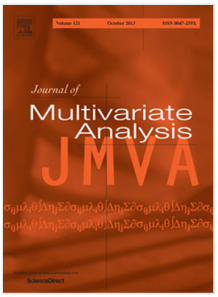
Accepted Manuscript



Estimation and model identification of longitudinal data time-varying nonparametric models

Shu Liu, Jinhong You, Heng Lian

PII: S0047-259X(17)30080-5
DOI: http://dx.doi.org/10.1016/j.jmva.2017.02.003
Reference: YJMVA 4220

To appear in: *Journal of Multivariate Analysis*

Received date: 13 April 2016

Please cite this article as: S. Liu, J. You, H. Lian, Estimation and model identification of longitudinal data time-varying nonparametric models, *Journal of Multivariate Analysis* (2017), http://dx.doi.org/10.1016/j.jmva.2017.02.003

Estimation and model identification of longitudinal data time-varying nonparametric models

Shu Liu*
School of Statistics and Information,
Shanghai University of International Business and Economics, Shanghai 201620, P.R. China
Jinhong You
School of Statistics and Management,
Shanghai University of Finance and Economics, Shanghai 200433, P.R. China
Heng Lian
Department of Mathematics,
City University of Hong Kong, Kowloon Tong, HK, 999077, Hong Kong

February 2, 2017

Abstract: In this paper, we consider nonparametric regression modeling for longitudinal data. An important modeling choice is that the covariate effect may change dynamically with time by using a bivariate link function. Comparing with Jiang and Wang [9, 10], and Zhang et al. [28] we make two distinct contributions to this important class of models. First, we show theoretically and empirically that taking the within-subject correlation into account can improve the estimation efficiency for the bivariate link function. Second, we propose a novel method involving a shrinkage estimation technique to identify consistently whether the effect of covariates is time-varying. Simulation studies are conducted to assess the finite-sample performance and a real data example is analyzed to illustrate the proposed methods.

AMS 2000 classifications: primary 62H12; secondary 62A10

Key words and phrases: Longitudinal data; Modified Cholesky decomposition; Model identification; Nonparametric regression; Time-varying.

*Corresponding author: Shu Liu, School of Statistics and Information, Shanghai University of International Business and Economics, Shanghai 201620, P.R. China. Email address: liu2008shu@126.com.

1 Introduction

In many biomedical and economic studies, measurements are collected over time on the same subject. The observations then form a longitudinal data set. For such data, standard parametric regression techniques that take into account within-subject correlation are well developed; see, e.g., [14, 18, 19, 20, 21, 27] and [1] for a summary of different types of parametric approaches. While parametric approaches are useful, issues may arise as to the adequacy of the model assumptions and the potential impact of model misspecifications on the analysis. This motivates the use of nonparametric approaches that we adopt in this paper.

Let $Y(t)$ and $Z(t)$ be the response variable and covariate, respectively, observed at time t. In traditional nonparametric regression, it is often assumed that $Y(t) = \mu(Z(t)) + \varepsilon(t)$, where $\varepsilon(t)$ represents the mean zero noise, which could be fitted by classic kernel, local polynomial or spline methods. However, in more complex data settings, the effect of the covariate may change with time, while in the standard univariate nonparametric regression the effect of time is entirely through the covariate $Z(t)$. Jiang and Wang [9] thus considered the time varying model

$$Y(t) = \mu(t, Z(t)) + \varepsilon(t), \tag{1}$$

which indicates that the effect of the covariate Z may change with time and as such is probably more realistic in many real situations. For related work, see [10, 28]. Jiang and Wang [9] extended the classical multivariate principal component analysis to accommodate covariate information for functional data and developed two estimators. Jiang and Wang [10] proposed a new single-index model to reflect the time-dynamic effects of the single index for longitudinal and functional response data with both longitudinal and time-invariant covariates. Zhang et al. [28] proposed a functional additive model with the components being time-dependent additive functions of the covariates. For the proposed functional additive model, they developed a backfitting algorithm to estimate the unknown regression functions.

