Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 1, *pp.* 123–138

New local estimation procedure for a non-parametric regression function for longitudinal data

Weixin Yao

Kansas State University, Manhattan, USA

and Runze Li

Pennsylvania State University, University Park, USA

[Received September 2010. Final revision February 2012]

Summary. The paper develops a new estimation of non-parametric 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 estimator proposed 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 procedure proposed, and to compare the procedure with the existing procedures. On the basis of our empirical studies, the newly proposed procedure works better than naive local linear regression with working independence error structure and the gain in efficiency can be achieved in moderate-sized samples. Our numerical comparison also shows that the newly proposed procedure outperforms some existing procedures. A real data set application is also provided to illustrate the estimation procedure proposed.

Keywords: Cholesky decomposition; Local polynomial regression; Longitudinal data; Profile least squares

1. Introduction

For clustered or longitudinal data, we know that the data that are 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 efficiency of estimation. This issue has been well studied for parametric regression models in the literature. See, for example, the generalized method of moments (Hansen, 1982), the generalized estimating equation (GEE) (Liang and Zeger, 1986) and quadratic inference function (Qu *et al.*, 2000).

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 modelling biases and lead to erroneous conclusions when there is model misspecification. In this paper, we focus on the non-parametric regression model for longitudinal data. Suppose that $\{(x_{ij}, y_{ij}), i=1,\ldots,n, j=1,\ldots,J_i\}$ is a random sample from the non-parametric regression model

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

Address for correspondence: Runze Li, Department of Statistics and the Methodology Center, Pennsylvania State University, University Park, PA 16802-2111, USA.
E-mail: rli@stat.psu.edu

1369–7412/13/75123

where $m(\cdot)$ is a non-parametric smoothing function, and ε_{ij} is a random error. Here (x_{ij}, y_{ij}) is the jth observation of the ith subject or cluster. Thus, $(x_{ij}, y_{ij}), j=1,\ldots,J_i$, are correlated. There has been substantial research interest in developing non-parametric estimation procedures for $m(\cdot)$ under the setting of clustered or 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 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 estimates than Lin and Carroll's (2000) kernel GEE. Linton *et al.* (2003) proposed a two-stage estimator to incorporate the correlation by using a linear transformation to transform the correlated data model into an 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 paper, 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 comparisons. We show that the newly proposed procedure is more efficient than Lin and Carroll's (2000) kernel GEE. In addition, we prove that the estimator proposed 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 the specification of 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 procedure proposed, and to compare the procedure proposed with the existing procedures. Results from our empirical studies suggest that the newly proposed procedure performs better than naive local linear regression and the gain in efficiency can be achieved in moderate-sized samples. We further conduct Monte Carlo simulation to compare the newly proposed procedure with the procedures that were proposed by Lin and Carroll (2000), Wang (2003), Chen and Jin (2005), Lin and Carroll (2006) and Chen *et al.* (2008). This numerical comparison shows that the newly proposed procedure may outperform the existing procedures. 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 asymptotic results for the estimator proposed. Finally, we present a numerical comparison and analysis of a real data example in Section 3. The proofs and the regularity conditions are given in Appendix A.

2. New estimation procedures

For ease of presentation, let us start with balanced longitudinal data. We shall 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 that $\{(x_{ij}, y_{ij}), i=1,\ldots,n, j=1,\ldots,J\}$ is a random sample from model (1). In this paper, we shall consider univariate x_{ij}. The newly proposed procedures are applicable for multivariate x_{ij} but are less useful practically because of the 'curse of dimensionality'.

