ and $\mathbf{Z}_{ij}$. This density is the same as the conditional density of Y_{ij} given $Q_{ij} = \beta_x(Y_{ij} - \beta_0 - \alpha Y_{i,j-1} - \boldsymbol{\beta}_z^T \mathbf{Z}_{ij})/\sigma_y^2 + W_{ij}/\sigma_u^2$, $Y_{i,j-1}$ and $\mathbf{Z}_{ij}$, whose logarithm is equal to

$$-\log\sqrt{2\pi\sigma_y^{*2}} - (2\sigma_y^{*2})^{-1}(Y_{ij} - \beta_0 - \alpha Y_{i,j-1} - \boldsymbol{\beta}_z^T \mathbf{Z}_{ij} - Q_{ij}\beta_x^*)^2, \quad j = 2, ..., m,$$

where $\beta_x^* = \beta_x/(\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)$ and $\sigma_y^{*2} = (\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)^{-1}\sigma_y^2/\sigma_u^2$. Differentiating the above function with respect to all the parameters then substituting the expression of Q_{ij}, we obtain the following pseudo conditional score equations

$$0 = \sum_{i=1}^{n} \sum_{j=2}^{m} \begin{pmatrix} 1 \\ Y_{i,j-1} \\ \mathbf{Z}_{ij} \end{pmatrix} \left\{ Y_{ij} - \beta_0 - \alpha Y_{i,j-1} - \boldsymbol{\beta}_z^T \mathbf{Z}_{ij} - \beta_x W_{ij} \right\},$$

$$
\begin{aligned}
0 &= \sum_{i=1}^{n}\sum_{j=2}^{m}\left\{(Y_{ij}-\beta_0-\alpha Y_{i,j-1}-\boldsymbol{\beta}_z^T\mathbf{Z}_{ij})\beta_x + W_{ij}\sigma_y^2/\sigma_u^2\right\} \\
&\quad \times(Y_{ij}-\beta_0-\alpha Y_{i,j-1}-\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}-\beta_x W_{ij}), \\
0 &= \sum_{i=1}^{n}\sum_{j=2}^{m}\left\{(Y_{ij}-\beta_0-\alpha Y_{i,j-1}-\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}-\beta_x W_{ij})^2-(\beta_x^2\sigma_u^2+\sigma_y^2)\right\}.
\end{aligned}
$$

Clearly, each term for i and j is the conditional score obtained for subject i at time j given the past history. Moreover, the first equation correspond to parameters $(\beta_0, \alpha, \boldsymbol{\beta}_z^T)$, the second equation corresponds to β_x, and the last equation is for σ_y^2.

Example 2. In this example, we consider a logistic transition model with $r = q = 1$, where Y_{ij} is a Bernoulli variable and satisfies

$$
\text{logit}P(Y_{ij}|H_{ij}) = \beta_0 + \alpha Y_{i,j-1} + \beta_x X_{ij} + \boldsymbol{\beta}_z^T\mathbf{Z}_{ij}. \tag{7}
$$

We can easily calculate that the sufficient statistic for x_{ij} is $T_{i1}^{(j)} = \beta_x Y_{ij} + W_{ij}/\sigma_u^2$ and that the logarithm of the conditional density $\widetilde{f}(Y_{ij}|T_{i1}^{(j)}, Y_{i,j-1}, \mathbf{Z}_{ij}; \boldsymbol{\theta})$ is

$$
\begin{aligned}
&-\frac{(T_{i1}^{(j)}-Y_{ij}\beta_x)^2\sigma_u^2}{2} + Y_{ij}(\beta_0+\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}+\alpha Y_{i,j-1}) \\
&-\log\left[\exp\left\{-\frac{(T_{i1}^{(j)}-\beta_x)^2\sigma_u^2}{2}+(\beta_0+\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}+\alpha Y_{i,j-1})\right\}+\exp\left\{-\frac{{T_{i1}^{(j)}}^2\sigma_u^2}{2}\right\}\right].
\end{aligned}
$$

After differentiating the above function with respect to all the parameters then substituting the expression of $T_{i1}^{(j)}$, we obtain the following pseudo conditional score equations

$$
\begin{aligned}
0 &= \sum_{i=1}^{n}\sum_{j=2}^{m}\begin{pmatrix}1\\ Y_{i,j-1}\\ \mathbf{Z}_{ij}\end{pmatrix} \\
&\quad \times\left[Y_{ij}-\frac{1}{1+\exp\left\{(1/2-Y_{ij})\beta_x^2\sigma_u^2-\beta_x W_{ij}-(\beta_0+\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}+\alpha Y_{i,j-1})\right\}}\right], \\
0 &= \sum_{i=1}^{n}\sum_{j=2}^{m}\left[Y_{ij}W_{ij}-\frac{(Y_{ij}\beta_x+W_{ij}/\sigma_u^2-\beta_x)\sigma_u^2}{1+\exp\left\{(1/2-Y_{ij})\beta_x^2\sigma_u^2-W_{ij}\beta_x-(\beta_0+\boldsymbol{\beta}_z^T\mathbf{Z}_{ij}+\alpha Y_{i,j-1})\right\}}\right].
\end{aligned}
$$

