Statistica Sinica **25** (2015), 507-527
doi:http://dx.doi.org/10.5705/ss.2013.073

SEMIPARAMETRIC LONGITUDINAL MODEL WITH IRREGULAR TIME AUTOREGRESSIVE ERROR PROCESS

Yang Bai[1,2], Jian Huang[3], Rui Li[1] and Jinhong You[1,2]

[1] *Shanghai University of Finance and Economics,*
[2] *Key Laboratory of Mathematical Economics (SUFE) and* [3] *University of Iowa*

Abstract: This paper considers semiparametric inference for longitudinal data collected at irregular and possibly subject-specific times. We propose an irregular time autoregressive model for the error process in a partially linear model and develop a unified semiparametric profiling approach to estimating the regression parameters and autoregressive coefficients. An appealing feature of the proposed method is that it can effectively accommodate irregular and subject-specific observation times. We establish the asymptotic normality of the proposed estimators and derive explicit forms of their asymptotic variances. For the nonparametric component, we construct a two-stage local polynomial estimator. Our method takes into account the autoregressive error structure and does not drop any observations. The asymptotic bias and variance of the estimator are derived. We report on simulation studies conducted to evaluate the finite sample performance of the proposed method. The analysis of a dataset of CD4 cell counts of HIV seroconverters demonstrates its application.

Key words and phrases: Asymptotic normality, irregular and subject-specific observation times, locally linear estimation, nonstationary autoregressive process, profile least squares.

1. Introduction

Longitudinal data arise in many applications, notably in biomedical studies. Typically, the main objectives of a longitudinal study are to estimate how the response variable is affected by covariates and how it changes over time. A distinguishing feature of longitudinal data is the correlation of repeated observations from the same subject over time. It is important to model within-subject covariances in the analysis. This increases the efficiency of the regression parameter estimator, enhances statistical power for hypothesis testing and reduces the bias of parameter estimation (Wang (2003)); Lin et al. (2004); Wang, Carroll, and Lin (2005)). The estimation of covariance itself can provide additional information on the association among observations measured over time.

The estimation of covariance functions with longitudinal data is a challenging problem due to the presence of a large number of parameters and the positive definite constraint. Another difficulty is that longitudinal data are often collected at irregular and subject-specific times. Wu and Pourahmadi (2003) proposed a method that transforms the problem of covariance estimation to a series of simpler regression problems. Huang, Liu, and Liu (2007) utilized a smoothing-based regularization approach combined with a modified Cholesky decomposition for the estimation of covariance matrices, these methods are suitable for longitudinal data with regular observation times. Fan, Huang, and Li (2007) and Fan and Wu (2008) proposed a semiparametric approach in which a parametric correlation structure is assumed while allowing a nonparametric variance function. Li (2011) studied a kernel-based bivariate nonparametric method for covariance estimation with longitudinal data. These methods can be used in analyzing longitudinal data with irregular and subject-specific observation times.

Different from these methods, we propose an irregular time autoregressive (AR) model aimed at directly modeling the error process itself but not the covariance function. This model can accommodate irregular and possibly subject-specific observation times. Some authors have studied the analysis technique for non-stationary and irregular time series (Salcedo et al. (2012)). Ours is a natural generalization of the standard AR model, which has been used in longitudinal analysis with equally-spaced observation times (Kenward (1987)). We adopt a partially linear model for the mean component with the proposed irregular time AR model for the error process. We propose a unified semiparametric profile approach to parameter estimation. An interesting aspect of this approach is that the regression parameters and AR coefficients are estimated simultaneously based on a single profile least squares criterion. Our method does not drop any observations and takes into account within-subject correlation structures. We establish the asymptotic normality of these estimators and derive explicit forms of their asymptotic covariance matrices. For the nonparametric component, we consider a two-stage local polynomial estimator that takes into account the AR error structure and does not drop any observations. Its asymptotic bias and variance are derived.

There is a large body of literature on longitudinal data regression models. Various parametric approaches have been developed for longitudinal data analysis (Liang and Zeger (1986) and Diggle, Liang, and Zeger (1994)). Ruckstuhl, Welsh, and Carroll (2000) and Wang (2003) proposed nonparametric methods that allow one to explore possible hidden structures in the data and to reduce possible modeling biases of the traditional parametric methods. Semiparametric longitudinal data models, especially partially linear models, have been studied by He, Zhu, and Fung (2002), Chen and Jin (2006), Fan, Huang, and Li (2007), Qin, Zhu, and Fung (2009), to mention only a few. Semiparametric models strike a balance between a general nonparametric approach and a fully parametric specification, and are now being widely used in longitudinal studies.

The remainder of the paper is organized as follows. In Section 2 we propose an irregular time AR process model for correlation structure in longitudinal data and develop a semiparametric profile least squares approach for parameter estimation. In Section 3 we study the asymptotic properties of the proposed estimator. A two-stage local linear estimator of the nonparametric component is constructed in Section 4. Section 5 presents results from numerical studies. These results show that the proposed method has good finite sample properties and performs better than the estimator without considering the correlation in the data. The proposed method is also applied to CD4 count data to illustrate its application. Concluding remarks are given in Section 6. The proofs of the main results are relegated to the Supplementary Material.

2. Model and Method

2.1. Motivation

To motivate the proposed method, we consider the dataset of CD4 cell counts among HIV seroconverters that has been analyzed by many authors, see for example Zeger and Diggle (1994), Wang, Carroll, and Lin (2005), Leng, Zhang, and Pan (2010), and Li (2011). In this dataset, there are 2,376 observations of CD4 cell counts on 369 men infected with the HIV virus, whose CD4 counts were measured during a period of 3 years before to 6 years after seroconversion. We take the root of CD4 cell counts as the response as in the previous studies. Several factors may affect the level of this count, and an important question is to estimate the effects of these factors and determine if they are significant. Specifically, the dataset includes the explanatory variables SMOKE (smoking status measured by packs of cigarettes), DRUG (drug use, yes, 1; no, 0), SEXP (number of sex partners), DEPRESSION (larger values indicate increased depressive symptoms), YEAR (the effect of time since seroconversion), and AGE (relative to a given time origin). Simple pairwise scatter plots suggest that YEAR affects the dependence nonlinearly and the others have linear effects on the CD4 counts. Therefore, we initially use a semiparametric regression model to fit this data,

$$\begin{aligned}\sqrt{CD_{i,j}} = \mathrm{AGE}_{i,j}\beta_1 + \mathrm{SMOKE}_{i,j}\beta_2 + \mathrm{DRUG}_{i,j}\beta_3 + \mathrm{PARTNERS}_{i,j}\beta_4 \\ +\mathrm{DEPRESSION}_{i,j}\beta_5 + g(\mathrm{YEAR}_{i,j}) + \varepsilon_{i,j},\end{aligned} \qquad (2.1)$$

where the β_l's (l=1, …, 5) are unspecified parameters and $g(\cdot)$ is a smooth function, all of which need to be estimated.

To explore possible correlations among the $\varepsilon_{i,j}$, we examine the residuals of the fit based on (2.1) via graphical tools. In Figure 1, we plot the jth ($j > 2$) residual versus the (j-1)th and (j-2)th residuals of the ith subject in the left two panels respectively. In addition, it is natural to consider whether the dependence

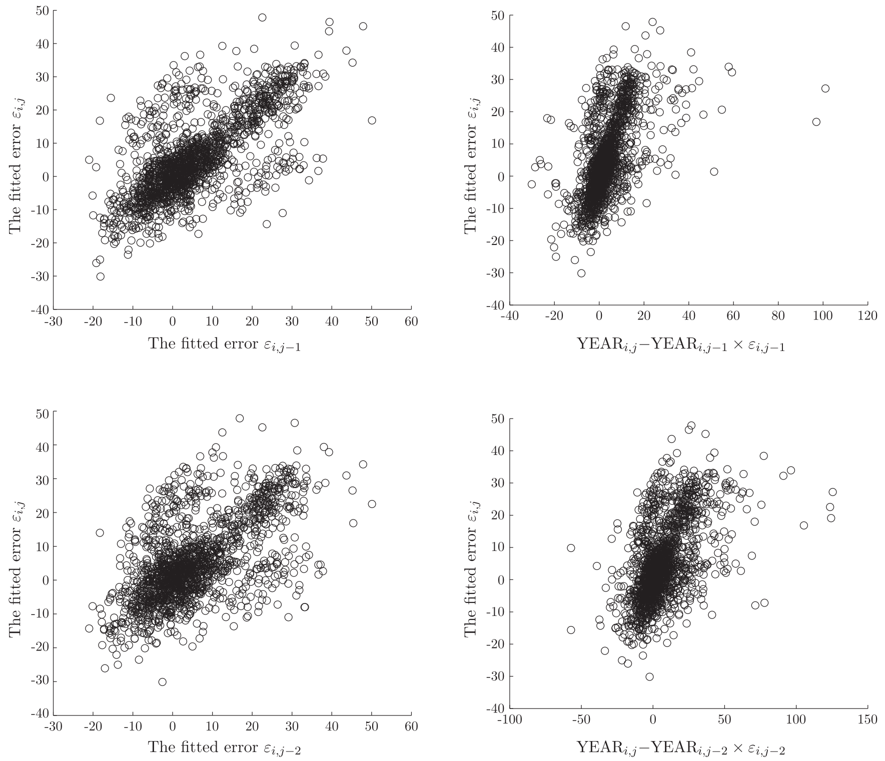

Figure 1. Plots for the residuals of model (2.1) that ignore correlations.

between the residuals also varies with their measurement time distance. So we further plot the jth ($j > 2$) residual against time-distance dependent residuals $(\text{YEAR}_{i,j} - \text{YEAR}_{i,j-1})\varepsilon_{i,j-1}$ and $(\text{YEAR}_{i,j} - \text{YEAR}_{i,j-2})\varepsilon_{i,j-2}$, respectively, in the bottom two panels in Figure 1.