It has been long recognized that the within-cluster correlation structure plays a very important role in a longitudinal data analysis and that it should be taken into account whenever possible. This usually improves estimation efficiency. Besides, the correlation structure may be of substantive interest by itself [11]. However, for model (1), previous work ignored such correlation structure, possibly due to the fact that explicit consideration of this aspect of the model is much more challenging and there are usually many more parameters in the covariance matrix and the positive definiteness of the covariance matrix has to be assured. One way to guarantee positive definiteness of the covariate matrix is to use a modified Cholesky decomposition that could be interpreted as an autoregression model,

with no constraint on the autoregression coefficients. Based on a modified Cholesky decomposition [11, 18, 19, 24], the within-subject covariance matrix is decomposed into a unit triangular matrix involving generalized autoregressive coefficients and a diagonal matrix involving innovation variances.

As a result, in order to construct a more efficient estimator of the mean function in model (1), we adopt a three-stage approach. First, an initial estimator of the mean function is obtained by ignoring the correlation. In the second stage, the estimated residuals are used to fit the autoregressive coefficients in the covariance. Finally, the estimated covariance is used to de-correlate the original observations and the final estimator is obtained. It is shown that the final estimator is more asymptotically efficient than estimators that ignore the within-cluster correlation. The large-sample properties of the proposed estimators are developed. We note that other ways of modeling the correlation structure are also possible, including [4, 12, 14, 29]. The advantage of the current approach is that the covariate/time effect on the correlation structure is explicitly produced, and nonparametric regression functions are avoided in the covariance modeling.

Modeling the time-varying effect requires estimating a link function with higher dimensionality, which adds unnecessary complexity if the covariate effect is not actually time-varying. Therefore, an interesting question is whether we can identify the situations when $\mu(t, Z(t))$ is actually only a function of its second argument. If so, a simpler model with mean function equal to $\mu(Z(t))$ can be adopted. Recently, Vogt [23] proposed a kernel-based L_2-test statistic to tackle a similar problem. However, his procedure is developed only for the case that the observations are made on a equally spaced grid densely over time. In this paper, we propose a novel method using the idea of shrinkage estimation [22]. We show that the shrinkage method can identify the true model structure (time-varying or non-time-varying) with probability approaching 1.

The rest of the paper is organized as follows. In Section 2 we describe estimators for the mean and the covariance structure, and establish their asymptotic properties. In particular, we show that the final estimator that takes into account the correlation structure is more asymptotically efficient than that which ignores the correlation structure. A shrinkage method for deciding whether a non-time-varying model is sufficient for the given data set is taken up in Section 3. In Section 4, we present Monte Carlo simulation results that empirically corroborate the theoretical results. We further illustrate the proposed procedure by analyzing a real data set in Section 5. Section 6 ends the article with a discussion. Technical proofs are presented in the Appendix.

2 Estimation methodology

2.1 Initial estimator of the mean function

Suppose there are n subjects, and for the ith subject there are m_i repeated measurements of $(Y(t), Z(t), t)$ over time. The jth observation of $(Y(t), Z(t), t)$ for the ith subject is denoted by (Y_{ij}, Z_{ij}, t_{ij}) with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$. Since subjects are often measured repeatedly over a given time period, the measurements of each subject are possibly correlated with each other but different subjects could be assumed to be independent.

We use local linear smoothing to obtain an initial estimator of the mean function, ignoring the within-subject correlation. In particular, the local linear fitting has several nice properties such as high statistical efficiency (in an asymptotic minimax sense), design adaptation and excellent boundary behavior. See Fan and Gijbels [3] for details. For any (t_{ij}, Z_{ij}) in a close neighborhood of (t, z), $\mu(t_{ij}, Z_{ij})$ can be approximated by

$$\mu(t_{ij}, Z_{ij}) \approx \mu(t, z) + \frac{\partial \mu(t, z)}{\partial t}(t_{ij} - t) + \frac{\partial \mu(t, z)}{\partial z}(Z_{ij} - z) \equiv \beta_0 + \beta_1(t_{ij} - t) + \beta_2(Z_{ij} - z).$$

As a result, the local linear smoother for the mean function $\mu(t, z)$ is $\widehat{\mu}(t, z) = \widehat{\beta}_0$ with $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)^\top$, and