3.3 Method for selecting transition orders

In practice, the transition orders (q, r) in the Y model (2) are often unknown. As our model is a semiparametric model, a full likelihood does not exist. Hence standard model selection methods are not directly applicable. We propose to choose (q, r) based on the pseudo log-likelihood function

$$\sum_{i=1}^{n} \ln \widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0); \boldsymbol{\theta}),$$

where $\widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0); \boldsymbol{\theta})$ is defined right after equation (4), i.e.,

$$\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0) = \left(Y_{i,m-1}, ..., Y_{i,m-q}, \mathbf{Z}_{im}, ..., \mathbf{Z}_{i,m-r+1}, T_{i1}^{(m)}(\boldsymbol{\theta}_0), ..., T_{ir}^{(m)}(\boldsymbol{\theta}_0)\right).$$

Here $\boldsymbol{\theta}_0$ denotes the parameter value under the true model and $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0)$ is $\mathbf{V}_{im}(\cdot)$ evaluated at the true value $\boldsymbol{\theta}_0$ under the model with transition orders (q, r).

The function $\widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0); \boldsymbol{\theta}_0)$ is the true density when (q, r) is equal to the true transition orders. Therefore, we are able to transform the selection of (q, r) in the original model (2) to the model selection in the new regression model given by $\widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0); \boldsymbol{\theta})$. Note that using $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0)$ instead of $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta})$ in the new model ensures that the covariate values do not vary with different (q, r). However, since $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0)$ is unknown, we propose to estimate $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0)$ at $\mathbf{V}_{im}^{(q,r)}(\widehat{\boldsymbol{\theta}}_F)$, where $\widehat{\boldsymbol{\theta}}_F$ is the parameter estimator under the full model with $q = m - 1$ and $r = m$ using the conditional score approach. Since $\widehat{\boldsymbol{\theta}}_F$ is consistent, $\mathbf{V}_{im}^{(q,r)}(\widehat{\boldsymbol{\theta}}_F)$ is a good approximation of $\mathbf{V}_{im}^{(q,r)}(\boldsymbol{\theta}_0)$.

Finally, we treat the pseudo log-likelihood function $\sum_{i=1}^{n} \ln \widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\widehat{\boldsymbol{\theta}}_F); \boldsymbol{\theta})$ like a "likelihood," and select (q, r) by minimizing the pseudo Akaike information criterion (P_AIC) defined as

$$-\sum_{i=1}^{n} 2 \ln \widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\widehat{\boldsymbol{\theta}}_F); \widehat{\boldsymbol{\theta}}) + 2\text{Card}(\widehat{\boldsymbol{\theta}}),$$

or the pseudo Bayesian information criterion (P_BIC) defined as

$$-\sum_{i=1}^{n} 2\ln \widetilde{f}(Y_{im}|\mathbf{V}_{im}^{(q,r)}(\widehat{\boldsymbol{\theta}}_F);\widehat{\boldsymbol{\theta}}) + \text{Card}(\widehat{\boldsymbol{\theta}})\log n,$$

where $\widehat{\boldsymbol{\theta}}$ is the estimate maximizing the pseudo likelihood function and $\text{Card}(\widehat{\boldsymbol{\theta}})$ denotes the number of parameters in the model.

The proposed method has been demonstrated to perform well in our numerical studies. However, it is not fully theoretically justified.

4. Numerical Results

4.1 Simulation studies

Corresponding to the two examples illustrated in the previous section, two simulation studies are conducted to examine the finite-sample performance of the proposed pseudo conditional score approach. Specifically, in the first simulation study, the longitudinal response Y_{ij} is generated from

$$Y_{ij} = -1+0.4Y_{i,j-1}+3X_{ij}+0.8Z_i+\epsilon_{ij},\ \ \epsilon_{ij} \sim N(0,1),\ \ i=1,...,n,\ \ j=2,...,m,$$

where Z_i is a Bernoulli variable with $P(Z_i = 1) = 0.5$ and X_{ij} follows the first order transition model

$$X_{ij} = 0.5 + 0.8X_{i,j-1} + \epsilon_{xij},\quad \epsilon_{xij} \sim N(0,1),\quad i=1,...,n,\quad j=2,...,m. \quad (8)$$

