New Local Estimation Procedure for Nonparametric Regression Function of Longitudinal Data

Article *in* Journal of the Royal Statistical Society Series B (Statistical Methodology) · March 2013

DOI: 10.1111/j.1467-9868.2012.01038.x · Source: PubMed

Weixin Yao
University of California, Riverside
88 PUBLICATIONS **902** CITATIONS

Runze Li
Pennsylvania State University
218 PUBLICATIONS **15,867** CITATIONS

New Local Estimation Procedure for Nonparametric Regression Function of Longitudinal Data

Weixin Yao and Runze Li

Abstract

This paper develops a new estimation of nonparametric regression functions for clustered or longitudinal data. We propose to use Cholesky decomposition and profile least squares techniques to estimate the correlation structure and regression function simultaneously. We further prove that the proposed estimator is as asymptotically efficient as if the covariance matrix were known. A Monte Carlo simulation study is conducted to examine the finite sample performance of the proposed procedure, and to compare the proposed procedure with the existing ones. Based on our empirical studies, the newly proposed procedure works better than the naive local linear regression with working independence error structure and the efficiency gain can be achieved in moderate-sized samples. Our numerical comparison also shows that the newly proposed procedure outperforms some existing ones. A real data set application is also provided to illustrate the proposed estimation procedure.

Key Words: Cholesky decomposition; Local polynomial regression; Longitudinal data; Profile least squares.

[1] Weixin Yao is Assistant Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506. Email: wxyao@ksu.edu. Runze Li is the corresponding author and Professor, Department of Statistics and the Methodology Center, The Pennsylvania State University Park, PA 16802-2111. Email: rli@stat.psu.edu. Li's research is supported by National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075 and a National Natural Science Foundation of China grant 11028103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

1 Introduction

For clustered or longitudinal data, we know that the data collected from the same subject at different times are correlated and that observations from different subjects are often independent. Therefore, it is of great interest to estimate the regression function incorporating the within-subject correlation to improve the estimation efficiency. This issue has been well studied for parametric regression models in the literature. See, for example, generalized method of moments (Hansen, 1982), the generalized estimating equation (Liang and Zeger, 1986), and quadratic inference function (Qu, Lindsay, and Li, 2000).

The parametric regression generally has simple and intuitive interpretations and provides a parsimonious description of the relationship between the response variable and its covariates. However, these strong assumption models may introduce modeling biases and lead to erroneous conclusions when there is model misspecification. In this article, we focus on the nonparametric regression model for longitudinal data. Suppose that $\{(x_{ij}, y_{ij}),\ i = 1, ..., n,\ j = 1, \ldots, J_i\}$ is a random sample from the following nonparametric regression model:

$$y_{ij} = m(x_{ij}) + \epsilon_{ij}, \tag{1}$$

where $m(\cdot)$ is a nonparametric smoothing function, and ϵ_{ij} is a random error. Here (x_{ij}, y_{ij}) is the j^{th} observation of the i^{th} subject or cluster. Thus, $(x_{ij}, y_{ij}),\ j = 1, \cdots, J_i$, are correlated. There has been substantial research interest in developing nonparametric estimation procedures for $m(\cdot)$ under the setting of clustered/longitudinal data. Lin and Carroll (2000) proposed the kernel GEE, an extension of the parametric GEE, for model (1) and showed that the kernel GEE works the best without incorporating the within-subject correlation. Wang (2003) proposed the marginal kernel method for longitudinal data and proved its efficiency by incorporating the true correlation structure. She also demonstrated that the marginal kernel method using the true correlation structure results in more efficient estimate than Lin and Carroll (2000)'s kernel GEE. Linton, Mammen, Lin, and Carroll (2003) proposed a two-stage estimator to incorporate the correlation by using a linear transformation to transform the correlated data model to uncorrelated data model if the working covariance matrix is known (up to some unknown parameters). They proved that their estimator has asymptotically smaller mean squared error than the regular working independence kernel estimator if the preliminary estimate is undersmoothed.

In this article, we propose a new procedure to estimate the correlation structure and regression function simultaneously, based on the Cholesky decomposition and profile least squares techniques. We derive the asymptotic bias and variance, and establish the asymptotic normality of the resulting estimator. We further conduct some theoretical comparison. We show that the newly proposed procedure is more efficient than Lin and Carroll (2000)'s kernel GEE. In addition, we prove that the proposed estimator is as asymptotically efficient as if the true covariance matrix were known a priori. Compared with the marginal kernel method of Wang (2003) and Linton et al. (2003), the newly proposed procedure does not require specifying a working correlation structure. This has appeal in practice because the true correlation structure is typically unknown. Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed procedure, and to compare the proposed procedure with the existing ones. Results from our empirical studies suggest that the newly proposed procedure performs better than the naive local linear regression and the efficiency gain can be achieved in moderate-sized samples. We further conduct Monte Carlo simulation to compare the newly proposed procedure with the procedures proposed by Lin and Carroll (2000), Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), and Chen, Fan, and Jin (2008). This numerical comparison shows that the newly proposed procedure may outperform the existing ones. We illustrate the proposed estimation method with an analysis of a real data set.

The remainder of this paper is organized as follows. In Section 2, we introduce the new estimation procedure based on the profile least squares and the Cholesky decomposition. We then provide the asymptotic results of the proposed estimator. Finally, we present numerical comparison and analysis of a real data example in Section 3. The proofs and the regularity conditions are given in the Appendix.

2 New estimation procedures

For ease of presentation, let us start with balanced longitudinal data. We will discuss how to use Cholesky composition to incorporate the within-subject correlation into the local estimation procedures for unbalanced longitudinal data in Section 2.2. Suppose $\{(x_{ij}, y_{ij}),\ i = 1, ..., n,\ j = 1, \ldots, J\}$ is a random sample from the model (1). In this paper, we will consider univariate x_{ij}. The newly proposed procedures are applicable for multivariate x_{ij}, but are practically less useful due to the "curse of dimensionality."

Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})^T$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$. Suppose $\text{cov}(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i) = \boldsymbol{\Sigma}$. Based on Cholesky decomposition, there exists a lower triangle matrix $\boldsymbol{\Phi}$ with diagonal ones such that

$$\text{cov}(\boldsymbol{\Phi}\boldsymbol{\epsilon}_i) = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T = \mathbf{D},$$

where $\mathbf{D}$ is a diagonal matrix. In other words, we have

$$\begin{aligned}&\epsilon_{i1} = e_{i1},\\ &\epsilon_{ij} = \phi_{j,1}\epsilon_{i,1} + \cdots + \phi_{j,j-1}\epsilon_{i,j-1} + e_{ij},\ i = 1, \ldots, n, j = 2, \ldots, J,\end{aligned}$$

where $\mathbf{e}_i = (e_{i1}, \ldots, e_{iJ})^T = \boldsymbol{\Phi}\boldsymbol{\epsilon}_i$, and $\phi_{j,l}$ is negative of $(j, l)-$element of $\boldsymbol{\Phi}$. Let $\mathbf{D} = \text{diag}(d_1^2, \ldots, d_J^2)$. Since $\mathbf{D}$ is a diagonal matrix, e'_{ij}s are uncorrelated and $\text{var}(e_{ij}) = d_j^2, j = 1, \ldots, J$. If $\{\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n\}$ were available, then we would work on the following partially linear model with uncorrelated error term e'_{ij}s:

$$\begin{aligned}&y_{i1} = m(x_{i1}) + e_{i1}\\ &y_{ij} = m(x_{ij}) + \phi_{j,1}\epsilon_{i,1} + \cdots + \phi_{j,j-1}\epsilon_{i,j-1} + e_{ij},\ i = 1, \ldots, n, j = 2, \ldots, J.\end{aligned} \tag{2}$$

However, in practice, ϵ_{ij} is not available, but it may be predicted by $\hat{\epsilon}_{ij} = y_{ij} - \hat{m}_I(x_{ij})$, where $\hat{m}_I(x_{ij})$ is a local linear estimate of $m(\cdot)$ based on model (1) pretending that the random error $\epsilon'_{ij}s$ are independent. As shown in Lin and Carroll (2000), $\hat{m}_I(x)$ under the working independence structure is a consistent estimate of $m(x)$.