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{Y_{ij} - \beta_0 - \beta_1(t_{ij} - t) - \beta_2(Z_{ij} - z)\}^2 K_{ht_{1,N}}(t_{ij} - t) K_{hz_{1,N}}(Z_{ij} - z), \qquad (2)$$

where K is a kernel function with $K_h(\cdot) = K(\cdot/h)/h$, $ht_{1,N}$ and $hz_{1,N}$ are bandwidths, and $N = m_1 + \cdots + m_n$. The solution to problem (2) is given by $\widehat{\boldsymbol{\beta}} = \mathbf{H}_{1,N}^{-1}(\mathbf{D}_{t,z}^\top \mathbf{W}_{t,z} \mathbf{D}_{t,z})^{-1} \mathbf{D}_{t,z}^\top \mathbf{W}_{t,z} \mathbf{Y}$, where $^\top$ denotes the transpose of a matrix or vector,

$$\mathbf{D}_{t,z} = \begin{pmatrix} 1 & \frac{t_{11} - t}{ht_{1,N}} & \frac{Z_{11} - z}{hz_{1,N}} \\ \vdots & \vdots & \vdots \\ 1 & \frac{t_{1,m_1} - t}{ht_{1,N}} & \frac{Z_{1,m_1} - z}{hz_{1,N}} \\ \vdots & \vdots & \vdots \\ 1 & \frac{t_{n,m_n} - t}{ht_{1,N}} & \frac{Z_{n,m_n} - z}{hz_{1,N}} \end{pmatrix},$$

$$\mathbf{W}_{t,z} = \mathrm{diag}\{K_{ht_{1,N}}(t_{11}-t)K_{hz_{1,N}}(Z_{11}-z), \ldots,$$
$$K_{ht_{1,N}}(t_{1,m_1}-t)K_{hz_{1,N}}(Z_{1,m_1}-z), \ldots, K_{ht_{1,N}}(t_{n,m_n}-t)K_{ht_{1,N}}(Z_{n,m_n}-z)\},$$

$\mathbf{Y} = (Y_{11}, \ldots, Y_{1,m_1}, \ldots, Y_{n,m_n})^\top$ and $\mathbf{H}_{1,N} = \mathrm{diag}(1, ht_{1,N}, hz_{1,N})$. Equivalently,

$$\widehat{\mu}(t,z) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} W_{h_1,ij}(t,z)Y_{ij},$$

where

$$W_{h_1,ij}(t,z) = \frac{K_{ht_{1,N}}(t_{ij}-t)K_{hz_{1,N}}(Z_{ij}-z)\{A_1 + (t_{ij}-t)A_2 + (Z_{ij}-z)A_3\}}{B},$$

with $A_1 = S_{n,20}S_{n,02} - S_{n,11}^2$, $A_2 = S_{n,01}S_{n,11} - S_{n,10}S_{n,02}$, $A_3 = S_{n,10}S_{n,11} - S_{n,01}S_{n,20}$, $B = A_1S_{n,00} + A_2S_{n,10} + A_3S_{n,01}$ and for all $\ell, k \in \{0,1,2\}$,

$$S_{n,lk}(t,u;\boldsymbol{\beta},h_1) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}(t_{ij}-t)^\ell(Z_{ij}-u)^k K_{ht_{1,N}}(t_{ij}-t)K_{hz_{1,N}}(Z_{ij}-u).$$

The following technical conditions are imposed to establish the asymptotic results. They may not be the weakest possible conditions.

Assumption 1: The density function $f(t,z)$ of (t_{ij}, Z_{ij}) is supported on $[0,1]\times\mathcal{Z}$, continuous and bounded away from zero and infinity.

Assumption 2: The kernel K is a density function with compact support and Lipschitz continuous.

Assumption 3: $\mu(t,z)$ is twice partially continuously differentiable on $[0,1]\times\mathcal{Z}$.

Assumption 4: The numbers of measurements m_i are uniformly bounded for all $1 \le i \le n$.

Assumption 5: The e_{ij} are iid random variables with mean zero and variance σ_e^2 .

Assumption 6: The bandwidth $ht_{s,N}$ satisfy $Nht_{s,N}^8/(\ln\ln N)^{1/2} \to 0$ and $Nht_{s,N}^2/(\ln N)^2 \to \infty$ as $n\to\infty$ and the bandwidth $hz_{s,N}$ satisfy $Nhz_{s,N}^8/(\ln\ln N)^{1/2} \to 0$ and $Nhz_{s,N}^2/(\ln N)^2 \to \infty$ as $n\to\infty$. Here, $s = 1, 2$ and 3. In addition, $ht_{1,N}/ht_{2,N} = o(1)$ and $hz_{1,N}/hz_{2,N} = o(1)$.

Remark 1. Assumption 1–3 are typical in the smoothing literature. Assumption 4 is widely used in the longitudinal data literature. In addition, the optimal asymptotic rate for two-dimensional smoothing trivially satisfies Assumption 6.