Here we assume the number of repeated measures per subject $m = 6$. We use $X_{i1} = 0.25$ and $Y_{i1} = -5/12 + 5Z_i/3$ as values at time one. The measurement error distribution in (3) has a variance 0.5. In the second simulation study, we generate binary responses from a logistic transition model with mean

$$\frac{\exp\{-1+0.5Y_{i,j-1}+X_{ij}+0.8Z_i\}}{1+\exp\{-1+0.5Y_{i,j-1}+X_{ij}+0.8Z_i\}},\quad i=1,...,n,\quad j=2,...,m,$$

where Z_i is generated from a Bernoulli distribution with $P(Z_i = 1) = 0.5$ and X_{ij} follows

$$X_{ij} = 0.4 + 0.5Z_i + 0.6X_{i,j-1} + \epsilon_{xij}, \;\; \epsilon_{xij} \sim N(0, 0.5), \;\; i = 1, ..., n, \;\; j = 2, ..., m.$$

The measure error variance is set to be 0.5. The initial states are given as $X_{i1} = 0.25$ and Y_{i1} from the Bernoulli distribution with probability 0.5. In both simulation studies, we solve the pseudo conditional score equations as given in Examples 1 and 2 to obtain the estimators and their asymptotic variances are estimated using the formula of $\widehat{\mathbf{\Sigma}}_n$. Table 1 summarizes the results from 1000 repetitions with sample sizes $n = 100$ or 200. The results show that in finite samples, the pseudo conditional score estimators have virtually no bias and the estimated standard errors agree well with the true standard errors.

We next conduct a simulation study to compare the robustness of the semiparametric pseudo conditional score method with the parametric maximum likelihood method as given in Pan et al. (2006) when the X model is misspecified. We use the same setting as in the first simulation study with $m = 6$. We consider three distribution scenarios for the X: (a) X_{ij} follows the first order transition model (8) with error ϵ_{xij} following a normal mixture, $0.5N(-0.5, 1) + 0.5N(0.5, 1)$; (b) X_{ij} follows the first order transition model (8) with error ϵ_{xij} following the extreme-value distribution; (c) X_{ij} follows a second-order transition model $X_{ij} = 0.5 + 0.8X_{i,j-2} + N(0, 1)$. For all three scenarios, the parametric maximum likelihood estimation(MLE) method treats X_{ij} from a first-order transition model with normal error distribution. We hence expect that the parametric MLE method would be biased because it misspecifies either the transition pattern or the error distribution.

Table 2 summarizes the robustness simulation results from 1000 repetitions with $n = 100$ and 200. The results show that the parametric MLE approach

gives biased estimates of the regression coefficients, especially α. The bias ranges from 3% to 10%. When the error distribution in the X model deviates slightly from normality as a normal mixture, the bias is small but the coverage probability can be poor. However, when the transition order in the X model is misspecified, the bias is more pronounced and is close to 10%, and the coverage probability becomes very poor. On the contrary, the pseudo conditional score approach always yields small bias and accurate coverage.

To evaluate the method in selecting transition orders as proposed in Section 3.3, we conduct another simulation study with dichotomous outcome. The setting is similar to our second simulation study except that the mean probability is

$$\frac{\exp\{-1 + Y_{i,j-2} + X_{i,j-1} + 0.8Z_i\}}{1 + \exp\{-1 + Y_{i,j-2} + X_{i,j-1} + 0.8Z_i\}}, \quad i = 1, ..., n, \quad j = 3, 4.$$

That is, the true transition order is $q = r = 2$. We apply the proposed method to fit models for all possible combinations of transition orders (q, r) with $q = 1, 2, 3$ and $r = 1, 2, 3, 4$. The pseudo AIC and the pseudo BIC are used for selecting the final orders. The result from 1000 repetitions with sample sizes 200 and 400 is given in Table 3. The result shows that the proposed method works well. Overall, the pseudo BIC outperforms the pseudo AIC, especially when sample size is large.

4.2 Numerical study on efficiency loss

The pseudo conditional score equation approach relies on the conditional likelihood function, so it does not utilize the full data information. Hence it may not give the efficient estimators. It is useful to know how much efficiency is lost when using such an approach. Since deriving the asymptotic efficiency bound for model (1) is generally difficult, we focus our discussion on the situation

where Y_{ij} is a normal outcome and $r = 1$ and $q = 1$ as in (6). Furthermore, we assume Z_{ij} and X_{ij} are independent but allow the repeated measures of X_{ij} to be correlated.