Motivated by the strong dependence between the lagged errors indicated in Figure 1, we consider an AR(2) structure for the errors $\varepsilon_{i,j}$ in (2.2) that accounts for irregular time intervals,

$$\varepsilon_{i,j} = (a_1 + b_1 dY_1)\varepsilon_{i,j-1} + (a_2 + b_2 dY_2)\varepsilon_{i,j-2} + e_{i,j}, \tag{2.2}$$

where $dY_k = \text{YEAR}_{i,j} - \text{YEAR}_{i,j-k}$ is the time distance between the jth and the $(j-k)$th measurements for the ith subject, the $e_{i,j}$'s are i.i.d random variables; the significance of these parameters is discussed further in Section 5.

2.2. Partially linear model with irregular time AR error process

Suppose there are n independent subjects, and the ith subject has m_i measurements at times $t_{i,1}, \ldots, t_{i,m_i}$, not necessarily equally spaced. These observation times are also possibly subject-specific. We consider a longitudinal data partially linear model

$$Y_{i,j} = X_{i,j}^{\top}\beta + g(t_{i,j}) + \varepsilon_{i,j}, \tag{2.3}$$

where $Y_{i,j}$ is the jth measurement of the ith subject, $X_{i,j} = (X_{i,j,1}, \ldots, X_{i,j,p})^{\top}$ consists of $p(p \ll n)$ covariates for the ith subject at times $t_{i,j}, 1 \le j \le m_i$, $\beta = (\beta_1, \ldots, \beta_p)^{\top}$ is a vector of regression coefficients, g is an unknown function that describes the average temporal profile of the subjects after adjusting for the covariate effects, and the $\varepsilon_{i,j}$'s are random errors.

Let $d_{i,j,k} = t_{i,j} - t_{i,j-k}$ be the difference between the jth and $(j-k)$th observation times of the ith subject. We propose to model the error process by the irregular time AR model,

$$\varepsilon_{i,j} = \sum_{k=1}^{q}(a_k + b_k d_{i,j,k})\varepsilon_{i,j-k} + e_{i,j}, j = q+1, \ldots, m_i, i = 1, \ldots, n, \tag{2.4}$$

where $a = (a_1, \ldots, a_q)^{\top}$ and $b = (b_1, \ldots, b_q)^{\top}$ are unknown parameters, the $e_{i,j}$'s are independent and identically distributed random error terms with mean 0 and variance σ_e^2. Here $q \ge 0$ is the lag order of the model that needs to be specified prior to the analysis, or can be determined based on an existing model selection criterion. This will be discussed in Section 5. The model consists of a stationary part $\sum_{k=1}^{q} a_k \varepsilon_{i,j-k}$ and a non-stationary part $\sum_{k=1}^{q} b_k d_{i,j,k}\varepsilon_{i,j-k}$. The latter accommodates irregular and subject-specific observation times in the data.

If $b_k = 0, 1 \le k \le q$, or $d_{i,j,1} \equiv d$, a constant, as in a balanced case, this model simplifies to a standard AR model. It can be considered a linearly varying-coefficient AR model, but the coefficients depend on the time differences instead of times. With this model, the covariance structure mainly depends on the time differences. This makes sense for longitudinal data because of the natural ordering of the observations. While there is flexibility in how to define $d_{i,j,k}$, for example, we can define it based a linear or nonlinear transformation of times, (2.4) appears to offer a reasonable trade-off between model complexity and the ability to describe possibly nonstationary correlation patterns.

A similar autoregressive error structure was proposed in Wei and He (2006), but theirs focused on an autoregressive structure among responses, whereas ours includes both dependence and the covariates simultaneously to describe the correlations in each individual.

2.3. Unified semiparametric profile least squares estimation

We describe a unified semiparametric profile least squares approach for estimating β and (a, b) simultaneously. For a given β, let $R_{i,j}(\beta) = Y_{i,j} - X_{i,j}^{\top}\beta$. We can write (2.3) as

$$R_{i,j}(\beta) = g(t_{i,j}) + \varepsilon_{i,j},\ j = 1, \ldots, m_i, i = 1, \ldots, n. \tag{2.5}$$

We use the local polynomial smoothing technique (e.g., Fan and Gijbels (1996)) for estimating g. For $t_{i,j}$ in a small neighborhood of a given t, we approximate $g(t_{i,j})$ using the first order Taylor expansion

$$g(t_{i,j}) \approx g(t) + g'(t)(t_{i,j} - t) \equiv \xi + \varsigma(t_{i,j} - t),$$

where g' is the first derivative of g. This leads to a local least-squares problem: finding (ξ, ς) to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{m_i} \big[R_{i,j}(\beta) - (\xi + \varsigma(t_{i,j} - t))\big]^2 K_{h_N}(t_{i,j} - t), \tag{2.6}$$

where $K(\cdot)$ is a kernel function, h_N is a bandwidth and $K_{h_N}(\cdot) = h_N^{-1} K(\cdot/h_N)$. Here $N = \sum_{i=1}^{n} m_i$ is the total number of observations. Let $X = (X_{1,1}, \ldots, X_{1,m_1}, \ldots, X_{n,m_n})^\top$ and $Y = (Y_{1,1}, \ldots, Y_{1,m_1}, \ldots, Y_{n,m_n})^\top$. Let $u_N = (1, \ldots, 1)^\top$ be an N-vector of 1's, and $t_N = (t_{1,1}, \ldots, t_{1,m_1}, \ldots, t_{n,1}, \ldots, t_{n,m_n})^\top$. Standard least squares calculation shows that the solution to (2.6) is

$$(\widehat{\xi}(t;\beta), \widehat{\varsigma}(t;\beta))^\top = (D_t^\top W_t D_t)^{-1} D_t^\top W_t R(\beta), \tag{2.7}$$

where $R(\beta) = Y - X\beta$, $D_t = (u_N, t_N - u_N t)$, and $W_t = \text{diag}(K_{h_N}(t_N - u_N t))$. Here $K_{h_N}(t_N - u_N t)$ means K_{h_N} operates on the vector $t_N - u_N t$ component-wise. For any given β in (2.7), g can be estimated by $\widehat{g}(t;\beta) \equiv \widehat{\xi}(t;\beta)$. We estimate the parameters β and (a, b) in (2.3) and (2.4) based on the profile residuals $\widehat{R}_{i,j}(\beta) \equiv Y_{i,j} - X_{i,j}\beta - \widehat{g}(t_{i,j};\beta)$ as follows.

Let $S = (S_{1,1}, \ldots, S_{1,m_1}, \ldots, S_{n,m_n})^\top$, where $S_{i,j} = (1, 0)(D_{t_{i,j}}^\top W_{t_{i,j}} D_{t_{i,j}})^{-1} D_{t_{i,j}}^\top W_{t_{i,j}}$. Let $\widehat{Y} = (I - S)Y$ and $\widehat{X} = (I - S)X$. Some algebra shows that $\widehat{R}_{i,j}(\beta) = \widehat{Y}_{i,j} - \widehat{X}_{i,j}^\top \beta$, where $\widehat{Y}_{i,j}$ is an element in $\widehat{Y}$ and $\widehat{X}_{i,j}$ is a column in $\widehat{X}$ corresponding to $Y_{i,j}$ in Y and $X_{i,j}$ in X position wise, respectively. As shown in the Supplementary Material, $\widehat{R}_{i,j}(\beta) \approx \varepsilon_{i,j} + O_p(1/\sqrt{N h_N})$ if β is the underlying value in the model. This and (2.4) for the $\varepsilon_{i,j}$ motivate us to propose the profile least squares criterion

$$Q(\beta, a, b) = \sum_{i=1}^{n} \Big[\sum_{j=q+1}^{m_i} \Big(\widehat{R}_{i,j}(\beta) - \sum_{k=1}^{q} (a_k + b_k d_{i,j,k}) \widehat{R}_{i,j-k}(\beta) \Big)^2 + \sum_{j=1}^{q} \widehat{R}_{i,j}^2(\beta) \Big]. \tag{2.8}$$

The estimator of (β, a, b) is the value $(\widehat{\beta}_N, \widehat{a}_N, \widehat{b}_N)$ that minimizes (2.8).