Replacing $\epsilon_{ij}'s$ in (2) with $\hat{\epsilon}_{ij}'s$, we have

$$y_{ij} = m(x_{ij}) + \phi_{j,1}\hat{\epsilon}_{i,1} + \cdots + \phi_{j,j-1}\hat{\epsilon}_{i,j-1} + e_{ij},\ i = 1, \ldots, n, j = 2, \ldots, J. \tag{3}$$

Let $\mathbf{Y} = (y_{12}, \ldots, y_{1J}, \ldots, y_{nJ})^T$, $\mathbf{X} = (x_{12}, \ldots, x_{1J}, \ldots, x_{nJ})^T$, $\boldsymbol{\phi} = (\phi_{21}, \ldots, \phi_{J,J-1})^T$, $\mathbf{e} = (e_{12}, \ldots, e_{nJ})^T$, and $\hat{\mathbf{F}}_{ij} = \left\{0^T_{(j-2)(j-1)/2}, \hat{\epsilon}_{i,1}, \ldots, \hat{\epsilon}_{i,j-1}, 0^T_{(J-1)J/2-(j-1)j/2}\right\}^T$, where 0_k is the k-dimension column vector with all entries 0. Then we can rewrite the model (3) with the following matrix format:

$$\mathbf{Y} = m(\mathbf{X}) + \hat{\mathbf{F}}_a\boldsymbol{\phi} + \mathbf{e}, \tag{4}$$

where $m(\mathbf{X}) = \{m(x_{12}), \ldots, m(x_{1J}), \ldots, m(x_{nJ})\}^T$ and $\hat{\mathbf{F}}_a = (\hat{\mathbf{F}}_{12}, \ldots, \hat{\mathbf{F}}_{1J}, \ldots, \hat{\mathbf{F}}_{nJ})^T$. Let $\mathbf{Y}^* = \mathbf{Y} - \hat{\mathbf{F}}_a\boldsymbol{\phi}$. Then

$$\mathbf{Y}^* = m(\mathbf{X}) + \mathbf{e}. \tag{5}$$

Note that e_{ij}'s in $\mathbf{e}$ are uncorrelated. Therefore, if $\boldsymbol{\Sigma}$ and thus $\boldsymbol{\phi}$ is known, we can use the Cholesky decomposition to transfer the correlated data model (1) to the uncorrelated data model (5) with the new response $\mathbf{Y}^*$.

For partial linear model (4), various estimation methods have been proposed. In this paper, we will employ the profile least squares techniques (Fan and Li, 2004) to estimate $\boldsymbol{\phi}$ and $m(\cdot)$ in (4).

2.1 Profile least squares estimate

Noting that (5) is a one-dimension nonparametric model, given $\boldsymbol{\phi}$, one may employ existing linear smoothers, such as local polynomial regression (Fan and Gijbels, 1996) and smoothing splines (Gu, 2002) to estimate $m(x)$. Here, we employ the local linear regression.

Let

$$\mathbf{A}_{x_0} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \\ x_{12} - x_0 & \cdots & x_{1J} - x_0 & \cdots & x_{nJ} - x_0 \end{pmatrix}^T,$$

and

$$\mathbf{W}_{x_0} = \mathrm{diag}\{K_h(x_{12} - x_0)/\hat{d}_1^2, \ldots, K_h(x_{1J} - x_0)/\hat{d}_J^2, \ldots, K_h(x_{nJ} - x_0)/\hat{d}_J^2\},$$

where $K_h(t) = h^{-1}K(t/h)$, $K(\cdot)$ is a kernel function, and h is the bandwidth, and $\hat{d}_j$ is any consistent estimate of d_j, the standard deviation of e_{1j}. Denote by $\hat{m}(x_0)$ the local linear regression estimate of $m(x_0)$. Then

$$\hat{m}(x_0) = \hat{\beta}_0 = [1, 0](\mathbf{A}_{x_0}{}^T \mathbf{W}_{x_0} \mathbf{A}_{x_0})^{-1} \mathbf{A}_{x_0}{}^T \mathbf{W}_{x_0} \mathbf{Y}^*.$$

Note that $\hat{m}(x_0)$ is a linear combination of $\mathbf{Y}^*$. Let $\mathbf{S}_h(x_0) = [1, 0](\mathbf{A}_{x_0}{}^T \mathbf{W}_{x_0} \mathbf{A}_{x_0})^{-1} \mathbf{A}_{x_0}{}^T \mathbf{W}_{x_0}$. Then $\hat{m}(\mathbf{X})$ can be represented by

$$\hat{m}(\mathbf{X}) = \mathbf{S}_h(\mathbf{X})\mathbf{Y}^*,$$

where $\mathbf{S}_h(\mathbf{X})$ is a $(J - 1)n \times (J - 1)n$ smoothing matrix, depending on $\mathbf{X}$ and the bandwidth h only. Substituting $m(\mathbf{X})$ in (5) by $\hat{m}(\mathbf{X})$, we obtain the linear regression model:

$$\{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}\, \mathbf{Y} = \{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}\, \hat{\mathbf{F}}_a \boldsymbol{\phi} + \mathbf{e},$$

where $\mathbf{I}$ is the identity matrix. Let

$$\hat{\mathbf{G}} = \mathrm{diag}(\hat{d}_2^2, \ldots, \hat{d}_J^2, \ldots, \hat{d}_2^2, \ldots, \hat{d}_J^2).$$

Then, the profile least squares estimator for $\boldsymbol{\phi}$ is

$$\hat{\boldsymbol{\phi}}_p = \left[\hat{\mathbf{F}}_a^T \{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}^T \hat{\mathbf{G}}^{-1} \{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}\, \hat{\mathbf{F}}_a\right]^{-1} \hat{\mathbf{F}}_a^T \{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}^T \hat{\mathbf{G}}^{-1} \{\mathbf{I} - \mathbf{S}_h(\mathbf{X})\}\, \mathbf{Y}. \qquad (6)$$

Let $\hat{\mathbf{Y}}^* = \mathbf{Y} - \hat{\mathbf{F}}_a \hat{\boldsymbol{\phi}}_p$, then

$$\hat{\mathbf{Y}}^* = m(\mathbf{X}) + \mathbf{e}, \qquad (7)$$

and $e_{ij}'s$ are uncorrelated. Note that when we estimate the regression function $m(x)$, we can also include the observations from the first time point. Therefore, for simplicity of notation, when estimating $m(x)$, we assume that $\hat{\mathbf{Y}}^*$ consists of all observations with $\hat{y}_{i1}^* = y_{i1}$. Similar changes are used for all other notation when estimating $m(x)$ in (7).

Since e_{ij}'s in (7) are uncorrelated, we can use the conventional local linear regression estimator:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} (\hat{\mathbf{Y}}^* - \mathbf{A}_{x_0}\boldsymbol{\beta})^T \mathbf{W}_{x_0} (\hat{\mathbf{Y}}^* - \mathbf{A}_{x_0}\boldsymbol{\beta}).$$

Then the local linear estimate of $m(x_0)$ is $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p) = \hat{\beta}_0$.