From Example 1, we have known that the $Q_{ij} = \beta_x(Y_{ij} - \beta_0 - \alpha Y_{i,j-1} - \boldsymbol{\beta}_z^T \mathbf{Z}_{ij})/\sigma_y^2 + W_{ij}/\sigma_u^2, j = 2, ..., m$ are sufficient statistics for $x_{ij}, j = 2, ..., m$. In fact, they are also complete and sufficient statistics. Therefore, following Bickel et al. (1993, Chap 4, pp.130), one can explicitly calculate the semiparametric efficiency bound (see the appendix). Thus, the efficiency loss in the pseudo conditional score estimator can be evaluated by comparing such efficiency bound versus $\boldsymbol{\Sigma}$ as given in Theorem 1.

We utilize a concrete example to illustrate the efficiency loss. Suppose that (Y_{ij}, W_{ij}) follows

$$Y_{ij} = -1 + 0.5Y_{i,j-1} + X_{ij} + 0.6Z_i + N(0, 2),$$

$$W_{ij} = X_{ij} + N(0, 0.5),$$

where Z_i is a Bernoulli variable with $P(Z_i = 1) = 0.5$ and X_{ij} is generated from the following transition model

$$X_{ij} = 0.4 + 0.5X_{i,j-1} + N(0, \sigma_x^2).$$

For different choices of $\sigma_x^2 = 0.3$ or 0.15 and different cluster sizes $m = 3$ or 4, we compute the asymptotic efficiency of the pseudo conditional score estimators for β_x, β_z, α relative to the semiparametric efficient bound. The results show that the efficiency loss increases with the decrease of σ_x^2; it varies from 10% to 20% in estimating β_x and α as m increases from 3 to 4; however, no efficiency is lost in estimating β_z.

4.3 Application to the ACSUS data

We apply our method to analyze the ACSUS data. Specifically, we restricted our attention to 533 patients who completed the first year interview. The participants were interviewed every 3 months for four times. The outcome was whether they had hospital admissions (yes/no) during the three months between two consecutive interviews. It is of scientific interest to study the effect of CD4 counts in predicting future hospitalization given the past history of hospitalization. As discussed in the introduction, CD4 counts were subject to considerable measurement error. Thus, a natural model for analyzing this data set is a prediction model by accounting for measurement errors in CD4 counts. A logistic transition model is used to fit the data with covariate $W = \log(CD4/100)$, a transformed variable that reduces the marked skewness of CD4 counts (Figure 1). We note that even after a log-transformation, the commonly used transformation for CD4 counts, CD4 counts still do not look normally distributed. This motivates us to leave the distribution of the true CD4 counts fully unspecified by considering the pseudo conditional score method. Other covariates include age (10 categories coded as 1-10), antiretroviral drug use, HIV-symptomatic at baseline, race, and gender. Additionally, the past hospitalization history is also adjusted for in the analysis. The size of the measurement error for W, $\sigma_u^2 = 0.38$, is set to be 1/3 of the variance of baseline W. This value is close to the estimated measurement error variance 0.39 by Wulfsohn and Tsiatis (1997), using data from another AIDS study conducted by Burroughs-Wellcome. In addition, we also fit model using $\sigma_u^2 = 0.18$ to obtain parameter estimates under a more conservative measurement error setting.

To select the best transition order (q, r), we apply the pseudo BIC method

proposed in Section 3.3. The result shows that $q = 1$ and $r = 1$ give the smallest value under the pseudo BIC criterion. This finding agrees with the result obtained from testing the significance of the extra terms when the highest order transitional model is fit: specifically, we fit the largest transition model with $q = 3$ and $r = 4$ and test for the significance of the higher than first-order terms, and we find they are highly insignificant. Hence our final model has the transition order $q = 1$ and $r = 1$. The parameter estimation result is given in Table 4, where the reported estimates are the estimated log-odds ratios of the covariates. Women have a significantly higher risk of future hospitalization than men. The effect of CD4 counts on the risk of future hospitalization is significant, given the previous hospitalization status. Subjects who had a previous hospital admission history and who had lower CD4 counts would be more likely to be hospitalized in the future.

We also fit the model by letting the measurement error σ_u^2 be 0.18, which corresponds to the situation when the coefficient of variation for the baseline W is 50%. The findings are similar but the estimated effect of W is slightly attenuated. We also present in Table 4 the naive estimators that are obtained by ignoring measurement error. The naive estimator of the CD4 count effect tends to bias towards zero.

5. Discussion

We consider in this paper transition measurement error models for longitudinal data. We propose a pseudo conditional score approach that does not require specifying the distribution of the unobserved covariate. Both numerical calculations and simulation studies show that the estimator using the pseudo conditional score method performs well.