Here we take advantage of the irregular time AR structure and estimate the regression parameters and the coefficients in the error model based on a single profile least squares criterion. This differs from the usual approach in longitudinal data models where the parameters in the mean function and correlation structure are estimated separately. All the observations are used in (2.8). For the first q observations, because their correlation structure cannot be estimated based on (2.4) with a lag order of q, we used the estimated residuals directly.

It is not easy to directly minimize Q in (2.8) jointly with respect to (β, a, b). We use an iterative procedure that alternately updates the regression parameter β and autoregressive parameter (a, b).

Step 1. For a given (a, b), minimize

$$Q_1(\beta) = \sum_{i=1}^{n} \Big[\sum_{j=q+1}^{m_i} \Big(\widehat{R}_{i,j}(\beta) - \sum_{k=1}^{q} (a_k + b_k d_{i,j,k}) \widehat{R}_{i,j-k}(\beta) \Big)^2 + \sum_{j=1}^{q} \widehat{R}_{i,j}^2(\beta) \Big] \text{ with fixed } a, b$$

with respect to β.

Step 2. For a given β, minimize

$$Q_2(a, b) = \sum_{i=1}^{n} \Big[\sum_{j=q+1}^{m_i} \Big(\widehat{R}_{i,j}(\beta) - \sum_{k=1}^{q} (a_k + b_k d_{i,j,k}) \widehat{R}_{i,j-k}(\beta) \Big)^2 + \sum_{j=1}^{q} \widehat{R}_{i,j}^2(\beta) \Big] \text{ with fixed } \beta$$

with respect to (a, b).

We start the iteration with $(a, b) = (0, 0)$, which yields the initial solution corresponding to the ordinary least squares estimate of β. We then carry out Steps 1 and 2 iteratively until convergence. Since Q_1 (with fixed (a, b)) and Q_2 (with fixed β) are quadratic functions, it is easy to compute their minimizers. In our numerical studies, this algorithm converges quickly in both simulation studies and data analysis.

3. Asymptotic Properties

We consider the theoretical properties of the profile least squares estimator. We show that $\widehat{\beta}_N$ and $(\widehat{a}_N, \widehat{b}_N)$ are asymptotically normal and asymptotically independent. We also propose consistent variance estimators for $\widehat{\beta}_N$ and $(\widehat{a}_N, \widehat{b}_N)$.

Let $\eta(t) = (\eta_1(t), \ldots, \eta_p(t))^\top$ be defined by

$$X_{i,j} = \eta(t_{i,j}) + \delta_{i,j}, j = 1, \ldots, m_i, \quad i = 1, \ldots, n,$$

where $\delta_{i,j} = (\delta_{i,j,1}, \ldots, \delta_{i,j,p})^\top$ satisfies $E(\delta_{i,j} | t_{i,j}) = \mathbf{0}$. Here the conditional expectation is taken componentwise.

Let $\delta^*_{i,j} = \delta_{i,j} - \sum_{k=1}^{q}(a_k + b_k d_{i,j,k})\delta_{i,j-k}, i = 1, \ldots, n$ and $j = q+1, \ldots, m_i$, and let

$$\boldsymbol{\zeta}_{ij} = (\varepsilon_{i,j-1}, \ldots, \varepsilon_{i,j-q}, \varepsilon_{i,j-1}d_{i,j,1}, \ldots, \varepsilon_{i,j-q}d_{i,j,q})^\top.$$

Assume

$$\frac{1}{N}\sum_{i=1}^{n}\Big(\sum_{j=1}^{q}\delta_{i,j}\delta_{i,j}^\top + \sum_{j=q+1}^{m_i}\delta^*_{i,j}\delta^{*\top}_{i,j}\Big) \rightarrow_p D > 0, \tag{3.1}$$

$$\frac{1}{N-nq}\sum_{i=1}^{n}\sum_{j=q+1}^{m_i}\zeta_{i,j}\zeta_{i,j}^\top \rightarrow_p \Lambda > 0, \tag{3.2}$$

$$\frac{1}{N}\sum_{i=1}^{n}\Big[\sigma_e^2\sum_{j=q+1}^{m_i}\delta^*_{i,j}\delta^{*\top}_{i,j} + (\delta_{i,1}, \ldots, \delta_{i,q})\mathrm{Cov}\{(\varepsilon_{i,1}, \ldots, \varepsilon_{i,q})^\top\}(\delta_{i,1}, \ldots, \delta_{i,q})^\top\Big] \rightarrow_p \Delta. \tag{3.3}$$

Here and in the sequel, any convergence statement is for $n \to \infty$.

Theorem 1. *Suppose the* $e_{i,j}$*'s are independent and identified distributed random variables with mean zero, variance* σ_e^2*, and finite fourth moment. Then, under* (A1)−(A5) *given in the on-line Supplementary Material and* (3.1) *to* (3.3)*, we have*

(i) $\sqrt{N}(\widehat{\beta}_N - \beta) \rightarrow_D N(0, D^{-1}\Delta D^{-1})$.

(ii) $\sqrt{N-nd}\{(\widehat{a}_N^\top, \widehat{b}_N^\top)^\top - (a^\top, b^\top)^\top\} \rightarrow_D N(0, \sigma_e^2\Lambda^{-1})$.

(iii) $\widehat{\beta}_N$ *and* $(\widehat{a}_N, \widehat{b}_N)$ *are asymptotically independent.*

Proofs can be found in the on-line supplement. The asymptotic normality results here can serve as the basis for the statistical inference for (β, a, b). The asymptotic covariance matrices for $\widehat{\beta}_N$ and $(\widehat{a}_N, \widehat{b}_N)$ have a relatively simple and explicit structure that enables us to directly construct variance estimators without having to resort to resampling-based methods. Part (iii) of Theorem 1 is similar to the standard result from linear regression where the estimated regression parameters and the estimator of the variance of the error distribution are asymptotically independent.

Consider the estimation of the covariance matrices of the estimators. This involves the estimation of σ_e^2, D, Δ, and Λ. We estimate σ_e^2 by

$$\widehat{\sigma}_{e,N}^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i - q}\sum_{j=q+1}^{m_i}\Big\{\widehat{\varepsilon}_{i,j} - \sum_{k=1}^{q}(\widehat{a}_{k,N} + \widehat{b}_{k,N}d_{i,j,k})\widehat{\varepsilon}_{i,j-k}\Big\}^2,$$

where $\widehat{\varepsilon}_{i,j} = \widehat{Y}_{i,j} - \widehat{X}_{i,j}^\top\widehat{\beta}_N, i = 1, \ldots, n, j = 1, \ldots, m_i$. Let $\widehat{X}^*_{i,j} = \widehat{X}_{i,j} - \sum_{k=1}^{q}(\widehat{a}_{k,N} + \widehat{b}_{k,N}d_{i,j,k})\widehat{X}_{i,j-k}$ and $\widehat{\boldsymbol{\zeta}}_{ij} = (\widehat{\varepsilon}_{i,j-1}, \ldots, \widehat{\varepsilon}_{i,j-q}, \widehat{\varepsilon}_{i,j-1}d_{i,j,1}, \ldots, \widehat{\varepsilon}_{i,j-q}$

$d_{i,j,q})^\top$, $i=1,\ldots,n$, $j=q+1,\ldots,m_i$. We estimate D,Λ and Δ by

$$\widehat{D}_N=\frac{1}{N}\sum_{i=1}^{n}\Big(\sum_{j=1}^{q}\widehat{X}_{i,j}\widehat{X}_{i,j}^\top+\sum_{j=q+1}^{m_i}\widehat{X}_{i,j}^*\widehat{X}_{i,j}^{*\top}\Big),$$

$$\widehat{\Lambda}_N=\frac{1}{N-nq}\sum_{i=1}^{n}\sum_{j=q+1}^{m_i}\widehat{\boldsymbol{\zeta}}_{i,j}\widehat{\boldsymbol{\zeta}}_{i,j}^\top,$$

$$\widehat{\Delta}_N=\frac{1}{N}\sum_{i=1}^{n}\Big[\widehat{\sigma}_{e,N}^2\sum_{j=q+1}^{m_i}\widehat{X}_{i,j}^*\widehat{X}_{i,j}^{*\top}+\Big(\sum_{j=1}^{q}\widehat{X}_{i,j}^\top\widehat{\varepsilon}_{i,j}\Big)\Big(\sum_{j=1}^{q}\widehat{X}_{i,j}^\top\widehat{\varepsilon}_{i,j}\Big)^\top\Big].$$

Theorem 2. *Suppose that the conditions of Theorem* 1 *hold. Then*

$$\sqrt{N-nq}(\widehat{\sigma}_{e,N}^2-\sigma_e^2)\xrightarrow{D}N(0,\,Var(e_{i,j}^2)),$$

$$\widehat{D}_N\rightarrow_p D,\ \widehat{\Lambda}_N\rightarrow_p\Lambda,\quad and\ \ \widehat{\Delta}_N\rightarrow_p\Delta.$$

Based on this result, a consistent estimator of the covariance matrix of $\widehat{\beta}_N$ is $\widehat{D}_N^{-1}\widehat{\Delta}_N\widehat{D}_N^{-1}/N$ and a consistent estimator of the covariance matrix of $(\widehat{a}_N,\widehat{b}_N)$ is $\widehat{\sigma}_{e,N}^2\widehat{\Lambda}_N^{-1}/(N-nq)$.