Bandwidth selection. To implement the newly proposed estimation procedure, we need to specify bandwidths. We use local linear regression with the working independent correlation matrix to estimate $\hat{m}_I(\cdot)$. The plug-in bandwidth selector (Ruppert, Sheather, and Wand, 1995) was applied for the estimation of $\hat{m}_I(\cdot)$. Then we calculate $\hat{\epsilon}_{ij} = y_{ij} - \hat{m}_I(x_{ij})$, and further calculate the difference-based estimate for $\boldsymbol{\phi}$ (Fan and Li, 2004), denoted by $\hat{\boldsymbol{\phi}}_{dbe}$. Using $\hat{\boldsymbol{\phi}}_{dbe}$ in (5), we select a bandwidth for the proposed profile least squares estimator using the plug-in bandwidth selector.

2.2 Theoretical comparison

The following notation is used in the asymptotic results below. Let $\mathbf{F}_i = (\mathbf{F}_{i1}, \ldots, \mathbf{F}_{iJ})^T$, where

$$\mathbf{F}_{ij} = \left\{ 0^T_{(j-2)(j-1)/2}, \epsilon_{i,1}, \ldots, \epsilon_{i,j-1}, 0^T_{(J-1)J/2-(j-1)j/2} \right\}^T,$$

and

$$\mu_j = \int t^j K(t) dt, \text{ and } \nu_0 = \int K^2(t) dt.$$

Denote by $f_j(x)$ the marginal density of X_{1j}. The asymptotic results of the profile least squares estimators $\hat{\boldsymbol{\phi}}_p$ and $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p)$ are given in the following theorem, whose proof can be found in the Appendix.

Theorem 2.1. *Supposing the regularity conditions A1—A6 in the Appendix hold, under the assumption of* $cov(\boldsymbol{\epsilon}_i \mid \boldsymbol{x}_i) = \boldsymbol{\Sigma}$, *we have* *(a) the asymptotic distribution of $\hat{\boldsymbol{\phi}}_p$ in (6) is given by*

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}) \to N(0, \mathbf{V}^{-1}),$$

where

$$\mathbf{V} = \frac{1}{J-1}\sum_{j=2}^{J} E(\mathbf{F}_{1j}\mathbf{F}_{1j}^T)/d_j^2,$$

and $var(e_{1j}) = d_j^2$.

(b) the asymptotic distribution of $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p)$, conditioning on $\{x_{11}, \dots, x_{nJ}\}$, is given below

$$\sqrt{Nh}\{\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p) - m(x_0) - \frac{1}{2}\mu_2 m''(x_0)h^2\} \to N\left(0, \frac{\nu_0}{\tau(x_0)}\right),$$

where $N = nJ$ and

$$\tau(x_0) = \frac{1}{J}\sum_{j=1}^{J}\frac{f_j(x_0)}{d_j^2}.$$

Under the same assumption of Theorem 2.1, the asymptotic variance of the local linear estimate with working independence correlation structure (Lin and Carroll, 2000) is

$$(Nh)^{-1}\nu_0\left[\frac{1}{J}\sum_{j=1}^{J} f_j(x_0)\sigma_j^{-2}\right]^{-1},$$

where $\text{var}(\epsilon_{1j}) = \sigma_j^2$. Based on the property of Cholesky's decomposition, we know that

$$\sigma_1^2 = d_1^2 \text{ and } \sigma_j^2 \geq d_j^2, j = 2, \dots, J.$$

The equality only holds when $\text{cov}(\boldsymbol{\epsilon} \mid \mathbf{x}) = \boldsymbol{\Sigma}$ is a diagonal matrix. Note that $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p)$ has the same asymptotic bias as the working independence estimate of $m(x_0)$ (Lin and Carroll, 2000). Therefore, if within-subject observations are correlated (i.e., the covariance matrix $\boldsymbol{\Sigma}$ is not diagonal), then our proposed estimator $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p)$ is asymptotically more efficient than local linear estimator with the working independence correlation structure.

We next introduce how to use the Cholesky decomposition in (7) for unbalanced longitudinal data, and investigate the performance of the proposed procedure when a working covariance matrix is used for calculating $\hat{\mathbf{Y}}^*$. We shall show that the resulting local linear estimator is also consistent with any working positive definite covariance matrix, and further show that its asymptotic variance is minimized when the covariance structure is correctly specified.

For unbalanced longitudinal data, let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ_i})^T$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ_i})$, where J_i is the number of observations for ith subject or cluster. Denoted by $\text{cov}(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i) = \boldsymbol{\Sigma}_i$, which is a $J_i \times J_i$ matrix and may depend on $\mathbf{x}_i$. Based on the Cholesky decomposition, there exists a lower triangle matrix $\boldsymbol{\Phi}_i$ with diagonal ones such that

$$\text{cov}(\boldsymbol{\Phi}_i\boldsymbol{\epsilon}_i) = \boldsymbol{\Phi}_i\boldsymbol{\Sigma}_i\boldsymbol{\Phi}_i' = \mathbf{D}_i, \tag{8}$$

where $\mathbf{D}_i$ is a diagonal matrix. Let $\phi_{j,l}^{(i)}$ be the negative of $(j,l)-$element of $\boldsymbol{\Phi}_i$. Similar to (3), we have for $i = 1, \ldots, n$, $j = 2, \ldots, J_i$,

$$\begin{aligned} y_{i1} &= m(x_{i1}) + e_{i1} \\ y_{ij} &= m(x_{ij}) + \phi_{j,1}^{(i)}\hat{\epsilon}_{i,1} + \cdots + \phi_{j,j-1}^{(i)}\hat{\epsilon}_{i,j-1} + e_{ij}, \end{aligned} \tag{9}$$

where $\mathbf{e}_i = (e_{i1}, \ldots, e_{iJ_i})^T = \boldsymbol{\Phi}_i\boldsymbol{\epsilon}_i$. Since $\mathbf{D}_i$ is a diagonal matrix, e_{ij}'s are uncorrelated. Therefore, if $\boldsymbol{\Sigma}_i$ were known, one could adapt the newly proposed procedure for unbalanced longitudinal data.

Following the idea of the generalized estimating equation (GEE, Liang and Zeger, 1986), we replace $\boldsymbol{\Sigma}_i$ with a **working covariance matrix**, denoted by $\tilde{\boldsymbol{\Sigma}}_i$, since the true covariance matrix is unknown in practice. A parametric working covariance matrix can be constructed as in GEE, and a semiparametric working covariance matrix may also be constructed following Fan, Huang, and Li (2007). Let $\tilde{\boldsymbol{\Phi}}_i$ be the corresponding lower triangle matrix with diagonal ones such that

$$\tilde{\boldsymbol{\Phi}}_i\tilde{\boldsymbol{\Sigma}}_i\tilde{\boldsymbol{\Phi}}_i' = \tilde{\mathbf{D}}_i,$$

where $\tilde{\mathbf{D}}_i$ is a diagonal matrix. Let $\tilde{\phi}_{j,l}^{(i)}$ be the negative of (j,l)-element of $\tilde{\mathbf{\Phi}}_i$. Let $\tilde{y}_{i1} = y_{i1}$ and $\tilde{y}_{ij} = y_{ij} - \tilde{\phi}_{j,1}^{(i)}\hat{\epsilon}_{i,1} - \cdots - \tilde{\phi}_{j,j-1}^{(i)}\hat{\epsilon}_{i,j-1}$. Then our proposed new local linear estimate $\tilde{m}(x_0) = \tilde{\beta}_0$ is the minimizer of the following weighted least squares:

$$(\tilde{\beta}_0, \tilde{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n}\sum_{j=1}^{J_i} K_h(x_{ij} - x_0)\tilde{d}_{ij}^{-2}\left\{\tilde{y}_{ij} - \beta_0 - \beta_1(x_{ij} - x_0)\right\}^2, \tag{10}$$

where $\tilde{d}_{ij}^2$ is the jth diagonal element of $\tilde{\mathbf{D}}_i$.