The approach extends the classical conditional score approach in Stephanski and Carroll (1987) in the following aspects. First, the classical conditional score approach relies on extracting sufficient statistics for the error-prone covariates in the full likelihood function so is impossible for the transition models; instead, our approach works on the conditional likelihood at each time point. Second, because the conditional scores from different timepoints are correlated, a sandwich variance estimator must be used for inference. Third, one specific question to the transition model is how to choose the transition orders and we have provided an innovative way for this purpose based on the pseudo-likelihood function. Furthermore, we note that the proposed approach is always applicable to the situations when $X_{i,j-k}$ enters expression (2) linearly no matter how $Y_{i,j-k}$ or its transformed value enters expression (2). Therefore, our approach can also be used for other transition models such as the ones proposed for count data in Diggle et al. (2002). Assigning different weights to the conditional scores from different timepoints might improve efficiency, but we have not as yet explored this refinement.

One important issue in fitting a transition model is the selection of transition orders (q, r). If one is willing to assume a parametric model for X, (q, r) can be selected using various model selection criteria, such as AIC and BIC. However, under the semiparametric model considered in this paper, there does not exist any literature on choosing q and r. In this paper, we propose to select (q, r) using the pseudo likelihood function and a small simulation study indicates that the method works pretty well. Theoretical justification of the proposed method needs more work.

Another important issue is to determine the size of measurement error, σ_u^2, which can be estimated using replication or validation data. In this case,

Theorem 1 needs to be slightly modified to account for the variability due to estimating σ_u^2. Particularly, following the same proof for Theorem 1, we obtain that the asymptotic variance equals the variance of

$$E_{\boldsymbol{\theta}_0}\left[\sum_{j=s+1}^{m}\nabla_{\boldsymbol{\theta}} g_{\sigma_{0u}^2}(Y_{ij}|V_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})\,\Big|_{\theta=\theta_0}\right]^{-1}\sum_{j=s+1}^{m} g_{\sigma_{0u}^2}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)+$$

$$E_{\boldsymbol{\theta}_0}\left[\sum_{j=s+1}^{m}\nabla_{\boldsymbol{\theta}} g_{\sigma_u^2}(Y_{ij}|V_{ij}(\boldsymbol{\theta});\boldsymbol{\theta})\,\Big|_{\theta=\theta_0}\right]^{-1}\frac{\partial}{\partial\sigma_u^2}\Big|_{\sigma_u^2=\sigma_{0u}^2} E_{\boldsymbol{\theta}_0}\left[\sum_{j=s+1}^{m} g_{\sigma_u^2}(Y_{ij}|\mathbf{V}_{ij}(\boldsymbol{\theta}_0);\boldsymbol{\theta}_0)\right] S_{\sigma_{0u}^2},$$

where $g_{\sigma_u^2}(Y_{ij}|\cdot)$ is the same as defined in Theorem 1 but indexed by σ_u^2, σ_{0u}^2 denotes the true value of σ_u^2, and $S_{\sigma_{0u}^2}$ is the influence from estimating σ_{0u}^2 using the validation sample. Clearly, the second part of the above expression reveals the influence on estimating θ_0 when σ_{0u}^2 is estimated. When neither validation data nor replications are available, one possible strategy is to conduct sensitivity analysis (e.g., Li and Lin, 2000) by varying the sizes of measurement error in a reasonable range.

ACKNOWLEDGMENT

Lin's research was supported by National Cancer Institute grant CA–76404 and National Heart, Lung and Blood Institute grant HL–58611.

REFERENCES

Berk, M. L., Maffeo, C., and Schur, C. L. (1993). *Research Design and Analysis Objectives*. AIDS Cost and Services Utilization Survey Report No. 1, Rockville, MD: Agency for Health Care Policy and Research.

Bickel, P. J., Klaassen, C. A. I., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semi-parametric Models*. John Hopkins University Press.

Biorn, E., and Klette, T. J. (1998). Panel data with errors-in-variables: essential and redundant orthogonality conditions in GMM-estimation. *Economics Letters* **59**, 275-282.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C (2006). *Measurement Error in Nonlinear Models*. Second edition. London: Chapman and Hall.

Carroll, R. J., and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of Royal Statistical Society B* **53**, 573-385.

Cook, J. R., and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models, *Journal of American Statistical Association* **89**, 1314-1328.

Diggle, P. J., Liang, K., Heagerty, P., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford Statistical Science.

Duncan, T.E., Duncan, S. C., and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modelling*. Routledge.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Griliches, Z., and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics* **31**, 93-118.