4. Two-stage Local Linear Estimator of Nonparametric Component

With the estimator $\widehat{\beta}_N$ of β, (2.7) gives the local polynomial (linear) estimator of $(g(t),g'(t))^\top$,

$$(\widehat{g}_N(t),\widehat{g}_N'(t))^\top=(D_t^\top W_tD_t)^{-1}D_t^\top W_t(Y-X\widehat{\beta}_N).$$

It may not be efficient since it does not take into account the AR error structure.

We describe a two-stage approach for estimating g that explicitly incorporates the AR error structure (2.4). With $R_{i,j}(\beta)=Y_{i,j}-X_{i,j}^\top\beta$, let

$$Y_{i,j}^*=R_{i,j}(\beta)-\sum_{k=1}^{q}(a_k+b_kd_{i,j,k})R_{i,j-k}(\beta)+\sum_{k=1}^{q}(a_k+b_kd_{i,j,k})g(t_{i,j-k}).$$

According to (2.3) and (2.4),

$$Y_{i,j}^*=g(t_{i,j})+e_{i,j},\quad i=1,\ldots,n \text{ and } j=q+1,\ldots,m_i. \tag{4.1}$$

Since $\mathrm{Var}(e_{i,j})$ is usually less than $\mathrm{Var}(\varepsilon_{i,j})$, based on (4.1) we can construct a more efficient estimator of $g(t_{i,j})$. As $Y_{i,j}^*$ contains unknown parameters (β,a,b) and the unknown function $g(t)$, we replace $Y_{i,j}^*$ by $\tilde{Y}_{i,j}^*$, where

$$\tilde{Y}_{i,j}^*=R_{i,j}(\widehat{\beta}_N)-\sum_{k=1}^{q}(\widehat{a}_{k,N}+\widehat{b}_{k,N}d_{i,j,k})R_{i,j-k}(\widehat{\beta}_N)+\sum_{k=1}^{q}(\widehat{a}_{k,N}+\widehat{b}_{k,N}d_{i,j,k})\widehat{g}_N(t_{i,j-k}),$$

for $i = 1, \ldots, n, j = q+1, \ldots, m_i$. For the first q observations, define $\tilde{Y}^*_{i,j} = R_{i,j}(\widehat{\beta}_N), i = 1, \cdots, n, j = 1, \cdots, q$. Denote the vector of all the $\tilde{Y}^*_{i,j}$ by

$$\tilde{Y}^* = (\tilde{Y}^*_{1,1}, \ldots, \tilde{Y}^*_{1,q}, \tilde{Y}^*_{1,q+1}, \ldots, \tilde{Y}^*_{1,m_1}, \ldots, \tilde{Y}^*_{n,1}, \ldots, \tilde{Y}^*_{n,q}, \tilde{Y}^*_{n,q+1}, \ldots, \tilde{Y}^*_{n,m_n})^\top.$$

Similar to (2.6), we consider the criterion

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\big[\tilde{Y}^*_{i,j} - (\xi + \varsigma(t_{i,j} - t))\big]^2 K_{h^*_N}(t_{i,j} - t),$$

and obtain a two-stage local linear estimator of $(g(t), g'(t))$,

$$(\widehat{g}_N^{TS}(t), \widehat{g}_N^{\prime TS}(t))^\top = (D_t^{*\top} W_t^* D_t^*)^{-1} D_t^{*\top} W_t^* \tilde{Y}^*, \tag{4.2}$$

where D_t^*, W_t^* have the same form as D_t, W_t except that h_N is replaced by h_N^*.

Let

$$\mu_k = \int_{-\infty}^{\infty} x^k K(x) dx, \quad \nu_k = \int_{-\infty}^{\infty} x^k K^2(x) dx, \quad k = 0, 1, 2, 3,$$

$$\tau^2 = \lim_{n\to\infty} \frac{\sigma_e^2 \sum_{i=1}^n (m_i - d) + \sum_{i=1}^n \sum_{j=1}^q \mathrm{Var}(\varepsilon_{i,j})}{N}.$$

Theorem 3. *Suppose that the $e_{i,j}$'s are independent and identified distributed random variables with mean zero, variance σ_e^2, and finite fourth moment. Then under* (A1)−(A6) *given in the Supplementary Material, we have*

$$\sqrt{Nh_N^*}\Big[H^{*-1}\Big\{\begin{pmatrix}\widehat{g}_N^{TS}(t)\\ \widehat{g}_N^{\prime TS}(t)\end{pmatrix} - \begin{pmatrix}g(t)\\ g'(t)\end{pmatrix}\Big\} - \frac{h_N^{*2}}{2}\begin{pmatrix}\kappa_1 g^{''}(t)\\ \kappa_2 g^{''}(t)\end{pmatrix} + o(h_N^{*2})\Big] \rightarrow_D N(0, \Gamma^{TS}),$$

where $H^* = diag(1, h_N^*)$, $g^{''}$ is the second derive of g and

$$\Gamma^{TS} = \tau^2\Big\{f(t)(\mu_2 - \mu_1^2)^2\Big\}^{-1}\begin{pmatrix}\gamma_{11}\ \gamma_{12}\\ \gamma_{21}\ \gamma_{22}\end{pmatrix},$$

$$\kappa_1 = \frac{\mu_2^2 - \mu_1\mu_3}{\mu_2 - \mu_1^2}, \qquad \kappa_2 = \frac{\mu_3 - \mu_1\mu_2}{\mu_2 - \mu_1^2},$$

$$\gamma_{11} = \mu_2^2\nu_0 - 2\mu_1\mu_2\nu_1 + \mu_1^2\nu_2, \quad \gamma_{12} = (\mu_1^2 + \mu_2)\nu_1 - \mu_1\mu_2\nu_0 - \mu_1\nu_2,$$
$$\gamma_{21} = (\mu_1^2 + \mu_2)\nu_1 - \mu_1\mu_2\nu_0 - \mu_1\nu_2, \qquad \gamma_{22} = \nu_2 - \mu_1(2\nu_1 + \mu_1\nu_0).$$

Corollary 1. *Suppose the conditions of Theorem* 3 *hold. Then*

$$\sqrt{Nh_N^*}\Big\{\widehat{g}_N^{TS}(t) - g(t) - \frac{h_N^{*2}}{2}\frac{\mu_2^2 - \mu_1\mu_3}{\mu_2 - \mu_1^2}g^{''}(t) + o(h_N^{*2})\Big\} \rightarrow_D N(0, \gamma_{TS}^2),$$

where $\gamma_{TS}^2 = \tau^2\{f(t)\}^{-1}(\kappa_3^2\nu_0 - 2\kappa_3\kappa_4\nu_1 + \kappa_4^2\nu_2)$ with $\kappa_3 = \mu_2/(\mu_2 - \mu_1^2)$ and $\kappa_4 = \mu_1/(\mu_2 - \mu_1^2)$.

To apply Corollary 1 or Theorem 3 to make statistical inference for $g(\cdot)$ or $(g(\cdot), g'(\cdot))^\top$, a consistent estimator of γ_{TS}^2 or Γ^{TS} is needed. Since μ_1, μ_2, μ_3, ν_0, ν_1 and ν_2 are known constants, we just need to estimate τ^2 and $f(t)$. Take

$$\widehat{\tau}_N^2 = \frac{\widehat{\sigma}_{e,N}^2 \sum_{i=1}^n (m_i - q) + \sum_{i=1}^n \sum_{j=1}^q (\widehat{\varepsilon}_{i,j})^2}{N} \text{ and } \widehat{f}_N(t) = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} K_{h_N^*}(t_{i,j} - t).$$

We can show that $\widehat{\tau}_N^2$ and $\widehat{f}_N(t)$ are consistent estimators of τ^2 and $f(t)$, respectively.

We note that $\widehat{g}_N^{TS}(\cdot)$ involves both of the smoothing parameters h_N and h_N^*. Condition (A6) given in the Supplementary Material requires the smoothing parameter h_N^* to be of the standard order and the smoothing parameter h_N for the initial estimators $\widehat{g}_N(\cdot)$ to be of a smaller order than the standard $O(N^{-1/5})$. This requirement is used to control the bias in the preliminary step of the estimation. Simulation experiments show that the final results are not very sensitive to the choice of the smoothing parameter h_N and that the usual optimal smoothing parameters divided by a constant, say 1.5 or 2, can be used.