Let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})^{\mathrm{T}}$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$. Suppose that $\mathrm{cov}(\varepsilon_i|\mathbf{x}_i) = \boldsymbol{\Sigma}$. On the basis of the Cholesky decomposition, there is a lower triangle matrix $\boldsymbol{\Phi}$ with 1s on the main diagonal such that

$$\mathrm{cov}(\boldsymbol{\Phi}\varepsilon_i) = \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}^{\mathrm{T}} = \mathbf{D},$$

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

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

where $\mathbf{e}_i = (e_{i1}, \ldots, e_{iJ})^{\mathrm{T}} = \boldsymbol{\Phi}\varepsilon_i$, and $\phi_{j,l}$ is the negative of the (j, l)-element of the $\boldsymbol{\Phi}$. Let $\mathbf{D} = \mathrm{diag}(d_1^2, \ldots, d_J^2)$. Since $\mathbf{D}$ is a diagonal matrix, the e'_{ij}s are uncorrelated and $\mathrm{var}(e_{ij}) = d_j^2$, $j = 1, \ldots, J$. If $\{\varepsilon_1, \ldots, \varepsilon_n\}$ were available, then we would work on the following partially linear model with uncorrelated error term e'_{ij}:

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

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

Replacing the ε'_{ij}s in model (2) with $\hat{\varepsilon}'_{ij}$s, we have

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

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

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

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

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

Note that the $e_{ij}s$ in $\mathbf{e}$ are uncorrelated. Therefore, if $\boldsymbol{\Sigma}$ and thus $\boldsymbol{\phi}$ are known, we can use the Cholesky decomposition to transform 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 shall employ the profile least squares techniques (Fan and Li, 2004) to estimate $\boldsymbol{\phi}$ and $m(\cdot)$ in approximation (4).

2.1. Profile least squares estimate

Noting that model (5) is a one-dimensional non-parametric model, given $\boldsymbol{\phi}$, we 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 local linear regression.

Let

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

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

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

Note that $\hat{m}(x_0)$ is a linear combination of $\mathbf{Y}^*$. Let $\mathbf{S}_h(x_0)=[1,0](\mathbf{A}_{x_0}^{\mathrm{T}}\mathbf{W}_{x_0}\mathbf{A}_{x_0})^{-1}\mathbf{A}_{x_0}^{\mathrm{T}}\mathbf{W}_{x_0}$. 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 model (5) by $\hat{m}(\mathbf{X})$, we obtain the linear regression model

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

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 ϕ is

$$\hat{\phi}_p=\{\hat{\mathbf{F}}_a^{\mathrm{T}}\,(\mathbf{I}-\mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\,\hat{\mathbf{G}}^{-1}\,(\mathbf{I}-\mathbf{S}_h(\mathbf{X}))\,\hat{\mathbf{F}}_a\}^{-1}\hat{\mathbf{F}}_a^{\mathrm{T}}\,(\mathbf{I}-\mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\,\hat{\mathbf{G}}^{-1}\,(\mathbf{I}-\mathbf{S}_h(\mathbf{X}))\,\mathbf{Y}. \tag{6}$$

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

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

and the e_{ij}'s are uncorrelated. 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 approximation (7).

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

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

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

2.1.1. 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 $\hat{m}_{\mathrm{I}}(\cdot)$. The plug-in bandwidth selector (Ruppert *et al.*, 1995) was applied for the estimation of $\hat{m}_{\mathrm{I}}(\cdot)$. Then we calculate $\hat{\varepsilon}_{ij}=y_{ij}-\hat{m}_{\mathrm{I}}(x_{ij})$, and further we calculate the difference-based estimate for ϕ (Fan and Li, 2004), denoted by $\hat{\phi}_{\mathrm{dbe}}$. Using $\hat{\phi}_{\mathrm{dbe}}$ in model (5), we select a bandwidth for the proposed profile least squares estimator by using the plug-in bandwidth selector.

2.2. Theoretical comparison

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

$$\mathbf{F}_{ij}=(\mathbf{0}_{(j-2)(j-1)/2}^{\mathrm{T}},\varepsilon_{i,1},\ldots,\varepsilon_{i,j-1},\mathbf{0}_{(J-1)J/2-(j-1)j/2}^{\mathrm{T}})^{\mathrm{T}},$$

and

$$\mu_j = \int t^j K(t)\,\mathrm{d}t,$$

$$\nu_0 = \int K^2(t)\,\mathrm{d}t.$$

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

Theorem 1. Supposing that the regularity conditions 1–6 in Appendix A hold, under the assumption of $\mathrm{cov}(\varepsilon_i|\mathbf{X}_i) = \mathbf{\Sigma}$, we have