Have, T. R. and Morabia, A. (2002). An assessment of non-randomized medical treatment of long-term schizophrenia relapse using bivariate binary-response transition models. *Biostatistics* **3**, 119-131.

Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58**, 342-351.

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

Li, Y. and Lin, X. (2000). Covariate measurement errors in frailty models for clustered survival data. *Biometrika* **87**, 849-866.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Pan, W. (2002). *Transition Measurement Error Models for Longitudinal Data*. Ph.D. Dissertation, University of Michigan.

Pan, W., Lin, X. and Zeng, D. (2006). Structural inference in transition measurement error models for longitudinal data. *Biometrics* **62**, 402-412.

Roy, J. and Lin, X. (2005). Missing covariates in longitudinal data with informative dropouts: Bias analysis and inference. *Biometrics* **61**, 837-846

Schmid, C. H. (1996). An EM algorithm fitting first-Order conditional autoregressive models to longitudinal data. *Journal of American Statistical Association* **91**, 1322-1330.

Schmid, C. H., Segal, M. R., and Rosner B. (1994). Incorporating measurement error in the estimation of autoregressive models for longitudinal data. *Journal of Statistical Planning and Inference* **42**, 1-18.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariates misclassification and measurement error in main study/validation study design. *Journal of American Statis-*

tical Association **95**, 51-61.

Stefanski, L. A., and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703-716.

Tsiatis, A. A., De Gruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data Measured with error applications to survival and CD4 counts in patients with AIDS. *Journal of American Statistical Association* **90**, 27-37.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of American Statistical Association* **93**, 249-261.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330-339.

Young, P. J., Weeden, S. and Kirwan, J. R. (1999). The analysis of a bivariate multi-state Markov transition model for rheumatoid arthritis with an incomplete disease history. *Statistics in Medicine* **18**, 1677-1690.

APPENDIX: Calculation of semiparametric efficiency bound in (6)

From Bickel et al. (1993), the semiparametric efficiency bound in (6) is given by $\mathbf{\Sigma}_e = \{E[\dot{l}^*_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\theta}, G)^{\otimes 2}]\}^{-1}$, where $a^{\otimes 2} = aa^T$ and

$$\dot{\ell}^*_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\theta}, G) = E[\dot{\ell}^c_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i, \mathbf{X}_i; \boldsymbol{\theta})|\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i] - E[\dot{\ell}^c_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\theta})|\mathbf{Q}_i, \mathbf{Z}_i].$$

Here, $\mathbf{Y}_i = (Y_{i2}, ..., Y_{im})^T$, $\mathbf{W}_i = (W_{i2}, ..., W_{im})^T$, $\mathbf{Z}_i = (\mathbf{Z}_{i2}, ..., \mathbf{Z}_{im})^T$, $\mathbf{X}_i = (X_{i2}, ..., X_{im})^T$, $\mathbf{Q}_i = (Q_{i2}, ..., Q_{im})^T$, $\dot{\ell}^c_{\boldsymbol{\theta}}$ is the score function for $\boldsymbol{\theta}$ with the

complete data $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, and $G(\cdot)$ denotes the joint distribution of $\mathbf{X}_i$. Particularly, direct calculations give $\dot{\ell}^*_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_i; \boldsymbol{\theta}, G)$ equal to

$$\frac{1}{\sigma_y^2}\sum_{j=2}^{m}\begin{pmatrix} \widetilde{\epsilon}_{ij} - E[\widetilde{\epsilon}_{ij}|Q_{ij}] \\ \mathbf{Z}_{ij}(\widetilde{\epsilon}_{ij} - E[\widetilde{\epsilon}_{ij}|Q_{ij}]) \\ Y_{i,j-1}\widetilde{\epsilon}_{ij} - E[Y_{j-1}\widetilde{\epsilon}_{ij}|Q_{ij}] - \beta_x(Y_{i,j-1} - E[Y_{i,j-1}|Q_{ij}])E[X_{ij}|Q_{ij}] \\ (\widetilde{\epsilon}_{ij} - E[\widetilde{\epsilon}_{ij}|Q_{ij}])E[X_{ij}|Q_{ij}] \\ (\widetilde{\epsilon}_{ij}^2 - E[\widetilde{\epsilon}_{ij}^2|Q_{ij}] - 2\beta_x(\widetilde{\epsilon}_{ij} - E[\widetilde{\epsilon}_{ij}|Q_{ij}])E[X_{ij}|Q_{ij}])/(2\sigma_y^2) \end{pmatrix},$$

where $\widetilde{\epsilon}_{ij} = Y_{ij} - \beta_0 - \alpha Y_{i,j-1} - \boldsymbol{\beta}_z^T \mathbf{Z}_{ij}$.