5. Numerical Studies

5.1. Determination of the AR order

In practice, the true AR order in the errors is not known a priori. We consider a penalized selection method for determining it. By minimizing

$$P(a,b) = \sum_{i=1}^n \sum_{j=q+1}^{m_i} \left\{ \widehat{R}_{i,j}(\tilde{\beta}_N) - \sum_{k=1}^q (a_k + b_k d_{i,j,k}) \widehat{R}_{i,j-k}(\tilde{\beta}_N) \right\}^2 + N \sum_{k=1}^q \lambda_k p(||\theta_k||),$$

we can specify the significant (a_k, b_k)'s and the autoregressive order correspondingly, where λ_k's are tuning parameters, $\theta_k = (a_k, b_k)^\top$ and the $p(\cdot)$ is a penalty function. Here, we use the smoothly clipped absolute deviation (SCAD, Fan and Li (1996)) penalty for its unbiasedness, sparsity, and continuity. We present simulation results in Example 3 to demonstrate its good selection performance.

5.2. Simulation studies

We conducted simulation studies to examine the finite sample performances of the proposed estimators. Consider first that the true AR order is known.

Example 1. The data were generated from

$$\begin{aligned} Y_{i,j} &= X_{i,j}^\top \beta + g(t_{i,j}) + \varepsilon_{i,j}, \\ \varepsilon_{i,j} &= \sum_{k=1}^2 (a_k + b_k d_{i,j,k}) \varepsilon_{i,j-k} + e_{i,j}, \end{aligned} \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m_i,$$

where $X_{i,j,1} = \sin(t_{i,j}) + \varrho_{i,j,1}$, $X_{i,j,2} = 1 + t_{i,j}^2 + \varrho_{i,j,2}$ and $g(t_{i,j}) = \sin^3(2\pi t_{i,j}) - e^{t_{i,j}} + t_{i,j}$ with $t_{i,j} \sim U(0,1)$, $\varrho_{i,j,1}, \varrho_{i,j,2}$, and the $e_{i,j}$ were $N(0,1)$. The coefficients were $\beta = (0.5, -0.8)^\top$ and $(a_1, a_2) = (0.6, -0.5), (b_1, b_2) = (-0.3, 0.4)$.

The sample size was $n = 100, 200, 300$ and the number of within-subject observations was $m_i = m = 5, 10, 15$ for each subject. In each scenario, the number of replications was 1,000. We used the truncated Gaussian kernel

$$K_{h_N}(\cdot) = \frac{1}{h_N\sqrt{2\pi}} \exp\Big\{\frac{-(\cdot)^2}{2h_N^2}\Big\} I(|\cdot| \le 1).$$

Bandwidth selection is important in nonparametric regression. Lin and Carroll (2001) suggested using cross-validation or the empirical bias bandwidth selection method of Ruppert (1997) with longitudinal data. We used the former for its simplicity. Half of the optimal bandwidth is used as the bandwidth in the first stage and the optimal bandwidth used in the second stage for the estimation of the nonparametric component $g(\cdot)$. Based on our experience, the two-stage estimator is not sensitive to the choice of the bandwidth of the first-stage estimator, see Table 2.

For the estimator $\widehat{\beta}_N$ of the regression parameter β, the average sample bias (bias), the empirical standard deviation (std) of the estimates of 1,000 replications, the mean of the estimate of the standard deviation (se) based on the asymptotic covariance matrix and the empirical coverage probability (cp) of the 95% confidence intervals are summarized in Table 1. We also present the corresponding bias, std, se, and cp of the estimator $\check{\beta}_N$ based on the criterion $\sum_{i=1}^n \sum_{j=1}^{m_i} \widehat{R}_{i,j}^2(\beta)$, where $\widehat{R}_{i,j}(\beta)$ is defined in Section 2. This estimator is consistent with an asymptotic normal distribution, but does not take into account the correlation structure in the data.

From Table 1, we see that the proposed estimator $\widehat{\beta}_N$ has small bias, even in the case of moderate sample size ($n = 100$ and $m = 5$). The biases decrease as the sample size increases. The estimated standard deviations approximate the empirical standard deviations well in all the cases. The empirical coverage probability of the confidence interval is close to the nominal level 95%. The proposed estimator $\widehat{\beta}_N$ also has smaller variance than the $\check{\beta}_N$.

For the estimators of the AR coefficients (a_1, b_1), their bias, std, se, and cp are also summarized in Table 1. The biases of the estimated AR coefficients are small and decrease as the sample increases. Also, on average, the estimated standard errors are close to the empirical standard deviations.

For the two-stage estimator $\widehat{g}_N$ of the nonparametric component g, we evaluated its performance using the square root of average squared errors (RASE)

$$\mathrm{RASE}(\widehat{g}_N) = \Big[\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\widehat{g}_N(t_{i,j}) - g(t_{i,j})\}^2\Big]^{1/2}.$$

Table 1. The average bias (bias), average empirical standard deviation (std) of the estimates, average of the estimated standard deviation (se) based on the asymptotic covariance matrix and empirical coverage probability (cp), calculated based on 1,000 replications.

	m	5			10			15		
	n	100	200	300	100	200	300	100	200	300
$\check{\beta}_{1,N}$	bias	0.0021	0.0006	-0.0006	-0.0002	0.0003	0.0003	-0.0001	-0.0005	0.0001
	std	0.0360	0.0268	0.0186	0.0228	0.0156	0.0129	0.0175	0.0120	0.0099
	se	0.0347	0.0260	0.0188	0.0227	0.0157	0.0128	0.0170	0.0121	0.0099
	cp	0.9410	0.9410	0.9480	0.9430	0.9530	0.9490	0.9410	0.9500	0.9500
$\widehat{\beta}_{1,N}$	bias	0.0022	0.0004	-0.0008	-0.0003	0.0001	0.0001	0.0000	-0.0002	-0.0001
	std	0.0286	0.0226	0.0159	0.0171	0.0116	0.0102	0.0131	0.0091	0.0076
	se	0.0287	0.0220	0.0162	0.0175	0.0121	0.0101	0.0127	0.0092	0.0074
	cp	0.9460	0.9480	0.9510	0.9430	0.9600	0.9480	0.9450	0.9540	0.9450
$\check{\beta}_{2,N}$	bias	-0.0028	0.0001	-0.0010	0.0003	-0.0002	0.0011	0.0008	0.0010	-0.0006
	std	0.0741	0.0523	0.0400	0.0442	0.0302	0.0249	0.0333	0.0233	0.0195
	se	0.0709	0.0500	0.0401	0.0444	0.0309	0.0251	0.0345	0.0239	0.0197
	cp	0.9360	0.9390	0.9450	0.9440	0.9640	0.9540	0.9600	0.9540	0.9500
$\widehat{\beta}_{2,N}$	bias	-0.0029	-0.0003	-0.0008	-0.0001	0.0003	0.0004	0.0001	0.0008	0.0004
	std	0.0635	0.0434	0.0331	0.0347	0.0239	0.0188	0.0256	0.0177	0.0144
	se	0.0604	0.0420	0.0333	0.0349	0.0238	0.0193	0.0264	0.0183	0.0149
	cp	0.9320	0.9460	0.9520	0.9510	0.9500	0.9580	0.9570	0.9510	0.9520
$\widehat{a}_{1,N}$	bias	-0.0138	-0.0078	-0.0055	-0.0113	-0.0054	-0.0048	-0.0084	-0.0059	-0.0035
	std	0.0736	0.0514	0.0413	0.0431	0.0302	0.0251	0.0333	0.0252	0.0195
	se	0.0734	0.0499	0.0397	0.0423	0.0298	0.0250	0.0343	0.0246	0.0197
	cp	0.9400	0.9370	0.9430	0.9400	0.9420	0.9390	0.9570	0.9420	0.9360
$\widehat{a}_{2,N}$	bias	0.0074	0.0063	0.0050	0.0010	0.0014	0.0027	0.0024	0.0015	0.0010
	std	0.0888	0.0598	0.0499	0.0494	0.0360	0.0300	0.0414	0.0301	0.0236
	se	0.0857	0.0592	0.0486	0.0496	0.0353	0.0295	0.0413	0.0296	0.0234
	cp	0.9330	0.9520	0.9330	0.9450	0.9520	0.9510	0.9580	0.9460	0.9450
$\widehat{b}_{1,N}$	bias	0.0037	0.0005	0.0011	0.0058	0.0036	0.0026	0.0069	0.0044	0.0020
	std	0.1046	0.0740	0.0602	0.0815	0.0577	0.0483	0.0743	0.0538	0.0420
	se	0.1020	0.0713	0.0570	0.0800	0.0572	0.0477	0.0748	0.0519	0.0429
	cp	0.9350	0.9490	0.9390	0.9560	0.9540	0.9590	0.9490	0.9490	0.9590
$\widehat{b}_{2,N}$	bias	-0.0031	-0.0033	-0.0018	0.0042	0.0004	-0.0012	0.0006	0.0012	0.0018
	std	0.1138	0.0758	0.0632	0.0801	0.0577	0.0483	0.0730	0.0526	0.0431
	se	0.1097	0.0761	0.0621	0.0804	0.0575	0.0471	0.0735	0.0518	0.0421
	cp	0.9430	0.9530	0.9440	0.9500	0.9480	0.9440	0.9520	0.9480	0.9460
$\widehat{g}_N(\cdot)$	Mean(RASE)	0.1429	0.1054	0.0888	0.0987	0.0737	0.0611	0.0820	0.0594	0.0502
	Std (RASE)	0.0313	0.0208	0.0173	0.0205	0.0137	0.0110	0.0153	0.0104	0.0084
$\widehat{g}_N^{TS}(\cdot)$	Mean(RASE)	0.1219	0.0915	0.0760	0.0856	0.0645	0.0534	0.0712	0.0524	0.0444
	Std(RASE)	0.0304	0.0203	0.0169	0.0196	0.0134	0.0107	0.0149	0.0101	0.0084

The empirical mean value (Mean) and standard deviation (Std) of RASE calculated over 1,000 replications are given in the bottom of Table 1. We see that the proposed two-stage estimator performs better than the standard one-stage estimator without considering the correlation structure.