(a) the asymptotic distribution of $\hat{\phi}_p$ in estimator (6) is given by

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

where

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

and $\mathrm{var}(e_{1j}) = d_j^2$, and

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

$$\{\hat{m}(x_0, \hat{\phi}_p) - m(x_0) - \tfrac{1}{2}\mu_2 m''(x_0)h^2\}\sqrt{(Nh)} \rightarrow N\{0, \nu_0/\tau(x_0)\},$$

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 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 $\mathrm{var}(\varepsilon_{1j}) = \sigma_j^2$. On the basis of the property of Cholesky's decomposition, we know that

$$\sigma_1^2 = d_1^2 \text{ and } \sigma_j^2 \geqslant d_j^2, \qquad j = 2, \ldots, J.$$

The equality holds only when $\mathrm{cov}(\varepsilon|\mathbf{x}) = \mathbf{\Sigma}$ is a diagonal matrix. Note that $\hat{m}(x_0, \hat{\phi}_p)$ 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 $\mathbf{\Sigma}$ is not diagonal), then our proposed estimator $\hat{m}(x_0, \hat{\phi}_p)$ is asymptotically more efficient than the local linear estimator with the working independence correlation structure.

We next introduce how to use the Cholesky decomposition in model (7) for unbalanced longitudinal data, and we 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 we further show that its asymptotic variance is minimized when the covariance structure is correctly specified.

For unbalanced longitudinal data, let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ_i})^{\mathrm{T}}$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ_i})$, where J_i is the number of observations for the ith subject or cluster. Denote $\mathrm{cov}(\varepsilon_i|\mathbf{x}_i) = \mathbf{\Sigma}_i$, which is a

$J_i \times J_i$ matrix and may depend on $\mathbf{x}_i$. On the basis of the Cholesky decomposition, there is a lower triangle matrix $\mathbf{\Phi}_i$ with diagonal 1s such that

$$\mathrm{cov}(\mathbf{\Phi}_i \varepsilon_i) = \mathbf{\Phi}_i \mathbf{\Sigma}_i \mathbf{\Phi}_i' = \mathbf{D}_i, \tag{8}$$

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

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

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

Following the idea of the GEE (Liang and Zeger, 1986), we replace $\mathbf{\Sigma}_i$ with a *working covariance matrix*, which is denoted by $\tilde{\mathbf{\Sigma}}_i$, since the true covariance matrix is unknown in practice. A parametric working covariance matrix can be constructed as in the GEE, and a semiparametric working covariance matrix may also be constructed following Fan *et al.* (2007). Let $\tilde{\mathbf{\Phi}}_i$ be the corresponding lower triangle matrix with 1s on the main diagonal such that

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

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

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

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

The asymptotic behaviour of $\tilde{m}(x_0)$ is given in theorem 2. Following Lin and Carroll (2000) and Wang (2003), we assume that $J_i = J < \infty$ simplify the presentation of the asymptotic results. Let $\phi_i = (\phi_{21}^{(i)}, \ldots, \phi_{J,J-1}^{(i)})^{\mathrm{T}}$ and $\tilde{\phi}_i = (\tilde{\phi}_{21}^{(i)}, \ldots, \tilde{\phi}_{J,J-1}^{(i)})^{\mathrm{T}}$.

Theorem 2. Suppose that the regularity conditions 1–6 in Appendix A hold and $\mathrm{cov}(\varepsilon_i | \mathbf{x}_i) = \mathbf{\Sigma}_i$. Let $\tilde{m}(x_0)$ be the solution of equation (10) by using the working covariance matrix $\tilde{\mathbf{\Sigma}}_i$.

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

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

and the asymptotic variance is given by

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

where

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

and

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

where c_j^2 is the jth diagonal element of $\mathrm{cov}\{\mathbf{F}(\tilde{\phi} - \phi) | \mathbf{X}\}$.