For specific example, the above semiparametric efficiency bound can be calculated explicitly in terms of the first two moments of $E[\mathbf{Y}_i|\mathbf{Q}_i]$, $E[\mathbf{X}_i|\mathbf{Q}_i]$, $E[\widetilde{\epsilon}_{ij}|\mathbf{Q}_i]$. For example, assume

(M.1) $(\mathbf{Y}_i, \mathbf{W}_i)$ follows $Y_{ij} = \beta_0 + \beta_z Z_{ij} + \beta_x X_{ij} + \alpha Y_{i,j-1} + \epsilon_{ij}$, $W_{ij} = X_{ij} + U_{ij}$;

(M.2). $\mathbf{X}$ is generated from the transition model $X_{ij} = \gamma_0 + \gamma_x X_{i,j-1} + \epsilon_{xij}$;

(M.3) $Z_{ij} = ... = Z_{i1}$ has mean m_z and variance v_z and it is independent of $\mathbf{X}_i$;

(M.4) Y_{i1} has mean m_y and variance v_y and X_{i1} has mean m_x and variance v_x;

(M.5) $(\epsilon_{ij}, U_{ij}, \epsilon_{xij})$ are independently from normal distribution with mean zero and variance $\sigma_y^2, \sigma_u^2, \sigma_x^2$ respectively.

Then under conditions (M.1) to (M.5), $\mathbf{X}_i$ given $\mathbf{Q}_i$ is a multivariate-normal distribution with mean $[\boldsymbol{\Sigma}_x^{-1} + (\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)\mathbf{I}_{m\times m}]^{-1}(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \mathbf{Q}_i)$, where $\boldsymbol{\mu}_x = E[\mathbf{X}_i]$ and $\boldsymbol{\Sigma}_x$ is the covariance matrix of $\mathbf{X}_i$ and both can be calculated from condition (M.2). Additionally, $E[\widetilde{\epsilon}_{ij}|\mathbf{Q}_i] = \beta_x(\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)^{-1}Q_{ij}$ and $E[Y_{i,j-1}|\mathbf{Q}_i] = \sum_{k=1}^{j-1}\alpha^{j-1-k}(\beta_0 + \beta_z m_z + \beta_x(\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)^{-1}Q_{ik}) + \alpha^{j-1}m_y$. Therefore, the moments of $E[\mathbf{Y}_i|\mathbf{Q}_i], E[\mathbf{X}_i|\mathbf{Q}_i]$, and $E[\widetilde{\epsilon}_{ij}|\mathbf{Q}_i]$ can be further calculated from the fact

$$\mathbf{Q}_i \sim \text{Multinormal}\left((\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)\boldsymbol{\mu}_x, \quad (\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)\mathbf{I}_{m\times m} + (\beta_x^2/\sigma_y^2 + 1/\sigma_u^2)^2\boldsymbol{\Sigma}_x\right).$$

Table 1: Simulation results for the pseudo conditional score estimates based on 1000 repetitions

Sample Size	Parameter	True Value	EST	EST_SE	EMP_SE	CP	MSE
			Linear transition model				
100	β_x	3.0	3.023	0.217	0.225	0.94	0.051
	β_z	0.8	0.804	0.322	0.329	0.95	0.108
	α	0.4	0.396	0.039	0.039	0.94	0.0016
200	β_x	3.0	3.017	0.152	0.150	0.95	0.023
	β_z	0.8	0.797	0.227	0.226	0.95	0.052
	α	0.4	0.397	0.027	0.027	0.95	0.0007
			Logistic transition model				
100	β_x	1.0	1.067	0.283	0.283	0.97	0.084
	β_z	0.8	0.796	0.384	0.398	0.95	0.158
	α	0.5	0.455	0.311	0.319	0.94	0.103
200	β_x	1.0	1.024	0.185	0.186	0.96	0.035
	β_z	0.8	0.812	0.262	0.258	0.96	0.067
	α	0.5	0.481	0.216	0.214	0.95	0.046

Note: EST is the mean of the estimates; EST_SE is the mean of the estimated standard errors; EMP_SE is the empirical standard error of the estimators; MSE is the mean square error; CP denotes the coverage proportion of the 95% confidence intervals.