In practice, it is perhaps impossible to correctly specify the correlation structure in the errors. However, our approach is robust to misspecification of the the

Table 2. The empirical mean value and average standard deviation (listed in parentheses) of the proposed estimates for parametric and nonparametric components calculated based on 1,000 replications with different bandwidth (h_N) in the first stage estimation.

m	5			10		
n	50	100	200	50	100	200
$h_N = 0.25h_{opt}$						
$\widehat{\beta}_{1,N}$	0.4989 (0.0442)	0.4991 (0.0305)	0.4990 (0.0209)	0.4997 (0.0275)	0.5007 (0.0184)	0.4997 (0.0124)
$\widehat{\beta}_{2,N}$	-0.7986 (0.0921)	-0.7985 (0.0565)	-0.7975 (0.0417)	-0.8018 (0.0499)	-0.8004 (0.0332)	-0.8000 (0.0242)
$\widehat{\beta}_{3,N}$	1.4992 (0.0493)	1.5003 (0.0300)	1.4997 (0.0221)	1.5002 (0.0263)	1.5003 (0.0175)	1.5001 (0.0126)
$\widehat{\beta}_{4,N}$	0.2001 (0.0243)	0.2001 (0.0099)	0.2000 (0.0009)	0.1999 (0.0086)	0.2004 (0.0084)	0.2000 (0.0018)
$\widehat{\beta}_{5,N}$	-1.2026 (0.0435)	-1.1993 (0.0300)	-1.1998 (0.0221)	-1.1995 (0.0246)	-1.2005 (0.0179)	-1.2001 (0.0129)
$\widehat{a}_{1,N}$	0.5432 (0.1174)	0.5765 (0.0736)	0.5844 (0.0530)	0.5716 (0.0649)	0.5840 (0.0436)	0.5910 (0.0297)
$\widehat{a}_{2,N}$	-0.4766 (0.1453)	-0.4933 (0.0963)	-0.4916 (0.0646)	-0.4845 (0.0748)	-0.4893 (0.0536)	-0.4936 (0.0363)
$\widehat{b}_{1,N}$	-0.2609 (0.1602)	-0.2851 (0.1036)	-0.2896 (0.0736)	-0.2799 (0.1255)	-0.2907 (0.0818)	-0.2955 (0.0564)
$\widehat{b}_{2,N}$	0.3884 (0.1847)	0.3972 (0.1222)	0.3936 (0.0841)	0.3925 (0.1200)	0.3935 (0.0848)	0.3964 (0.0571)
$\widehat{g}_N(\cdot)$	0.1959 (0.0501)	0.1511 (0.0346)	0.1186 (0.0229)	0.1459 (0.0294)	0.1115 (0.0204)	0.0869 (0.0146)
$h_N = 0.5h_{opt}$						
$\widehat{\beta}_{1,N}$	0.4997 (0.0394)	0.4996 (0.0327)	0.4993 (0.0213)	0.5010 (0.0262)	0.5008 (0.0170)	0.5003 (0.0122)
$\widehat{\beta}_{2,N}$	-0.7949 (0.0917)	-0.8033 (0.0578)	-0.8022 (0.0419)	-0.7989 (0.0481)	-0.8013 (0.0327)	-0.8007 (0.0234)
$\widehat{\beta}_{3,N}$	1.5015 (0.0447)	1.5010 (0.0286)	1.4999 (0.0216)	1.5012 (0.0249)	1.5007 (0.0166)	1.4991 (0.0118)
$\widehat{\beta}_{4,N}$	0.1999 (0.0139)	0.2003 (0.0099)	0.2001 (0.0077)	0.1998 (0.0081)	0.2005 (0.0080)	0.2004 (0.0038)
$\widehat{\beta}_{5,N}$	-1.2033 (0.0394)	-1.1987 (0.0285)	-1.1992 (0.0210)	-1.1996 (0.0254)	-1.2000 (0.0186)	-1.2003 (0.0120)
$\widehat{a}_{1,N}$	0.5669 (0.1039)	0.5879 (0.0696)	0.5915 (0.0536)	0.5826 (0.0643)	0.5900 (0.0454)	0.5955 (0.0302)
$\widehat{a}_{2,N}$	-0.4999 (0.1392)	-0.5003 (0.0851)	-0.4968 (0.0624)	-0.4977 (0.0720)	-0.4982 (0.0506)	-0.4983 (0.0361)
$\widehat{b}_{1,N}$	-0.2760 (0.1470)	-0.2910 (0.1011)	-0.2952 (0.0725)	-0.2811 (0.1155)	-0.2947 (0.0844)	-0.2970 (0.0570)
$\widehat{b}_{2,N}$	0.4076 (0.1794)	0.4023 (0.1122)	0.3962 (0.0788)	0.3979 (0.1200)	0.4012 (0.0859)	0.3987 (0.0559)
$\widehat{g}_N(\cdot)$	0.1970 (0.0492)	0.1496 (0.0350)	0.1160 (0.0230)	0.1467 (0.0268)	0.1125 (0.0193)	0.0870 (0.0136)
$h_N = h_{opt}$						
$\widehat{\beta}_{1,N}$	0.5065 (0.0456)	0.5050 (0.0324)	0.5018 (0.0226)	0.5021 (0.0260)	0.5021 (0.0167)	0.5018 (0.0122)
$\widehat{\beta}_{2,N}$	-0.8037 (0.0870)	-0.7938 (0.0596)	-0.7950 (0.0417)	-0.8066 (0.0527)	-0.7964 (0.0352)	-0.7997 (0.0248)
$\widehat{\beta}_{3,N}$	1.4947 (0.0489)	1.5051 (0.0320)	1.5022 (0.0194)	1.4989 (0.0245)	1.4993 (0.0163)	1.5016 (0.0120)
$\widehat{\beta}_{4,N}$	0.2018 (0.0180)	0.2006 (0.0124)	0.1997 (0.0065)	0.1996 (0.0076)	0.1999 (0.0074)	0.1999 (0.0031)
$\widehat{\beta}_{5,N}$	-1.2122 (0.0447)	-1.2016 (0.0345)	-1.1981 (0.0214)	-1.2024 (0.0253)	-1.1992 (0.0173)	-1.1989 (0.0127)
$\widehat{a}_{1,N}$	0.5883 (0.1106)	0.5893 (0.0733)	0.5971 (0.0488)	0.5950 (0.0649)	0.6007 (0.0429)	0.6000 (0.0303)
$\widehat{a}_{2,N}$	-0.5025 (0.1446)	-0.4955 (0.0914)	-0.4946 (0.0629)	-0.5011 (0.0785)	-0.4977 (0.0524)	-0.5006 (0.0382)
$\widehat{b}_{1,N}$	-0.2984 (0.1602)	-0.2991 (0.1010)	-0.3056 (0.0703)	-0.3129 (0.1196)	-0.3136 (0.0812)	-0.3091 (0.0578)
$\widehat{b}_{2,N}$	0.4051 (0.1836)	0.3981 (0.1145)	0.3942 (0.0803)	0.4028 (0.1253)	0.4016 (0.0842)	0.4041 (0.0600)
$\widehat{g}_N(\cdot)$	0.1996 (0.0522)	0.1616 (0.0365)	0.1186 (0.0236)	0.1531 (0.0305)	0.1151 (0.0194)	0.0936 (0.0137)

correlation structure in the sense that it still leads to consistent estimate of the mean component, although there is a reduction of efficiency in estimating the mean component parameters.

Example 2. Consider the model

$$Y_{i,j} = \sum_{s=1}^{5} X_{i,j,s}\beta_s + g(t_{i,j}) + \varepsilon_{i,j},$$

where $\{X_{i,j,1}, X_{i,j,2}, \varepsilon_{i,j}$ are as in Example 1, $X_{i,j,3} \sim \text{Bernoulli}(1, 0.5)$, $X_{i,j,4} = \cos(t_{i,j} - 0.5) + \varrho_{i,j,4}$ and $X_{i,j,5} = \exp(\sin(t_{i,j})) + \varrho_{i,j,5}$ with $\varrho_{i,j,4} \sim t(2)$ and $\varrho_{i,j,5} \sim N(0, 1)$. Let $\beta = (0.5, -0.8, 1.5, 0.2, -1.2)^\top$ and other notations are as in Example 1. The empirical means and standard deviations of parametric and nonparametric components based on 1,000 replications are reported in Table 2.