(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

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

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

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

Theorem 2, part (a), implies that the leading term of the 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, part (a), also implies that the resulting estimate is consistent for any positive definite working covariance matrix. Theorem 2, part (b), implies that the asymptotic variance of $\tilde{m}(x_0)$ in equation (10) is minimized when the working correlation matrix is equal to the true correlation matrix. Comparing theorem 1, part (b), with theorem 2, part (b), we know that the proposed profile least square estimate $\hat{m}(x_0, \hat{\phi}_p)$ for balanced longitudinal data is as asymptotically efficient as if we knew the true covariance matrix.

It is of great interest to compare the performance of the proposed procedure with the existing procedures in terms of the asymptotic mean-squared error, which equals the sum of the asymptotic variance and the square of the asymptotic bias. As pointed out in Chen *et al.* (2008), it is difficult to compare the performance of estimation procedures for longitudinal or clustered data on the basis of 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 the 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 that was proposed in Wang (2003). From a numerical comparison in Table 1 of Chen *et al.* (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 in terms of mean integrated squared errors.

It is very difficult to compare the asymptotic variance given in theorem 2 with that for existing variances under general settings. We shall provide a numerical comparison between the newly proposed method and existing procedures proposed in Wang (2003), Chen and Jin (2005), Lin and Carroll (2006) and Chen *et al.* (2008) in the next section. It is possible to make some comparisons for some simple cases. For balanced longitudinal data with $J_i = J$, denote σ^{jj} as the jth diagonal element of Σ^{-1}. Then the asymptotic variance of Wang's (2003) 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, $\mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{D}$, 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 in theorem 2, part (b), is greater than that of the procedure proposed in Wang (2003). This motivates us to improve the proposed procedure further. It is known that the Cholesky decomposition depends on the order of within-subject

Table 1. Comparison of methods for various cases and sample sizes for the *balanced* data of example 1 based on 1000 replicates†

Case	*Method*	*Parameter*	*Results for the following values of n:*			
			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

†'New' stands for the newly proposed procedure, and 'oracle' for the oracle estimator. Bias is the average of absolute values of biases at 101 grid points. SD is the average of the standard deviations at 101 grid points. RMISE is the relative MISE between two other estimators and the working independence method of Lin and Carroll (2000).

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 $\mathbf{D}$ by using $\hat{\varepsilon}_{ij} = y_{ij} - \hat{m}_I(x_{ij})$. This enables us to estimate the factor $J^{-1}\Sigma_{j=1}^{J} f_j(x_0)d_j^{-2}$ 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 $J^{-1}\Sigma_{k=1}^{J} f_{j_k}(x_0)\tilde{d}_{j_k}^{-2}$is as large as possible with respect to the new order $\{j_1, \ldots, j_J\}$, where the $\tilde{d}_{j_k}^2$s are the diagonal elements of $\mathbf{D}$ in the corresponding Cholesky decomposition. On the basis of our limited experience, we recommend arranging the order so that $\tilde{d}_1^2 \geqslant \ldots \geqslant \tilde{d}_J^2$ (i.e. the diagonal elements of $\mathbf{D}$ from largest to the smallest). We shall 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 profile least squares estimator proposed, compare the newly proposed method with some existing methods and illustrate the newly proposed procedure with an empirical analysis of a real data example.

3.1. 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

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

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

(a) case I, the $\varepsilon'_{ij}s$ are independent;
(b) case II, $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{ik})$ equals 0.6, when $j \neq k$ and 1 otherwise;
(c) case III, $\mathbf{\Sigma}_i = \mathrm{cov}(\varepsilon_i|\mathbf{x}_i)$ is an auto-regressive AR(1) correlation structure with $\rho = 0.6$.

For the 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 profile least squares procedure proposed with the oracle estimator by 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 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 procedure proposed. For each scenario, we conduct 1000 simulations.

Following Chen *et al.* (2008), we use the the mean integrated squared errors MISE defined below as a criterion for comparison:

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

where $T = 1000$, the number of simulations,

$$D_t \triangleq \int_{0.1}^{0.9} \{m(x) - \hat{m}_t(x)\}^2 \,\mathrm{d}x$$

is the integrated squared error for the tth simulation, $\hat{D}_t$ estimates D_t by replacing the integration with 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 the tth 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, which is defined by the ratio of MISE for the two other estimators to that for the working independence method of Lin and Carroll (2000). Thus, if RMISE >1, then the corresponding method performs better than the working independence method.