Denote $\mu_j = \int_{-\infty}^{\infty} u^j K(u)du$ and $\nu_j = \int_{-\infty}^{\infty} u^j K^2(u)du$. The asymptotic behavior of $\widehat{\boldsymbol{\beta}}$ is described in the following theorem, with necessary regularity conditions and proofs given in the Appendix.

Theorem 1. *If Assumption 1 to 6 hold, then we have, as* $n \to \infty$,

$$\sqrt{Nht_{1,N}hz_{1,N}}\Big[\mathbf{H}_{1,N}\left\{\begin{pmatrix}\widehat{\mu}(t,z)\\ \frac{\partial\widehat{\mu}(t,z)}{\partial t}\\ \frac{\partial\widehat{\mu}(t,z)}{\partial z}\end{pmatrix}-\begin{pmatrix}\mu(t,z)\\ \frac{\partial\mu(t,z)}{\partial t}\\ \frac{\partial\mu(t,z)}{\partial z}\end{pmatrix}\right\}-\frac{1}{2}\begin{pmatrix}\Big(ht_{1,N}^2\frac{\partial^2\mu(t,z)}{\partial t}+hz_{1,N}^2\frac{\partial^2\mu(t,z)}{\partial z}\Big)\mu_2\\ hz_{1,N}^2\frac{\partial^2\mu(t,z)}{\partial z^2}\\ ht_{1,N}^2\frac{\partial^2\mu(t,z)}{\partial t^2}\end{pmatrix}$$
$$+\,o_p(ht_{1,N}^2+hz_{1,N}^2)\Big]\rightsquigarrow\mathcal{N}(0,\boldsymbol{\Sigma}),$$

where $\rightsquigarrow$ *denotes convergence in distribution and* $\boldsymbol{\Sigma}=\{f(t,z)\}^{-1}\omega^2\mathfrak{S}$ *with* $f(t,z)$ *being the joint density function of* (t_{ij},Z_{ij}),

$$\omega^2=\lim_{n\to\infty}\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\varepsilon_{ij}^2\quad\textit{and}\quad\mathfrak{S}=\begin{pmatrix}\nu_0^2 & \frac{\nu_0\nu_1}{\mu_2} & \frac{\nu_0\nu_1}{\mu_2}\\ \frac{\nu_0\nu_1}{\mu_2} & \frac{\nu_0\nu_2}{\mu_2^2} & \frac{\nu_1^2}{\mu_2^2}\\ \frac{\nu_0\nu_1}{\mu_2} & \frac{\nu_1^2}{\mu_2^2} & \frac{\nu_0\nu_2}{\mu_2^2}\end{pmatrix}.$$

Due to the fact that the correlation between $\varepsilon_i(t_{ij_1})$ and $\varepsilon_i(t_{ij_2})$ is not considered in the estimators $\widehat{\mu}(t,z)$, $\partial\widehat{\mu}(t,z)/\partial t$ and $\partial\widehat{\mu}(t,z)/\partial z$, one cannot expect them to be asymptotically efficient. In the following sections, based on the estimator $\widehat{\mu}(t,z)$ we propose an improved estimator for $\mu(t,z)$, $\partial\mu(t,z)/\partial t$ and $\partial\mu(t,z)/\partial z$.

2.2 Fitting subject correlation structure

In order to construct improved estimators of $\mu(t,z)$, $\partial\mu(t,z)/\partial t$ and $\partial\mu(t,z)/\partial z$ we need to fit the within subject correlation structure first. Denote $\mathrm{cov}(\boldsymbol{\varepsilon}_i|\mathbf{Z}_i)=\boldsymbol{\Sigma}_i$ where $\boldsymbol{\varepsilon}_i=(\varepsilon_{i1},\ldots,\varepsilon_{im_i})^\top$ with $\varepsilon_{ij}=\varepsilon_i(t_{ij})$, $\mathbf{Z}_i=(Z_{i1},\ldots,Z_{im_i})^\top$, and $\boldsymbol{\Sigma}_i$ is an $m_i\times m_i$ matrix and may depend on $\mathbf{Z}_i$. Then by the Cholesky decomposition technique, proceeding as in [18], there exists a lower triangular matrix $\boldsymbol{\Phi}_i$ with diagonal elements being ones such that

$$\mathrm{cov}(\boldsymbol{\Phi}_i\boldsymbol{\varepsilon}_i|\mathbf{Z}_i)=\boldsymbol{\Phi}_i\boldsymbol{\Sigma}_i\boldsymbol{\Phi}_i^\top=\mathrm{diag}(\sigma_{e_{i1}}^2,\ldots,\sigma_{e_{im_i}}^2).$$