The asymptotic behavior of $\tilde{m}(x_0)$ is given in Theorem 2.2. Following Lin and Carroll (2000) and Wang (2003), we assume that $J_i = J < \infty$ in order to simplify the presentation of the asymptotic results. Let $\boldsymbol{\phi}_i = (\phi_{21}^{(i)}, \ldots, \phi_{J,J-1}^{(i)})^T$ and $\tilde{\boldsymbol{\phi}}_i = (\tilde{\phi}_{21}^{(i)}, \ldots, \tilde{\phi}_{J,J-1}^{(i)})^T$.

Theorem 2.2. *Suppose the regularity conditions A1—A6 in the Appendix hold and cov*$(\boldsymbol{\epsilon}_i \mid \boldsymbol{x}_i) = \mathbf{\Sigma}_i$. *Let* $\tilde{m}(x_0)$ *be the solution of (10) using the working covariance matrix* $\tilde{\mathbf{\Sigma}}_i$.

(a) The asymptotic bias of $\tilde{m}(x_0)$ *is given by*

$$bias\{\tilde{m}(x_0)\} = \frac{1}{2}\mu_2 m''(x_0)h^2(1 + o_p(1))$$

and the asymptotic variance is given by

$$var\{\tilde{m}(x_0)\} = (Nh)^{-1}\frac{\nu_0\gamma(x_0)}{\tilde{\tau}^2(x_0)}(1 + o_p(1)),$$

where

$$\tilde{\tau}(x_0) = \frac{1}{J}\sum_{j=1}^{J} f_j(x_0)E(\tilde{d}_j^{-2} \mid X_j = x_0),$$

and

$$\gamma(x_0) = \frac{1}{J}\sum_{j=1}^{J} f_j(x_0)E\left\{(c_j^2 + d_j^2)\tilde{d}_j^{-4} \mid X_j = x_0\right\},$$

where c_j^2 *is the* j*th diagonal element of cov*$\left\{\boldsymbol{F}(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) \mid \boldsymbol{X}\right\}$.

(b) The asymptotic variance of $\tilde{m}(x_0)$ *is minimized only when* $\tilde{\mathbf{\Sigma}}_i = k\mathbf{\Sigma}_i$ *is correctly specified for*

a positive constant k. It can then be simplified to

$$var\{\tilde{m}(x_0)\} \approx (Nh)^{-1}\nu_0 \left\{ \frac{1}{J}\sum_{j=1}^{J} f_j(x_0)E(d_j^{-2} \mid X_j = x_0) \right\}^{-1}.$$

For balanced longitudinal data, if $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ for all i and does not depend on $\boldsymbol{X}$, then

$$var\{\tilde{m}(x_0)\} \approx (Nh)^{-1}\nu_0 \left\{ \frac{1}{J}\sum_{j=1}^{J} f_j(x_0)d_j^{-2} \right\}^{-1}. \tag{11}$$

Theorem 2.2 (a) implies that the leading term of asymptotic bias does not depend on the working covariance matrix. This is expected since the bias is caused by the approximation error of local linear regression. Theorem 2.2 (a) also implies that the resulting estimate is consistent for any positive definite working covariance matrix. Theorem 2.2 (b) implies that the asymptotic variance of $\tilde{m}(x_0)$ in (10) is minimized when the working correlation matrix is equal to the true correlation matrix. Comparing Theorem 2.1 (b) to the Theorem 2.2 (b), one knows that the proposed profile least square estimate $\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p)$ for balanced longitudinal data is as asymptotically efficient as if one knew the true covariance matrix.

It is of great interest to compare the performance of the proposed procedure with the existing ones in terms of asymptotic mean squared errors, which equals the summation of asymptotic variance and the square of asymptotic bias. As pointed out in Chen, Fan and Jin (2008), it is difficult to compare the performance of estimation procedures for longitudinal or clustered data based on local linear regression. For example, as shown in Wang (2003), her proposal has the minimal asymptotic variance. This has been further confirmed by the numerical comparison in Table 3 given in next section. However, the asymptotic bias term of Wang's proposal cannot be easily evaluated since the bias can only be expressed as the solution of a Fredholm-type equation. As a result, it is very difficult to evaluate the asymptotic mean squared errors of the procedure proposed in Wang (2003). From numerical comparison in Table 1 of Chen, Fan and Jin (2008), Wang's procedure has the minimal variance across all bandwidths used in the comparison, but the bias of Wang's procedure is slightly greater than that of other methods. As a result, her procedure is not always the best one in terms of mean integrated squared errors.

It is very difficult to compare the asymptotic variance given in Theorem 2.2 with that for existing ones under general settings. We shall provide a numerical comparison between the newly proposed method and existing ones proposed in Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), and Chen, Fan, and Jin (2008) in next section. It is possible to make some comparisons for some simple cases. For balanced longitudinal data with $J_i = J$, denote σ^{jj} to be the j-th diagonal element of Σ^{-1}. Then the asymptotic variance of Wang (2003)'s estimator can be written as $(Nh)^{-1}\nu_0\{J^{-1}\sum_{j=1}^{J}\sigma^{jj}f_j(x_0)\}^{-1}$. Using the definition of Cholesky's decomposition: $\mathbf{\Phi\Sigma\Phi}^T = \mathbf{D}$, it follows that $\sigma^{jj} = d_j^{-2} + \sum_{k=j+1}^{J} d_k^{-2}\phi_{kj}^2$, which implies that the asymptotic variance given Theorem 2.2 (b) is greater than that of the procedure proposed in Wang (2003). This motivates us to further improve the proposed procedure. It is known that the Cholesky's decomposition depends on the order of within-subject observations. Since we can estimate $f_j(x_0)$ by using a kernel estimate, assume that $f_j(x)$ is known for simplicity of presentation. We may estimate $\mathbf{D}$ using $\hat{\epsilon}_{ij} = y_{ij} - \hat{m}_I(x_{ij})$. This enables us to estimate the factor $J^{-1}\sum_{j=1}^{J} f_j(x_0)d_j^{-2}$ before implementing the proposed profile least squares procedure. Thus, for balanced longitudinal data and for $\mathbf{\Sigma}$ not depending on $\mathbf{X}$, we may change the order of within-subject observations (i.e, the order of js) such that $J^{-1}\sum_{k=1}^{J} f_{j_k}(x_0)\tilde{d}_{j_k}^{-2}$ is as large as possible with respect to the new order $\{j_1, \cdots, j_J\}$, where $\tilde{d}_{j_k}^2$'s are the diagonal elements of $\mathbf{D}$ in the corresponding Cholesky's decomposition. Based on our limited experience, we recommend to arrange the order so that $\tilde{d}_1^2 \geq \cdots \geq \tilde{d}_J^2$ (i.e., the diagonal elements of $\mathbf{D}$ from largest to the smallest). We will give a detailed demonstration of this strategy in Example 2.

3 Simulation results and real data application

In this section, we conduct a Monte Carlo simulation to assess the performance of the proposed profile least estimator, compare the newly proposed method with some existing ones, and illustrate the newly proposed procedure with an empirical analysis of a real data example.

Example 1. This example is designed to assess the finite sample performance of the proposed estimator for both balanced and unbalanced longitudinal data. In this example, data $\{(x_{ij}, y_{ij}), i = 1, \ldots, n, j = 1, \ldots, J_i\}$ are generated from the model

$$y_{ij} = 2\sin(2\pi x_{ij}) + \epsilon_{ij},$$

where $x_{ij} \sim U(0,1)$ and $\epsilon_{ij} \sim N(0,1)$. Let $\epsilon_i = (\epsilon_{i1} \ldots \epsilon_{iJ_i})^T$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ_i})^T$. We consider the following three cases:

Case I: $\epsilon_{ij}' s$ are independent.