Table 1 shows that the new and oracle methods have smaller MISE than the independence model when the data are correlated (cases II and III) and the gain in efficiency can be achieved even for moderate sample size. For independent data (case I), the new method does not lose much efficiency for estimating the correlation structure when compared with the independence model. Furthermore, Table 1 shows that, when the sample size is large, the new method performs as well as the oracle method, 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 the 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 are unbalanced. To see how well the method proposed 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 by using expressions (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 the Cholesky decomposition of Σ_i.

Table 2. Comparison of methods for various cases and sample sizes for the *unbalanced* data of example 1 based on 1000 replicates†

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

†The methods and definitions of the parameters are the same as for Table 1.

Table 2 shows the comparison results. Since, for the independence case, the newly proposed method will essentially provide the same result as the working independence procedure, we report only the results for cases II and III in the top panel of Table 2, from which it can be seen that the newly proposed procedure works well and provides a 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 procedure proposed when the correlation structure is misspecified. For this, we conduct a simulation by swapping the correlation structures of cases II and III, i.e. we use an 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 procedure proposed still has some gain in efficiency over the working independence method, although the gain is not as much as that with true correlation structure.

3.2. Example 2

In this example, we compare the performance of the procedure proposed with those developed in Lin and Carroll (2000), Wang (2003), Chen and Jin (2005), Lin and Carroll (2006) and Chen *et al.* (2008). Since Chen *et al.* (2008) made a similar numerical comparison between those methods, we use the same simulation setting as in Chen *et al.* (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

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

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 $(\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3},\varepsilon_{i4})$ are generated from the multivariate normal distribution 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 1000.

We first illustrate how to change the order of within-subject observations to obtain a smaller asymptotic variance of the resulting estimate. Note that the $f_j(x)$s are the same for all js. Thus, we want to change the order of within-subject observations such that $J^{-1}\Sigma_{j=1}^{J}d_j^{-2}$ is as large as possible. Note that the diagonal elements of $\mathbf{\Sigma}^{-1}$ are (49.1071, 21.8254, 196.4286, 12.2768), and $J^{-1}\Sigma_{j=1}^{J}\sigma^{jj}=69.9095$, and the 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 $J^{-1}\Sigma_{j=1}^{J}d_j^{-2}=59.1140$. 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 $J^{-1}\Sigma_{j=1}^{J}\sigma^{jj}$ still equals 69.9095, whereas the corresponding $\mathbf{D}=\text{diag}(0.1600, 0.0576, 0.0220, 0.0051)$, $(\tilde{d}_1^{-2},\tilde{d}_2^{-2},\tilde{d}_3^{-2},\tilde{d}_4^{-2})=(6.2500, 17.3611, 45.4545, 196.4286)$ and $J^{-1}\Sigma_{j=1}^{J}\tilde{d}_j^{-2}=66.3736$. This implies that we can reduce the asymptotic variance of the least squares estimate proposed via changing the order of within-subject observations. In our simulation, we shall change the order of within-subject observation so that $\tilde{d}_1^2\geqslant\tilde{d}_2^2\geqslant\tilde{d}_3^2\geqslant\tilde{d}_4^2$.

Following Chen *et al.* (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 *et al.* (2008) and the newly proposed method. The Epanechnikov kernel is used in all the methods.

We use MISE defined in expression (12) to compare methods. To calculate MISE in this example, we set

$$D_t=\int_{-0.8}^{0.8}\{m(x)-\hat{m}_t(x)\}^2\,\mathrm{d}x,$$

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, which is 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 effort and to make a fair comparison, the RMISEs for the procedures that were developed in Wang (2003), Chen and Jin (2005), Lin and Carroll (2006) and Chen *et al.* (2008) have been extracted from Table 1 of Chen *et al.* (2008). From Table 3, we can see that the procedure that was proposed in Wang (2003) with full iteration has the smallest variance across all bandwidths, whereas its bias is greater than that of the newly proposed procedure. In terms of RMISE, the newly proposed method is comparable with the others for bandwidths 0.02, 0.05 and 0.06, and outperforms the others for bandwidths 0.03 and 0.04. Note that here the bandwidth 0.04 provides the smallest MISE for all methods.