Namely, for all $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, m_i\}$,

$$\varepsilon_{i1} = e_{i1}, \quad \varepsilon_{ij} = \phi_{j,1}^{(i)} \varepsilon_{i1} + \cdots + \phi_{j,j-1}^{(i)} \varepsilon_{i,j-1} + e_{ij}, \tag{3}$$

where $(e_{i1}, \ldots, e_{im_i})^\top = \boldsymbol{\Phi}_i \boldsymbol{\varepsilon}_i$, $\phi_{j,\ell}^{(i)}$ is the negative of the (j, ℓ) element of $\boldsymbol{\Phi}_i$. Obviously, the e_{ij} are uncorrelated with $\mathrm{E}(e_{ij}) = 0$ and $\mathrm{var}(e_{ij}) = \sigma_{e_{ij}}^2$ for all $j \in \{1, \ldots, m_i\}$. For simplicity, we assume that $\sigma_{e_{ij}}^2 = \sigma_e^2$ for all $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, m_i\}$. Following [11, 15, 24], we assume that

$$\phi_{j,\ell}^{(i)} = \mathbf{W}_{j,\ell}^{(i)\top} \boldsymbol{\theta}, \tag{4}$$

where $\mathbf{W}_{j,\ell}^{(i)} = (W_{j,\ell,1}^{(i)}, \ldots, W_{j,\ell,q}^{(i)})^\top$ is the $q \times 1$ vector of covariates which may contain time, time difference, other baseline covariates, as well as their interactions, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ is the regression coefficients. Based on the estimated residuals $\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{\mu}(t_{ij}, Z_{ij})$, and applying the least square technique we can obtain an estimator of $\boldsymbol{\theta}$, namely

$$\widehat{\boldsymbol{\theta}} = \Big(\sum_{i=1}^{n} \sum_{j=2}^{m_i} \widehat{\boldsymbol{\Pi}}_{ij} \widehat{\boldsymbol{\Pi}}_{ij}^\top \Big)^{-1} \sum_{i=1}^{n} \sum_{j=2}^{m_i} \widehat{\boldsymbol{\Pi}}_{ij} \widehat{\varepsilon}_{ij},$$

where $\widehat{\boldsymbol{\Pi}}_{ij} = (\sum_{k=1}^{j-1} \widehat{\varepsilon}_{i,k} W_{j,k,1}^{(i)}, \ldots, \sum_{k=1}^{j-1} \widehat{\varepsilon}_{i,k} W_{j,k,q}^{(i)})^\top$. The following theorem shows that $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal.

Theorem 2. *If Assumption 1 to 6 hold, then we have, as* $n \to \infty$, $\sqrt{N-n}\,(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{\Lambda}^{-1})$, *where*

$$\boldsymbol{\Lambda} = \lim_{n \to \infty} \frac{1}{N-n} \sum_{i=1}^{n} \sum_{j=2}^{m_i} E(\boldsymbol{\Pi}_{ij} \boldsymbol{\Pi}_{ij}^\top)$$

with $\boldsymbol{\Pi}_{ij} = (\sum_{k=1}^{j-1} \varepsilon_{i,k} W_{j,k,1}^{(i)}, \ldots, \sum_{k=1}^{j-1} \varepsilon_{i,k} W_{j,k,q}^{(i)})^\top$.

2.3 Improved estimator of the mean function and its partial derivatives

The estimators $\widehat{\mu}(t, z)$, $\partial \widehat{\mu}(t, z)/\partial t$ and $\partial \widehat{\mu}(t, z)/\partial z$ proposed in Section 2.1 do not take into account the correlations within the subjects. We now construct new estimators by accounting for these correlations. If the ε_{ij} were available, then model (1) would become the following partially linear model with

bivariate nonparametric component and uncorrelated error terms:

$$Y_{i1} = \mu(t_{i1}, Z_{i1}) + e_{i1},$$
$$Y_{ij} = \mu(t_{ij}, Z_{ij}) + \phi_{j,1}^{(i)}\varepsilon_{i1} + \cdots + \phi_{j,j-1}^{(i)}\varepsilon_{i,j-1} + e_{ij},$$