Case II: $\text{cov}(\epsilon_{ij}, \epsilon_{ik}) = 0.6$, when $j \neq k$ and 1 otherwise.

Case III: $\mathbf{\Sigma}_i = \text{cov}(\boldsymbol{\epsilon}_i \mid \mathbf{x}_i)$ is AR(1) correlation structure with $\rho = 0.6$.

For balanced data case, we let $J_i = J = 6$, $i = 1, \ldots, n$. To investigate the effect of errors in estimating the covariance matrix, we compare the proposed profile least squares procedure with the oracle estimator using the true covariance matrix. The oracle estimator serves as a benchmark for the comparison. In this example, we also compare the newly proposed procedure with the local linear regression using the working independence correlation structure (Lin and Carroll, 2000). The sample size n is taken to be 30, 50, 100 and 400 to examine the finite sample performance of the proposed procedure. For each scenario, we conduct $1,000$ simulations.

Following Chen, Fan, and Jin (2008), we use the the mean integrated squared errors (MISE) defined below as a criterion for comparison:

$$\text{MISE}\{\hat{m}(\cdot)\} = \frac{1}{T}\sum_{t=1}^{T} \hat{D}_t, \tag{12}$$

where $T = 1000$, the number of simulations, $D_t \triangleq \int_{0.1}^{0.9} \{m(x) - \hat{m}_t(x)\}^2 dx$ is the integrated squared error for t^{th} simulation, $\hat{D}_t$ estimates D_t by replacing the integration with the summation over the grid points $x_g = 0.1 + 0.008g$ ($g = 0, \ldots, 100$), and $\hat{m}_t(x)$ is the estimate of $m(x)$ for t^{th} simulation. Table 1 depicts simulation results. In Table 1 and in the discussion below, "New" stands for the newly proposed procedure, and "Oracle" for the oracle estimator. Table 1 depicts the relative

MISE (RMISE), defined by the ratio of MISE of the two other estimators to that for the working independence method of Lin and Carroll (2000). Thus, if a RMISE is greater than 1, then the corresponding method performs better than the working independence method.

Table 1: Comparison of methods for different cases and sample sizes for **balanced** data of Example 1 based on 1,000 replicates. "New" stands for the newly proposed procedure, and "Oracle" for the oracle estimator. Bias: average of absolute values of biases at 101 grid points. SD: average of standard deviations at 101 grid points. RMISE: the relative MISE between two other estimators and the working independence method of Lin and Carroll (2000).

Case	Method		$n = 30$	$n = 50$	$n = 150$	$n = 400$
I	New	Bias	0.076	0.065	0.039	0.028
		SD	0.204	0.155	0.094	0.062
		RMISE	0.901	0.945	0.985	0.989
I	Oracle	Bias	0.077	0.065	0.039	0.028
		SD	0.190	0.149	0.093	0.062
		RMISE	1.000	1.000	1.000	1.000
II	New	Bias	0.067	0.055	0.035	0.023
		SD	0.212	0.162	0.099	0.060
		RMISE	1.155	1.194	1.278	1.362
II	Oracle	Bias	0.065	0.053	0.035	0.023
		SD	0.204	0.159	0.098	0.060
		RMISE	1.235	1.256	1.294	1.367
III	New	Bias	0.070	0.054	0.036	0.026
		SD	0.199	0.152	0.094	0.060
		RMISE	1.127	1.187	1.244	1.256
III	Oracle	Bias	0.069	0.054	0.035	0.026
		SD	0.190	0.149	0.094	0.060
		RMISE	1.223	1.232	1.266	1.266

Table 1 shows that the "New" and "Oracle" have smaller MISE than the "Independence" when the data are correlated (Cases II and III) and the efficiency gain can be achieved even for moderate sample size. For independent data (Case I), the "New" does not lose much efficiency for estimating the correlation structure when compared to the "Independence". Furthermore, Table 1 shows that when sample size is large, the "New" performs as well as the "Oracle", which uses the true correlation structure. The simulation results confirm the theoretical findings in Section 2.

Next we assess our proposed estimator for unbalanced longitudinal data. Let J_i, the number of observations for ith subject, be the uniform discrete random variable taking values among

$\{1, 2, \ldots, 12\}$. Since J_i can be different for each i, the data is unbalanced. To see how well the proposed method can incorporate the within-subject correlation, we first consider the situation in which the true within-subject correlation structure is known. We transform the correlated data to uncorrelated data using (8) and (9). Then we apply the existing local linear regression to the transformed data with weights d_{ij}^{-2}, where d_{ij} is the jth element of D_i and D_i is the diagonal matrix of Cholesky decomposition of Σ_i.

Table 2 shows the comparison results. Since for independence case, the newly proposed method will essentially provide the same result as working independence procedure, we only report the results for Cases II and III in top panel of Table 2, from which it can be seen that the newly proposed procedure works well and provides better estimator than the working independence procedure in terms of MISE for unbalanced data.

In practice, it may not be realistic to assume that the correlation structure is known. Thus, it is of interest to assess the performance of the proposed procedure when the correlation structure is mis-specified. To this end, we conduct a simulation by swapping the correlation structures of Cases II and III. That is, we use AR(1) correlation structure for Case II, and compound symmetric correlation structure for Case III. The corresponding simulation results are reported in the bottom panel of Table 2. As expected, the simulation result implies that the proposed procedure still have some efficiency gain over the working independence method, although the gain is not as much as that with true correlation structure.

Example 2. In this example, we compare the performance of the proposed procedure with those developed in Lin and Carroll (2000), Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), and Chen, Fan, and Jin (2008). Since Chen, Fan, and Jin (2008) has made a similar numerical comparison among those methods, we use the same simulation setting in Chen, Fan, and Jin (2008) to make a comparison in this example for fairness. Specifically, the data $\{(x_{ij}, y_{ij}), i = 1, \ldots, n, j = 1, \ldots, 4\}$ are generated from the model

$$y_{ij} = m(x_{ij}) + \epsilon_{ij},$$

Table 2: Comparison of methods for different cases and sample sizes for **unbalanced** data of Example 1 based on 1,000 replicates. Caption is the same as that in Table 1.

Case	Method		$n = 30$	$n = 50$	$n = 150$	$n = 400$
Correlation structure is **correctly** specified						
		Bias	0.058	0.047	0.032	0.024
II	New	SD	0.214	0.167	0.101	0.065
		RMISE	1.283	1.241	1.304	1.320
		Bias	0.058	0.047	0.032	0.024
II	Oracle	SD	0.213	0.166	0.101	0.065
		RMISE	1.294	1.247	1.311	1.322
		Bias	0.057	0.055	0.036	0.024
III	New	SD	0.190	0.151	0.092	0.059
		RMISE	1.220	1.246	1.245	1.239
		Bias	0.056	0.054	0.035	0.024
III	Oracle	SD	0.189	0.151	0.092	0.059
		RMISE	1.229	1.252	1.248	1.242
Correlation structure is **incorrectly** specified						
		Bias	0.062	0.050	0.035	0.025
II	New	SD	0.224	0.172	0.106	0.068
		RMISE	1.167	1.160	1.187	1.206
		Bias	0.062	0.058	0.037	0.026
III	New	SD	0.212	0.168	0.100	0.065
		RMISE	1.012	1.043	1.062	1.071

where $m(x) = 1 - 60x \exp\{-20x^2\}$, x_{i1} and x_{i3} are independently generated as $U[-1, 1]$, $x_{i2} = x_{i1}$, $x_{i4} = x_{i3}$, and errors $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i2}, \epsilon_{i4})$ are generated from the multivariate normal with mean 0, correlation 0.6, and marginal variances 0.04, 0.09, 0.01, and 0.16, respectively. The sample size $n = 150$ and the number of replicates is 1,000.