3.3. Example 3

In this example, we illustrate the methodology proposed with an empirical analysis of a data set that was collected from the Web site of Pennsylvania–New Jersey–Maryland Interconnections, which is the largest regional transmission organization in the US electricity market. The data set includes hourly electricity price and electricity load in the Allegheny Power Service

Table 3. Comparison of methods for various choices of bandwidth based on 1000 replicates†

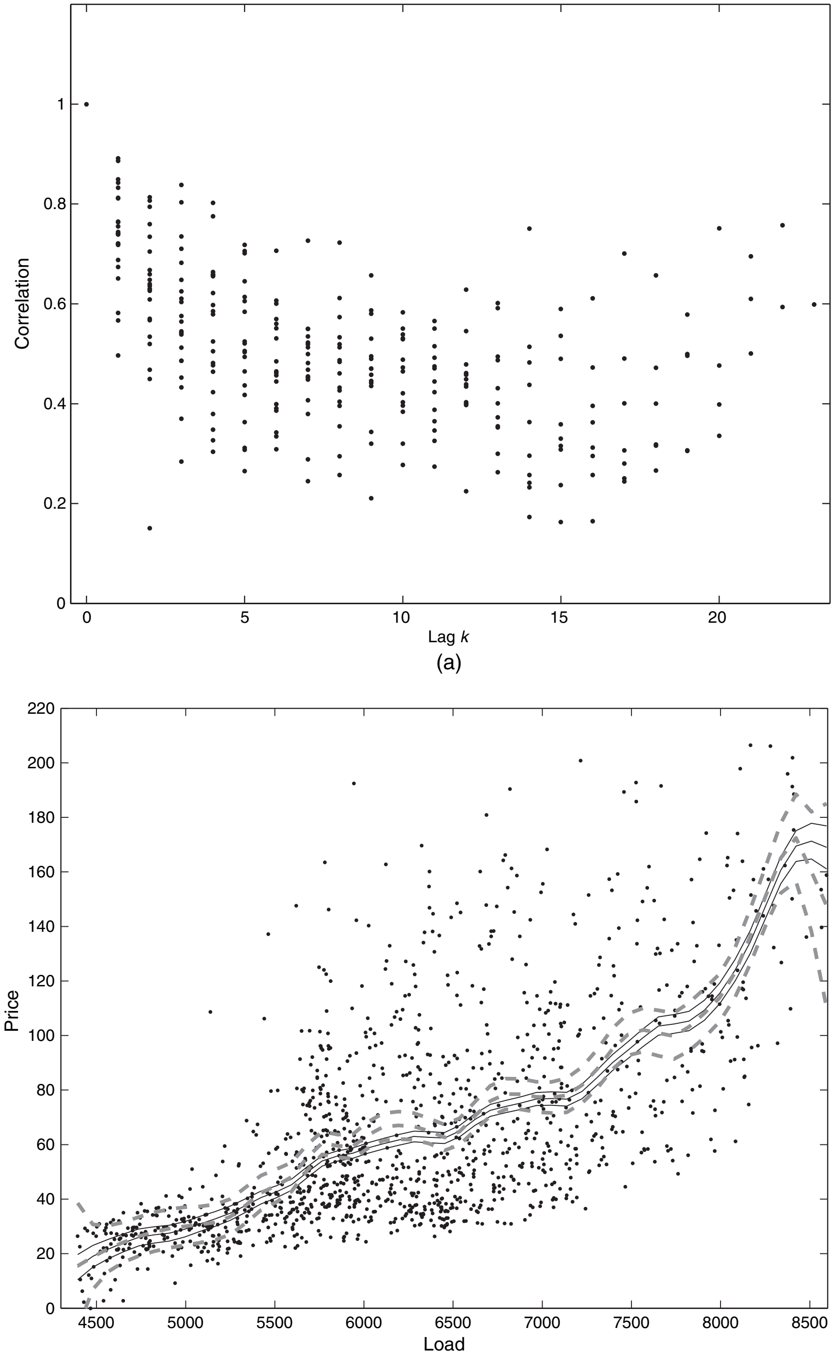
Method	*Parameter*	*Results for the following values of h:*				
		$h=0.02$	$h=0.03$	$h=0.04$	$h=0.05$	$h=0.06$
Wang's first	Bias	0.029	0.013	0.024	0.038	0.056
	SD	0.711	0.082	0.041	0.035	0.035
	RMISE	4.900	1.607	1.373	1.027	0.878
Wang's full	Bias	0.027	0.014	0.026	0.040	0.058
	SD	0.625	0.076	0.039	0.035	0.035
	RMISE	6.068	1.803	1.340	0.949	0.811
Chen and Jin (2005)	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 and Carroll (2006)	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 *et al.* (2008)	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