Table 3. The mean value (mean) and the average empirical standard deviation (std) of the estimates are calculated based on 1,000 replications. The italicized columns give the finite sample performance under the misspecified model that ignores the quadratic components.

	m	5						10					
	n	100		200		300		100		200		300	
$\check{\beta}_{1,N}$	mean	-1.1986	*-1.2000*	-1.2013	*-1.1988*	-1.2004	*-1.2002*	-1.1996	*-1.2003*	-1.2008	*-1.1997*	-1.1996	*-1.1992*
	std	0.0396	*0.0405*	0.0272	*0.0292*	0.0224	*0.0211*	0.0267	*0.0291*	0.0166	*0.0180*	0.0147	*0.0145*
$\widehat{\beta}_{1,N}$	mean	-1.1995	*-1.2005*	-1.2015	*-1.2007*	-1.2004	*-1.1998*	-1.1997	*-1.1998*	-1.2005	*-1.2000*	-1.1998	*-1.1997*
	std	0.0227	*0.0224*	0.0159	*0.0161*	0.0130	*0.0122*	0.0154	*0.0153*	0.0095	*0.0103*	0.0084	*0.0086*
$\check{\beta}_{2,N}$	mean	0.6941	*0.6980*	0.6963	*0.7009*	0.6988	*0.6999*	0.7022	*0.6977*	0.6985	*0.7015*	0.7006	*0.6995*
	std	0.0775	*0.0786*	0.0531	*0.0553*	0.0455	*0.0424*	0.0509	*0.0521*	0.0372	*0.0365*	0.0284	*0.0301*
$\widehat{\beta}_{2,N}$	mean	0.6978	*0.7003*	0.6969	*0.7015*	0.6995	*0.6999*	0.7020	*0.6975*	0.7001	*0.6999*	0.7002	*0.7000*
	std	0.0458	*0.0461*	0.0310	*0.0312*	0.0260	*0.0251*	0.0295	*0.0290*	0.0201	*0.0207*	0.0162	*0.0164*
$\widehat{g}_N(\cdot)$	Mean	0.1324	*0.1343*	0.1000	*0.0975*	0.0831	*0.0811*	0.1038	*0.1072*	0.0750	*0.0754*	0.0624	*0.0631*
	Std	0.0405	*0.0464*	0.0315	*0.0319*	0.0256	*0.0239*	0.0339	*0.0351*	0.0224	*0.0231*	0.0175	*0.0185*
$\widehat{g}_N^{TS}(\cdot)$	Mean	0.1114	*0.1139*	0.0839	*0.0830*	0.0704	*0.0688*	0.0927	*0.0949*	0.0673	*0.0680*	0.0557	*0.0571*
	Std	0.0392	*0.0419*	0.0294	*0.0300*	0.0239	*0.0230*	0.0332	*0.0330*	0.0227	*0.0230*	0.0174	*0.0186*

Table 4. Results of the order determination for error AR process.

	m	15				20			
	n	50	100	150	200	50	100	150	200
Under		198	133	52	10	233	221	61	7
Correct		321	550	673	802	474	585	790	868
Over		481	317	275	188	293	194	149	125

We observe that all the estimations are unbiased with small standard deviations, and the estimated function $\widehat{g}_N(\cdot)$ has similar performance for the different values of h_N.

Example 3. The data were generated from
$Y_{i,j} = X_{i,j}^{\top}\beta + g(t_{i,j}) + \varepsilon_{i,j}$, $\varepsilon_{i,j} = \phi(d_{i,j,1})\varepsilon_{i,j-1} + e_{i,j}, i = 1, \ldots, n, j = 2, \ldots, m_i$,
where $\beta = (-1.2, 0.7)^{\top}$, $X_{i,j,1} = 1 + t_{i,j}^2 + \varrho_{i,j,1}$, $X_{i,j,2} \sim$ Bernoulli$(1, 0.5)$ and $g(t_{i,j}) = t_{i,j}\sin(2\pi t_{i,j})$ with $t_{i,j} \sim U(0,1)$, the $\varrho_{i,j,1}$ and $e_{i,j}$ were$N(0,1)$. We considered generation processes $\phi(d_{i,j,1}) \equiv 0.8$ and $\phi(d_{i,j,1}) = 0.8 - 0.5d_{i,j,1} + 0.5d_{i,j,1}^2$ as examples. If the above function is misspecified as our proposed linear structure of $d_{i,j,1}$, the resulting estimates are reported in Table 3 in the form of italics. From that, all the estimates for the mean component are consistent but with larger standard deviations (or standard errors) which will bring efficiency loss.

Example 4. The data were generated as in Example 1, but with

$$\varepsilon_{i,j} = \sum_{k=1}^{10}(a_k + b_k d_{i,j,k})\varepsilon_{i,j-k} + e_{i,j},$$

Table 5. CD4 data analysis results: estimates of the regression parameters.

	Independence	Proposed	Fan (2007)	Leng (2010)	Li (2011)
β_1	0.0154 (0.0360)	-0.0030 (0.0146)	0.016 (0.032)	0.005 (0.030)	0.0182 (0.0019)
β_2	1.0181 (0.1908)	0.6796 (0.0646)	0.665 (0.152)	0.768 (0.130)	1.1185 (0.0075)
β_3	1.1167 (0.5352)	0.7926 (0.1204)	0.700 (0.358)	0.821 (0.345)	0.5402 (0.0309)
β_4	-0.0498 (0.0625)	0.0168 (0.0239)	0.011 (0.040)	0.044 (0.038)	-0.0021 (0.0027)
β_5	-0.0450 (0.0216)	-0.0675 (0.0085)	-0.034 (0.014)	-0.030 (0.014)	-0.0508 (0.0007)

with $a = (0.6, 0, 0, 0, 0, 0, -0.5, 0, 0, 0)^\top$, $b = (-0.3, 0, 0, 0, 0, 0, 0.4, 0, 0, 0)^\top$, and other notations defined in the same way. The determination of the AR order is then the specification of the signified autoregressive coefficient, and the selection results are in Table 4.

5.3. CD4 data

Since the structure of the data is completely unknown, we took an initial AR model with order $q = 4$, giving the model

$$\varepsilon_{i,j} = \sum_{k=1}^{4} (a_k + b_k d_{i,j,k})\varepsilon_{i,j-k} + e_{ij} \quad \text{with} \quad d_{i,j,k} = \text{YEAR}_{i,j} - \text{YEAR}_{i,j-1}.$$

The estimated parametric coefficients β_l, $l = 1, \ldots, 5$, and the corresponding standard errors are listed in Table 5. From Table 5, we can see that the effects of all covariates are significant. The effects of covariates AGE and DEPRESSION are negative and the effects of covariates SMOKE, DRUG, and PARTNERS are positive. Table 5 also lists the results obtained by Fan, Huang, and Li (2007), Leng, Zhang, and Pan (2010), and Li (2011). The effect of covariate AGE based on our method is negative and significant. However, the effects of covariate AGE based on the methods developed by Fan, Huang, and Li (2007) and Leng, Zhang, and Pan (2010) are not significant, while that based on Li (2011) is positive and significant. For the covariate SMOKE, the results obtained by our proposed method are comparable with those of the other studies. For the covariate DRUG, our estimated coefficient is positive, significant, and comparable with that from Li (2011). For the covariate PARTNERS, our estimated coefficient is positive and significant, while those based on Fan, Huang, and Li (2007), Leng, Zhang, and Pan (2010), and Li (2011) are not significant. For the covariate DEPRESSION, our estimated coefficient is negative and significant, which is comparable with those of the others. The standard errors of the estimated coefficients based on the proposed method tend to be smaller, indicating that there is gain in efficiency by using this approach for modeling the correlation structure in the errors.

We used the SCAD penalized approach of Fan and Li (1996) to determine the order of the error model. The error model with lag order 2 was selected. The fitting results based on the selected model are shown in the bottom of Table 6.

Table 6. CD4 data analysis results: estimates of the autoregressive coefficients in the error structure and corresponding 95% confidence intervals.