We first illustrate how to change the order of within-subject observations in order to obtain a smaller asymptotic variance of the resulting estimate. Note that $f_j(x)$s are the same for all js. Thus, we want to change the order of within-subject observations such that $J^{-1} \sum_{j=1}^{J} d_j^{-2}$ is as large as possible. Note that the diagonal elements of $\mathbf{\Sigma}^{-1}$ is (49.1071, 21.8254, 196.4286, 12.2768), and $J^{-1} \sum_{j=1}^{J} \sigma^{jj} = 69.9095$, and corresponding $\mathbf{D} = \text{diag}\{0.0400, 0.0576, 0.0055, 0.0815\}$. Thus, $(d_1^{-2}, d_2^{-2}, d_3^{-2}, d_4^{-2}) = (25.0000, 17.3611, 181.8182, 12.2768)$, and therefore $J^{-1} \sum_{j=1}^{J} d_j^{-2} = 59.1140$. Now we put the data from subject i in order as (x_{i4}, y_{i4}), (x_{i2}, y_{i2}), (x_{i1}, y_{i1}), (x_{i3}, y_{i3}). The corresponding $J^{-1} \sum_{j=1}^{J} \sigma^{jj}$ still equals 69.9095, while the corresponding $\mathbf{D}$ = diag{0.1600, 0.0576,

Table 3: Comparison of methods for different choices of bandwidth based on 1,000 replicates. Caption is the same as that in Table 1.

Method		$h = 0.02$	$h = 0.03$	$h = 0.04$	$h = 0.05$	$h = 0.06$
Wang's first	Bias	0.029	0.013	0.024	0.038	0.056
	SD	0.711	0.082	0.041	0.035	0.035
	RMISE	4.900	1.607	1.373	1.027	0.878
Wang's full	Bias	0.027	0.014	0.026	0.040	0.058
	SD	0.625	0.076	0.039	0.035	0.035
	RMISE	6.068	1.803	1.340	0.949	0.811
Chen-Jin(05)	Bias	0.036	0.012	0.021	0.033	0.048
	SD	1.217	0.109	0.049	0.042	0.040
	RMISE	1.217	0.786	1.223	1.117	1.064
Lin-Carroll(06)	Bise	0.031	0.012	0.022	0.034	0.049
	SD	0.778	0.084	0.041	0.035	0.035
	RMISE	3.461	1.158	1.481	1.195	1.057
Chen-Fan-Jin(08)	Bias	0.027	0.012	0.021	0.033	0.047
	SD	0.863	0.093	0.046	0.040	0.038
	RMISE	2.876	1.215	1.340	1.178	1.110
New	Bias	0.025	0.014	0.023	0.035	0.050
	SD	0.624	0.079	0.042	0.037	0.036
	RMISE	4.946	2.251	1.712	1.132	1.034

0.0220, 0.0051}, $(\tilde{d}_1^{-2}, \tilde{d}_2^{-2}, \tilde{d}_3^{-2}, \tilde{d}_4^{-2})$= (6.2500, 17.3611, 45.4545, 196.4286), and $J^{-1}\sum_{j=1}^{J} \tilde{d}_j^{-2} = 66.3736$. This implies that we can reduce asymptotic variance of the proposed least squares estimate via changing the order of within-subject observations. In our simulation, we will change the order of within-subject observation so that $\tilde{d}_1^2 \geq \tilde{d}_2^2 \geq \tilde{d}_3^2 \geq \tilde{d}_4^2$.

Following Chen, Fan, and Jin (2008), the curve estimate $\hat{m}(x)$ is computed on the grid points $x_g = -0.8+0.016g, g = 0, 1, \ldots, 100$, with various global fixed bandwidths. Seven different methods are considered: the working independence method of Lin and Carroll (2000), the one (first) step estimation of Wang (2003), the full iterated estimation of Wang (2003), the local linear method of Chen and Jin (2005), the closed-form method of Lin and Carroll (2006), the method of Chen, Fan, and Jin (2008), and the newly proposed method. The Epanechnikov kernel is used in all methods.

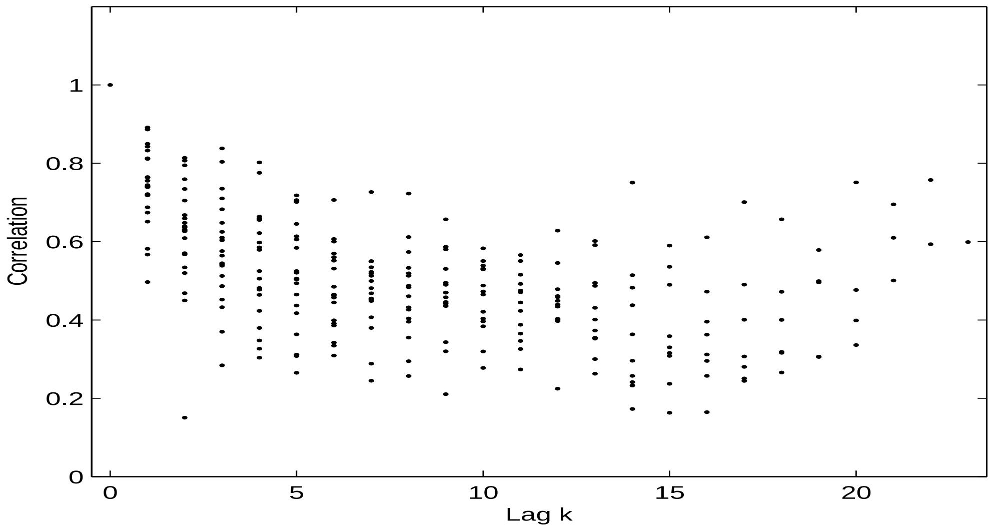
We use MISE defined in (12) to compare different methods. To calculate the MISE in this example, we set $D_t = \int_{-0.8}^{0.8}\{m(x) - \hat{m}_t(x)\}^2 dx$, and $\hat{D}_t$ estimates D_t by replacing the integration with the summation over the grid points $x_g = -0.8 + 0.016g$ $(g = 0, \ldots, 100)$.

Table 3 depicts RMISE, defined by the ratio of MISE of the six other estimators to that for the working independence method of Lin and Carroll (2000). To avoid duplicate efforts and make a fair comparison, the RMISE values for procedures developed in Wang (2003), Chen and Jin (2005), Lin and Carroll (2006), and Chen, Fan, and Jin (2008) are extracted from the Table 1 of Chen, Fan, and Jin (2008). From Table 3, we can see that the procedure proposed in Wang (2003) with full iteration has the smallest variance across all bandwidths, while its bias is greater than the newly proposed procedure. In terms of RMISE, the newly proposed method is comparable to the others for bandwidths 0.02, 0.05, and 0.06, and outperforms the others for bandwidth 0.03 and 0.04. Note that here the bandwidth 0.04 provides the smallest MISE for all methods.