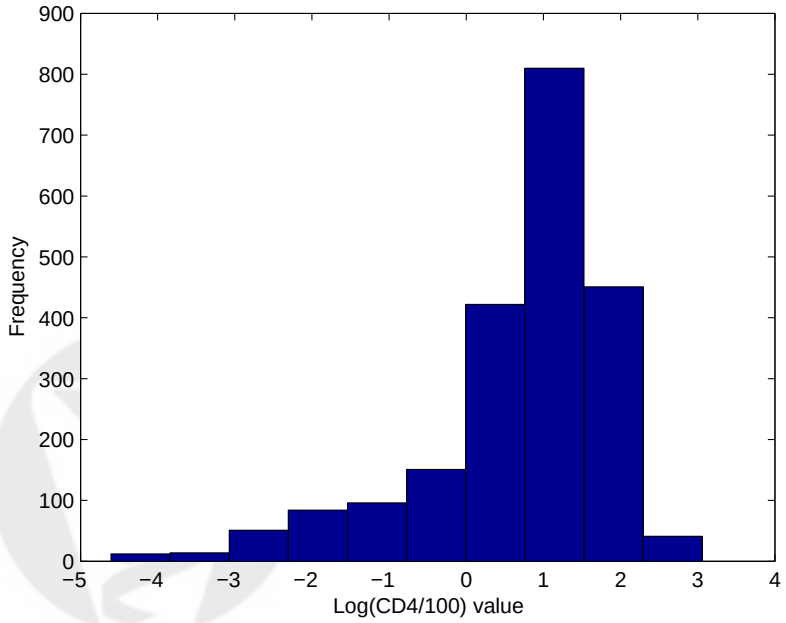

Figure 1: Histogram of log-transformed CD4 counts in the ACSUS data

Table 2: Robustness comparison between the pseudo conditional score method and the parametric maximum likelihood method when the X model is mis-specified

			Pseudo conditional score			Parametric method		
n	Par.	True	Rel. Bias(%)	EMP_SE	CP	Rel. Bias(%)	EMP_SE	CP
			X from 1st-order transition model with mixture normal error					
100	β_x	3	0.47	0.116	0.95	-2.50	0.099	0.88
	β_z	0.8	-0.87	0.231	0.96	-3.87	0.217	0.95
	α	0.4	-0.75	0.025	0.94	3.75	0.022	0.89
200	β_x	3	0.23	0.082	0.95	-2.63	0.069	0.79
	β_z	0.8	0.62	0.166	0.94	-2.37	0.158	0.95
	α	0.4	-0.25	0.018	0.95	4.00	0.016	0.82
			X from 1st-order transition model with extreme-value error					
100	β_x	3	1.53	0.221	0.94	-2.83	0.127	0.93
	β_z	0.8	2.75	0.251	0.94	-3.75	0.229	0.95
	α	0.4	-1.75	0.049	0.93	7.75	0.033	0.58
200	β_x	3	0.53	0.160	0.96	-3.27	0.094	0.88
	β_z	0.8	0.25	0.173	0.95	-5.75	0.158	0.94
	α	0.4	-0.50	0.034	0.94	8.25	0.023	0.31
			X from 2nd-order transition model with normal error					
100	β_x	3	0.40	0.102	0.97	1.87	0.105	0.89
	β_z	0.8	0.00	0.235	0.95	-6.37	0.238	0.93
	α	0.4	0.00	0.023	0.95	9.75	0.024	0.84
200	β_x	3	0.20	0.071	0.95	1.80	0.073	0.80
	β_z	0.8	-0.12	0.169	0.95	-6.62	0.170	0.94
	α	0.4	0.00	0.016	0.95	9.75	0.017	0.72

Note: see Table 1.

Table 3: Frequency table of transition orders selected using the pseudo-likelihood function from 1000 repetitions

		selection use P_AIC				selection use P_BIC			
$n = 200$		$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
	$q = 1$	5	87	42	55	43	340	64	35
	$q = 2$	28	318	129	161	90	334	39	35
	$q = 3$	9	79	37	50	2	14	1	3
$n = 400$		$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 1$	$r = 2$	$r = 3$	$r = 4$
	$q = 1$	0	14	10	19	3	160	28	16
	$q = 2$	1	462	200	151	12	678	61	18
	$q = 3$	0	83	30	30	0	14	2	2

Table 4: Application of the pseudo conditional score method to analysis of the ACSUS data

	$\sigma_u^2 = 0.38$		$\sigma_u^2 = 0.18$		Naive Estimate	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
$\log(CD4/100)$ (β_x)	-0.460	0.072	-0.416	0.067	-0.383	0.063
age	0.030	0.056	0.031	0.055	0.032	0.054
antireviral drug use	0.051	0.235	0.077	0.232	0.097	0.230
HIV symptomatic	0.086	0.191	0.069	0.188	0.058	0.187
race	0.208	0.214	0.209	0.211	0.209	0.210
sex (female vs. male)	0.621	0.243	0.577	0.239	0.545	0.237
previous hospitalization (α)	1.838	0.253	1.865	0.250	1.885	0.248