	Ignore Correlation			Include Correlation		
	Estimate	SE	CI	Estimate	SE	CI
Before order determination						
$\widehat{a_1}$	0.5198	0.0839	[0.3553, 0.6842]	0.5052	0.0805	[0.3475, 0.6629]
$\widehat{a_2}$	0.4443	0.1147	[0.2194, 0.6692]	0.4623	0.1115	[0.2439, 0.6808]
$\widehat{a_3}$	-0.0757	0.1022	[-0.2760, 0.1246]	-0.0937	0.0976	[-0.2851, 0.0977]
$\widehat{a_4}$	0.5909	0.1569	[0.2835, 0.8984]	0.5836	0.1538	[0.2823, 0.8850]
$\widehat{b_1}$	-0.0078	0.1375	[-0.2773, 0.2616]	0.0063	0.1348	[-0.2580, 0.2705]
$\widehat{b_2}$	-0.2793	0.1748	[-0.6220, 0.0633]	-0.2966	0.1709	[-0.6315, 0.0383]
$\widehat{b_3}$	0.1886	0.1362	[-0.0783, 0.4555]	0.2103	0.1320	[-0.0484, 0.4690]
$\widehat{b_4}$	-0.6142	0.1944	[-0.9952, -0.2332]	-0.6050	0.1900	[-0.9775, -0.2326]
After order determination						
$\widehat{a_1}$	0.5866	0.0522	[0.4842, 0.6890]	0.5839	0.0511	[0.4837, 0.6840]
$\widehat{a_2}$	0.3892	0.0755	[0.2412, 0.5372]	0.3887	0.0739	[0.2438, 0.5336]
$\widehat{b_1}$	-0.1206	0.0738	[-0.2650, 0.0239]	-0.1146	0.0722	[-0.2561, 0.0268]
$\widehat{b_2}$	-0.1999	0.0944	[-0.3850, -0.0149]	-0.1984	0.0924	[-0.3795, -0.0173]

The estimates of a_1 and a_2 are positive and significant, and those of b_1 and b_2 are negative,with that of b_2 significant. This suggests that the CD4 cell counts, after adjusting for the covariates within the same subject, are positively correlated, and that the correlation tends to decrease as the observed time distance increases. Taking the AR error structure into account leads to more efficient estimates with smaller standard errors.

Figure 2 shows the local linear and two-stage local linear estimates of the unknown function g. The 95% pointwise confidence bands of the local polynomial estimates, and the 95% pointwise confidence bands of two-stage local polynomial estimates are shown in Figure 2. The two-stage local linear estimate has the narrower pointwise confidence bands.

6. Concluding Remarks

We have introduced an irregular time AR error model for longitudinal analysis with irregular and possibly subject-specific observation times, and proposed a unified profile least squares approach to estimating the regression coefficients and the parameters in the error model. The proposed error model can be generalized in many different ways. For example, a nonparametric generalization of (2.4) is

$$\varepsilon_{i,j} = \sum_{k=1}^{q} c_k(d_{i,j,k})\varepsilon_{i,j-k} + e_{i,j}, j = q+1, \ldots, m_i, i = 1, \ldots, n,$$

where the c_k's are unknown functions. Model (2.4) can be viewed a linear approximation to this model. This type of model has been proposed in the context of time series analysis (Cai, Fan, and Yao (2000)) but it appears it has not been used in longitudinal studies.

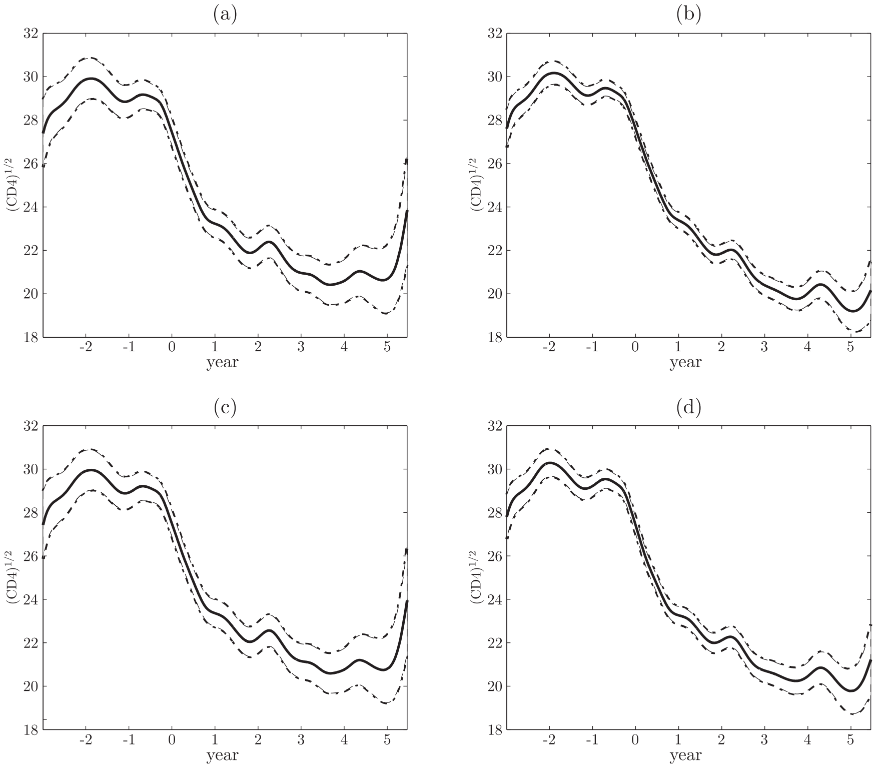

Figure 2. Plots of the estimated function $g(\cdot)$ and corresponding pointwise confidence band (a): that ignore the error correlated structure and fit the initial model; (b): that take the error correlated structure into account and fit the initial model; (c): that ignore the error correlated structure and fit the selected model; (d): that take the error correlated structure into account and fit the selected model.

The estimation method is based on the least-squares principle that is not robust to outliers in the data. He, Zhu, and Fung (2002) studied robust methods for estimation in partially linear regression models, but did not consider within-subject correlation. It would be interesting to combine the proposed method with theirs to construct robust and more efficient estimators for both the parametric and nonparametric components.

We have only considered that the number of covariates in the parametric component is fixed. Large data set and high dimensionality are characteristics of many contemporary problems. And, when the number of explanatory variables is large, it is more realistic to regard it as growing with sample size. Lam and Fan (2008) and Xie and Huang (2009) considered estimation in partially linear models with the number of covariates diverging with sample size. They focused on the cross sectional data structure. It would be interesting to investigate the statistical properties with diverging numbers of parameters in the present setting.

We have focused on irregular observations for the sparse longitudinal data with the observations of each individual fixed. In such disciplines as meteorology and economics, one has measurements to an extent that one speaks of dense longitudinal data. There has not been much attention paid to irregular autoregressive time series analysis of dense longitudinal data, and further study here would be interesting.

Acknowledgements

Bai and You's research was supported by grants from the National Natural Science Foundation of China (NSFC) (No.11071154 and No.11471203, respectively). Li's research was supported by SHUFE Graduate Innovation and Creativity Funds. The work is also partially supported by the Program for Changjiang Scholars and Innovative Research Team in University (IRT13077).

References

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coeffcient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95**, 941-956.

Chen, K. N. and Jin, Z. Z. (2006). Partial linear regression models for clustered data. *J. Amer. Statist. Assoc.* **101**, 195-204.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Fan, J. (2007). Variable screening in high-dimensional feature space. *Proceedings of the* 4*th International Congress of Chinese Mathematicians*, Vol. II, 735-747.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J., Huang, T. and Li, R. Z. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102**, 632-641.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Amer. Statist. Assoc.* **103**, 1520-1523.

He, X., Zhu, Z. Y. and Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579-590.

Huang, J. Z., Liu, L. and Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *J. Comput. Graph. Statist.* **16**, 189-209.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296-308.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191-219.

Lam, C. and Fan, J. (2008). Profile-Kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36**, 2232-2260.

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statist. Sinica* **20**, 167-181.

Leng, C. L., Zhang, W. P. and Pan, J. X. (2010). Semiparametric mean-covariance regression analysis for longitudinal data. *J. Amer. Statist. Assoc.* **105**, 181-193.

Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98**, 355-370.

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

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96**, 1045-1056.

Lin, X., Wang, N., Welsh, A., and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered data. *Biometrika* **91**, 177-193.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61**, 405-415.

Qin, G. Y., Zhu, Z. Y. and Fung, W. K. (2009). Robust sstimation of covariance parameters in partial linear model for longitudinal data. *J. Statist. Plann. Inference* **139**, 558-570.

Ruckstuhl, A. F., Welsh, A. and Carroll, R. J. (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statist. Sinica* **10**, 51-71.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92**, 1049-1062.

Salcedo, G., Porto, R., Roa, S. and Momo, F. (2012). A wavelet-based time-varying autoregressive model for non-stationary and irregular time series. *J. Appl. Stat.* **39**, 2313-2325.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90**, 43-52.

Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100**, 147-157.

Wei, Y. and He, X. (2006). Conditional Growth Charts (with discussions). *Ann. Statist.* **34**, 2069-2097.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90**, 831-844.

Xie, H. L. and Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

Zeger, S. L. and Diggle, P. G. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689-699.

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, P.R. China.

Key Laboratory of Mathematical Economics (SUFE), Ministry of Education of China.

E-mail: statbyang@mail.shufe.edu.cn

Department of Statistics and Actuarial Science, and Department of Biostatistics, University of Iowa, Iowa City, Iowa 52242, USA.

E-mail: jian-huang@uiowa.edu

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P.R. China.

E-mail: s0502062lirui@126.com

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P.R. China.

Key Laboratory of Mathematical Economics (SUFE), Ministry of Education of China.

E-mail: johnyou07@163.com

(Received March 2013; accepted January 2014)