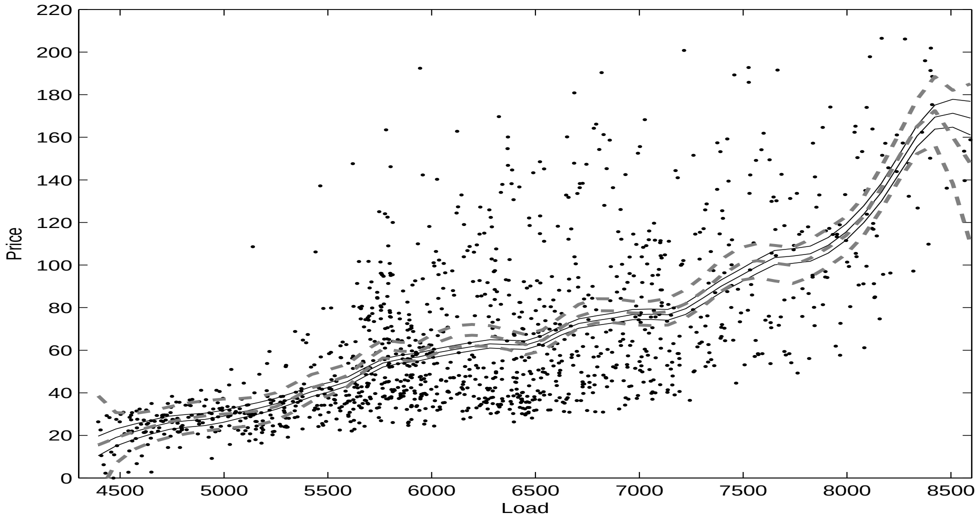
Example 3. In this example, we illustrate the proposed methodology with an empirical analysis of a data set collected from the website of Pennsylvania-New Jersey-Maryland Interconnections (PJM), the largest regional transmission organization (RTO) in the U.S. electricity market. The data set includes hourly electricity price and electricity load in the Allegheny Power Service district on each Wednesday of 2005. We studied the effect of the electricity load on the electricity price. As an illustration, we treated day as subject, and set the electricity price as the response variable and the electricity load as the predictor variable. Thus, the sample size n equals 52, and each subject has $J = 24$ observations. The scatter plot of observations is depicted in Figure 1 (b).

We first used local linear regression with working independence covariance matrix to estimate the regression. The plug-in bandwidth selector (Ruppert, Sheather, and Wand, 1995) yields a bandwidth of 89. The dashed lines in Figure 1 (b) are the resulting estimate along its 95% pointwise confidence interval. Based on the resulting estimate, we further obtain the residuals, and estimat the correlation between $\epsilon_{i,j}$ and $\epsilon_{i,j+k}$ for $j = 1, \cdots, 23$ and $1 \leq k \leq 24 - j$. The plot of estimated correlations is depicted in Figure 1 (a), which shows that the within-subject correlation is moderate. Thus, our proposed method may produce a more accurate estimate than the local linear regression, ignoring the within-subject correlation.

Next, we apply the newly proposed procedure for this data set. The bandwidth selected by the plug-in bandwidth selector equals 91. The solid curves in Figure 1 (b) are the fitted regression curves along with 95% pointwise confidence interval by the newly proposed procedure. Figure 1



(a)



(b)

Figure 1: (a) Plot of estimated correlation between $\epsilon_{i,j}$ and $\epsilon_{i,j+k}$ versus the lag k. For example, dots at $k = 1$ correspond to the correlations between $\epsilon_{i,j}$ and $\epsilon_{i,j+1}$ for $j = 1, 2, \cdots, 23$. (b) Scatter plot of observations and the plot of fitted regression curves. The solid curve is the fitted regression by the proposed method and the corresponding 95% point-wise confidence interval. The dash-dot curve is the local linear fit ignoring the within- subject correlation.

(b) shows that the newly proposed procedure provides a smaller confidence interval than the one ignoring the within-subject correlation. In addition, the fitted curve by the proposed method is much smoother than the working independence local linear fit, because the new method can borrow the information from more observations by taking the correlation into account. From Figure 1(b), it can be seen that the relationship between the electricity load and the electricity price is nonlinear. In general, the price increases as the load increases. However, the price change rate seems to remain almost constant when the load is from 4500-7000, but the price change rate is much larger when the load is larger than 7500.

4 Concluding remark

We have developed a new local estimation procedure for regression functions of longitudinal data. The proposed procedure uses the Cholesky decomposition and profile least squares techniques to estimate the correlation structure and regression function simultaneously. We demonstrate that the proposed estimator is as asymptotically efficient as an oracle estimator which uses the true covariance matrix to take into account the within-subject correlation. In this paper, we focus on nonparametric regression models. The proposed methodology can be easily adapted for other regression models, such as additive models and varying coefficient models. Such extensions are of great interest for future research.

APPENDIX: PROOFS

Define $\mathbf{B} = \hat{\mathbf{F}}_a - \mathbf{F}_a$. Since $\mathbf{G}$ can be estimated by a parametric rate, we will assume that $\mathbf{G}$ is known in our proof, without loss of generality. Our proofs use a strategy similar to that in Fan and Huang (2005). The following conditions are imposed to facilitate the proof and are adopted from Fan and Huang (2005). They are not the weakest possible conditions.

A1. The random variable x_{ij} has a bounded support Ω. Its density function $f_j(\cdot)$ is Lipschitz continuous and bounded away from 0 on its support. The x_{ij}'s are allowed to be correlated for different j's

A2. $m(\cdot)$ has the continuous second derivative in $x \in \Omega$.

A3. The kernel $K(\cdot)$ is a bounded symmetric density function with bounded support and satisfies the Lipschitz condition.

A4. $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

A5. There is a $s > 2$ such that $\mathrm{E}||F_{1j}||^s < \infty, \forall j$ and for some $\xi > 0$ such that $n^{1-2s^{-1}-2\xi}h \to \infty$.

A6. $\sup_{x\in\Omega}|\hat{m}_I(x) - m(x)| = o_p(n^{-1/4})$, where $\hat{m}_I(x)$ is obtained by local linear regression pretending that the data are independent and identically distributed (i.i.d.).

Lemma A.1. *Under Conditions A1—A6, it follows that*

(a) Let $\tilde{\mathbf{V}} = J^{-1}\sum_{j=1}^{J} E(\mathbf{F}_{1j}\mathbf{F}_{1j}^T)/d_j^2$. *Then*

$$\frac{1}{N}\hat{\mathbf{F}}_a^T\{I - \boldsymbol{S}_h(\boldsymbol{X})\}^T\mathbf{G}^{-1}\{I - \boldsymbol{S}_h(\boldsymbol{X})\}\hat{\mathbf{F}}_a \xrightarrow{P} \tilde{\mathbf{V}}.$$

(b) $N^{-1/2}\hat{\mathbf{F}}_a^T\{I - \boldsymbol{S}_h(\boldsymbol{X})\}^T\mathbf{G}^{-1}\{I - \boldsymbol{S}_h(\boldsymbol{X})\}\boldsymbol{m} = o_p(1)$ *and* $N^{-1/2}\hat{\mathbf{F}}_a^T\{I - \boldsymbol{S}_h(\boldsymbol{X})\}^T\mathbf{G}^{-1}\{I - \boldsymbol{S}_h(\boldsymbol{X})\}\boldsymbol{B\phi} = o_p(1)$.

(c) Let $\boldsymbol{e} = (e_{11}, \ldots, e_{nJ})^T$. *Then*

$$\sqrt{N}\left[\hat{\mathbf{F}}_a^T\{I - \boldsymbol{S}_h(\boldsymbol{X})\}^T\mathbf{G}^{-1}\{I - \boldsymbol{S}_h(\boldsymbol{X})\}\hat{\mathbf{F}}_a\right]^{-1}\hat{\mathbf{F}}_a^T\{I - \boldsymbol{S}_h(\boldsymbol{X})\}^T\mathbf{G}^{-1}\{I - \boldsymbol{S}_h(\boldsymbol{X})\}\boldsymbol{e} = N(0, \tilde{\mathbf{V}}^{-1}).$$

The proofs of Lemma A.1 is given in the earlier version of this paper, and is available from the authors upon request.