†The definitions of the parameters are the same as for Table 1.

district on each Wednesday of 2005. We studied the effect of electricity load on electricity price. As an illustration, we treated day as the 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 Fig. 1(b).

We first used local linear regression with working independence covariance matrix to estimate the regression. The plug-in bandwidth selector (Ruppert *et al.*, 1995) yields a bandwidth of 89. The broken curves in Fig. 1(b) are the resulting estimate along its 95% pointwise confidence interval. On the basis of the resulting estimate, we further obtain the residuals and estimate the correlation between $\varepsilon_{i,j}$ and $\varepsilon_{i,j+k}$ for $j=1,\ldots,23$ and $1 \leqslant k \leqslant 24-j$. The plot of estimated correlations is depicted in Fig. 1(a), which shows that the within-subject correlation is moderate. Thus, our proposed method may produce a more accurate estimate than local linear regression, ignoring the within-subject correlation.

Next, we apply the newly proposed procedure to this data set. The bandwidth that is selected by the plug-in bandwidth selector equals 91. The full curves in Fig. 1(b) are the fitted regression curves along with 95% pointwise confidence interval for the newly proposed procedure. Fig. 1(b) shows that the newly proposed procedure provides a smaller confidence interval than the interval by ignoring the within-subject correlation. In addition, the fitted curve by the method proposed is much smoother than the working independence local linear fit, because the new method can borrow information from more observations by taking the correlation into account. From Fig. 1(b), it can be seen that the relationship between electricity load and electricity price is non-linear. In general, the price increases as the load increases. However, the price change rate seems to remain almost constant when the load is 4500–7000, but the price change rate is much larger when the load is larger than 7500.



(a)

(b)

Fig. 1. (a) Plot of the estimated correlation between $\varepsilon_{i,j}$ and $\varepsilon_{i,j+k}$ *versus* the lag k (for example, dots at $k = 1$ correspond to the correlations between $\varepsilon_{i,j}$ and $\varepsilon_{i,j+1}$ for $j = 1, \ldots, 23$) and (b) scatter plot of observations and the plot of fitted regression curves (———, fitted regression by the method proposed and the corresponding 95% pointwise confidence interval; -------, local linear fit ignoring the within-subject correlation)

4. Concluding remarks

We have developed a new local estimation procedure for regression functions of longitudinal data. The procedure proposed uses the Cholesky decomposition and profile least squares techniques to estimate the correlation structure and regression function simultaneously. We demonstrate that the estimator proposed 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 non-parametric regression models. The methodology proposed 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.

Acknowledgements

The authors are grateful to the Associate Editor and referees for their helpful and constructive comments that substantially improved an earlier draft. 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 National Institute on Drug Abuse or the National Institutes of Health.

Appendix A: Proofs

Define $\mathbf{B}=\hat{\mathbf{F}}_a-\mathbf{F}_a$. Since $\mathbf{G}$ can be estimated by a parametric rate, we shall assume that $\mathbf{G}$ is known in our proof, without loss of generality. Our proofs use a strategy which is 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.

Condition 1. 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 js.

Condition 2. $m(\cdot)$ has the continuous second derivative in $x\in\Omega$.

Condition 3. The kernel $K(\cdot)$ is a bounded symmetric density function with bounded support and satisfies the Lipschitz condition.