Proof of Theorem 2.1: Let us first show the asymptotic normality of $\hat{\boldsymbol{\phi}}_p$. According to the expression of $\hat{\boldsymbol{\phi}}_p$ in (6), we can break $\sqrt{N}(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi})$ into the sum of following three terms A, B and

C

$$A = \sqrt{N}\left[\left\{\hat{\mathbf{F}}_a^T(I-\mathbf{S}_h(\mathbf{X}))^T\mathbf{G}^{-1}(I-\mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\right\}^{-1}\hat{\mathbf{F}}_a^T\{I-\mathbf{S}_h(\mathbf{X})\}^T\mathbf{G}^{-1}\{I-\mathbf{S}_h(\mathbf{X})\}\mathbf{m}\right],$$
$$B = -\sqrt{N}\left[\left\{\hat{\mathbf{F}}_a^T(I-\mathbf{S}_h(\mathbf{X}))^T\mathbf{G}^{-1}(I-\mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\right\}^{-1}\hat{\mathbf{F}}_a^T\{I-\mathbf{S}_h(\mathbf{X})\}^T\mathbf{G}^{-1}\{I-\mathbf{S}_h(\mathbf{X})\}\mathbf{B}\boldsymbol{\phi}\right],$$
$$C = \sqrt{N}\left[\left\{\hat{\mathbf{F}}_a^T(I-\mathbf{S}_h(\mathbf{X}))^T\mathbf{G}^{-1}(I-\mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\right\}^{-1}\hat{\mathbf{F}}_a^T\{I-\mathbf{S}_h(\mathbf{X})\}^T\mathbf{G}^{-1}\{I-\mathbf{S}_h(\mathbf{X})\}\mathbf{e}\right].$$

From Lemma A.1(a) and (b), the asymptotic properties of these two terms lead to the conclusion that $A = o_p(1)$. Similarly, applying Lemma A.1 (a) and (b) on two product components of term B results in $B = o_p(1)$, as well. In addition, Lemma A.1 (c) states that term C converges to $N(0, \tilde{\mathbf{V}}^{-1})$. Noting that $\hat{\boldsymbol{\phi}}_p$ does not use the observations from the first time points, we should replace J by $J-1$ for $\hat{\boldsymbol{\phi}}_p$. Putting A, B, and C together, we get the asymptotic distribution of $\hat{\boldsymbol{\phi}}_p$.

Next we derive the asymptotic bias and variance of $\hat{m}(\cdot)$. Note that,

$$\begin{aligned}\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p) &= [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}(\mathbf{m}+\mathbf{e}+\mathbf{F}_a\boldsymbol{\phi}-\hat{\mathbf{F}}_a\hat{\boldsymbol{\phi}}_p)\\ &= [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}(\mathbf{m}+\mathbf{e})\{1+o_p(1)\}.\end{aligned}$$

Note that $\mathrm{E}\{\mathbf{e} \mid \mathbf{X}\} = 0$. Therefore,

$$\begin{aligned}\mathrm{bias}\{\hat{m}(x_0, \hat{\boldsymbol{\phi}}_p) \mid \mathbf{X}\} &= [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{m}\{1+o_p(1)\} - m(x_0)\\ &= [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\left\{\mathbf{m} - A_{x_0}[m(x_0), hm'(x_0)]^T\right\}\{1+o_p(1)\}.\end{aligned}$$

Similar to the arguments in Fan and Gijbels (1996, §3.7), we can prove that the asymptotic bias is $\frac{1}{2}m''(x_0)h^2\mu_2$.

In addition, note that

$$[1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0} = \frac{1}{N\tau(x_0)}[K_h(x_{11}-x_0)/d_1^2, \ldots, K_h(x_{nJ}-x_0)/d_J^2].$$

Therefore,

$$\begin{aligned}\text{var}\{\hat{m}(x_0,\hat{\boldsymbol{\phi}}_p) \mid \mathbf{X}\} &= [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\text{cov}(\mathbf{e})\mathbf{W}_{x_0}\mathbf{A}_{x_0}\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}[1,0]^T\{1+o_p(1)\}\\ &= \frac{1}{Nh\tau(x_0)}\int K^2(x)dx\{1+o_p(1)\}.\end{aligned}$$

As to the asymptotic normality,

$$\hat{m}(x_0,\hat{\boldsymbol{\phi}}_p) - \text{E}\{\hat{m}(x_0,\hat{\boldsymbol{\phi}}_p) \mid \mathbf{X}\} = [1,0]\{\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{A}_{x_0}\}^{-1}\mathbf{A}_{x_0}^T\mathbf{W}_{x_0}\mathbf{e}\{1+o_p(1)\}$$

Thus, conditioning on $\mathbf{X}$, the asymptotic normality can be established using the CLT since given j, $e_{ij}'s$ are identically and independent distributed with mean zero and variance d_j^2.

Proof of Theorem 2.2:

1.) The proof can be done in a similar to the proof of Theorem 2.1.

2.) When $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, we have $c_j^2 = 0$ and $\tilde{d}_j^2 = d_j^2$. Hence

$$\text{var}\{\tilde{m}(x_0) \mid \mathbf{X}\} \approx (Nh)^{-1}\nu_0\left\{\frac{1}{J}\sum_{j=1}^J f_j(x_0)\text{E}(d_j^{-2} \mid X_j = x_0)\right\}^{-1}.$$

By noting that

$$\gamma(x_0) \geq \left\{\frac{1}{J}\sum_{j=1}^J f_j(x_0)\text{E}(d_j^2\tilde{d}_j^{-4} \mid X_j = x_0)\right\}, \tag{13}$$

and

$$\left\{\sum_{j=1}^J f_j(x_0)\text{E}(d_j^2\tilde{d}_j^{-4} \mid X_j = x_0)\right\}\left\{\sum_{j=1}^J f_j(x_0)\text{E}(d_j^{-2} \mid X_j = x_0)\right\} \geq \left\{\sum_{j=1}^J f_j(x_0)\text{E}(\tilde{d}_j^{-2} \mid X_j = x_0)\right\}^2. \tag{14}$$

we can obtain the result. For result (13), the equality only holds when $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}$. For the second inequality (14), based on the Cauchy-Schwarz Inequality, the equality only holds when $\tilde{d}_j/d_j$ are all equal. Based on the Cholesky decomposition result, $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}$ and $\tilde{d}_j/d_j$ are all equal only when $\tilde{\boldsymbol{\Sigma}} = k\boldsymbol{\Sigma}$ and thus $\tilde{\boldsymbol{\Sigma}}_i = k\boldsymbol{\Sigma}_i$, for some constant k.

References

Chen, K., Fan, J., and Jin, Z. (2008). Design-adaptive minimax local linear regression for longitudinal/clustered data. *Statistica Sinica*, 18, 515-534.

Chen, K. and Jin, Z. (2005). Local polynomial regression analysis for clustered data. *Biometrika*, 92, 59-74.

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

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11, 1031-1057.

Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, 102, 632-641.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of American Statistical Association*, 99, 710-723.

Gu, C. (2002). *Smoothing Spline Anova Models.* Springer-Verlag, New York.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

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. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of American Statistical Association*, 95, 520-534.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of Royal Statistical Society*, B, 68, 69-88.

Linton, O. B., Mammen, E., Lin, X., and Carroll, R. J. (2003). Accounting for correlation in marginal longitudinal nonparametric regression. *Second Seattle Symposium on Biostatistics.*

Qu, A. and Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equatinos using quadratic inference functions. *Biometrika*, 87, 823-836.

Ruppert, D., Sheather S. J., and Wand M. P. (1995). An effective bandwidth selector for local least sqrares regression. *Journal of American Statistical Association*, 90, 1257-1270.

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