Condition 4. $nh^8\rightarrow 0$ and $nh^2/\log(n)^2\rightarrow\infty$.

Condition 5. There is an $s>2$ such that $E\|F_{1j}\|^s<\infty, \forall j$, and for some $\xi>0$ such that $n^{1-2s^{-1}-2\xi}h\rightarrow\infty$.

Condition 6. $\sup_{x\in\Omega}|\hat{m}_{\mathrm{I}}(x)-m(x)|=o_p(n^{-1/4})$, where $\hat{m}_{\mathrm{I}}(x)$ is obtained by local linear regression pretending that the data are independent and identically distributed.

Lemma 1. Under conditions 1–6, we have the following results.

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

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

(b) $N^{-1/2}\hat{\mathbf{F}}_a^{\mathrm{T}}(I-\mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I-\mathbf{S}_h(\mathbf{X}))\mathbf{m}=o_p(1)$ and $N^{-1/2}\hat{\mathbf{F}}_a^{\mathrm{T}}(I-\mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I-\mathbf{S}_h(\mathbf{X}))\mathbf{B}\phi=o_p(1)$.
(c) Let $\mathbf{e}=(e_{11},\ldots,e_{nJ})^{\mathrm{T}}$. Then

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

The proof of lemma 1 is available from the authors on request.

A.1. Proof of theorem 1

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

$$A = \{\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\}^{-1}\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\mathbf{m}\surd N,$$
$$B = -\{\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\}^{-1}\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\mathbf{B}\phi\surd N,$$
$$C = \{\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\hat{\mathbf{F}}_a\}^{-1}\hat{\mathbf{F}}_a^{\mathrm{T}}(I - \mathbf{S}_h(\mathbf{X}))^{\mathrm{T}}\mathbf{G}^{-1}(I - \mathbf{S}_h(\mathbf{X}))\mathbf{e}\surd N.$$

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

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

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

Note that $E(\mathbf{e}|\mathbf{X}) = 0$. Therefore,

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

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

In addition, note that

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

Therefore,

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

Regarding the asymptotic normality,

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

Thus, conditioning on $\mathbf{X}$, the asymptotic normality can be established by using the central limit theorem since, given j, the e'_{ij}s are independent and identically distributed with mean 0 and variance d_j^2.

A.2. Proof of theorem 2

(a) The proof can be done in a similar way to the proof of theorem 1.

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

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

By noting that

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

and

$$\left\{ \sum_{j=1}^{J} f_j(x_0) E(d_j^2 \tilde{d}_j^{-4} | X_j = x_0) \right\} \sum_{j=1}^{J} f_j(x_0) E(d_j^{-2} | X_j = x_0) \geqslant \left\{ \sum_{j=1}^{J} f_j(x_0) E(\tilde{d}_j^{-2} | X_j = x_0) \right\}^2, \tag{14}$$

we can obtain the result. For result (13), the equality holds only when $\tilde{\phi} = \phi$. For the second inequality (14), on the basis of the Cauchy–Schwarz inequality, the equality holds only when $\tilde{d}_j/d_j$ are all equal. On the basis of the Cholesky decomposition result, $\tilde{\phi} = \phi$ 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. *Statist. Sin.*, **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*. London: Chapman and Hall.

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. *J. Am. Statist. Ass.*, **102**, 632–641.

Fan, J. and Li, R. (2004) New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Am. Statist. Ass.*, **99**, 710–723.

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

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. *J. Am. Statist. Ass.*, **95**, 520–534.

Lin, X. and Carroll, R. J. (2006) Semiparametric estimation in general repeated measures problems. *J. R. Statist. Soc.* B, **68**, 69–88.

Linton, O. B., Mammen, E., Lin, X. and Carroll, R. J. (2003) Accounting for correlation in marginal longitudinal nonparametric regression. *2nd Seattle Symp. Biostatistics*.

Qu, A., Lindsay, B. G. and Li, B. (2000) Improving generalised estimating equations 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 squares regression. *J. Am. Statist. Ass.*, **90**, 1257–1270.

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