where $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, m_i\}$. Combining the fact $\phi_{j,\ell}^{(i)} = \mathbf{W}_{j,\ell}^{(i)\top}\boldsymbol{\theta}$, it is easy to see that

$$Y_{ij} - \sum_{r=1}^{q}\Big(\sum_{k=1}^{j-1}\varepsilon_{ik}W_{j,k,r}^{(i)}\Big)\theta_r = \mu(t_{ij}, Z_{ij}) + e_{ij}$$

for all $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, m_i\}$. In addition, it is easy to see that (3) is an autoregressive model. Therefore, $\mathrm{var}(e_{i1}) = \mathrm{var}(\varepsilon_{i1})$ and $\mathrm{var}(e_{ij}) \le \mathrm{var}(\varepsilon_{ij})$ for all $j \in \{2, \ldots, m_i\}$. As a result, if we take $Y_{ij} - \sum_{r=1}^{q}(\sum_{k=1}^{j-1}\varepsilon_{ik}W_{j,k,r}^{(i)})\theta_r$ as pseudo-responses, these can be used to construct more efficient estimators of $\mu(t,z)$, $\partial\mu(t,z)/\partial t$ and $\partial\mu(t,z)/\partial z$ than the initial estimators $\widehat{\mu}(t,z)$, $\partial\widehat{\mu}(t,z)/\partial t$ and $\partial\widehat{\mu}(t,z)/\partial z$.

In practice, ε_{ik} and θ_r are unknown. Therefore, we replace them by $\widehat{\varepsilon}_{ik}$ and $\widehat{\theta}_r$. For each $i \in \{1, \ldots, n\}$ and $j \in \{2, \ldots, m_i\}$, define

$$Y_{ij}^{*} = Y_{ij} - \sum_{r=1}^{q}\Big(\sum_{k=1}^{j-1}\widehat{\varepsilon}_{ik}W_{j,k,r}^{(i)}\Big)\widehat{\theta}_r. \tag{5}$$

Then we propose an improved estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (\mu(t,z), \partial\mu(t,z)/\partial t, \partial\mu(t,z)/\partial z)^\top$ as the solution of the following problem:

$$\arg\min_{\boldsymbol{\beta}^I}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij}^{*} - \beta_0 - \beta_1(t_{ij}-t) - \beta_2(Z_{ij}-z)\}^2 K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z).$$

The minimizer is

$$\widehat{\boldsymbol{\beta}}^I = (\widehat{\beta}_0^I, \widehat{\beta}_1^I, \widehat{\beta}_2^I)^\top = (\widehat{\mu}^I(t,z), \partial\widehat{\mu}^I(t,z)/\partial t, \partial\widehat{\mu}^I(t,z)/\partial z)^\top = \mathbf{H}_{2,N}^{-1}(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{Y}^{*},$$

where $\mathbf{Y}^{*} = (Y_{11}, \ldots, Y_{1,m_1}^{*}, \ldots, Y_{n,1}, \ldots, Y_{n,m_n}^{*})^\top$ and $\mathbf{H}_{2,N}$, $\mathbf{D}_{t,z}^{*}$ and $\mathbf{W}_{t,z}^{*}$ have the same definitions as $\mathbf{H}_{1,N}$, $\mathbf{D}_{t,z}$ and $\mathbf{W}_{t,z}$ but with $ht_{2,N}$ and $hz_{2,N}$ in place of $ht_{1,N}$ and $hz_{1,N}$, respectively.

For $\widehat{\boldsymbol{\beta}}^I$, we have the following the asymptotic result.

Theorem 3. *If Assumption 1 to and 6 hold, then we have, as* $n \to \infty$,

$$\sqrt{Nht_{2,N}hz_{2,N}}\Big[\mathbf{H}_{2,N}\left\{\begin{pmatrix}\widehat{\mu}^I(t,z)\\ \frac{\partial\widehat{\mu}^I(t,z)}{\partial t}\\ \frac{\partial\widehat{\mu}^I(t,z)}{\partial z}\end{pmatrix}-\begin{pmatrix}\mu(t,z)\\ \frac{\partial\mu(t,z)}{\partial t}\\ \frac{\partial\mu(t,z)}{\partial z}\end{pmatrix}\right\}-\frac{1}{2}\begin{pmatrix}\left(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t}+hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z}\right)\mu_2\\ hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z^2}\\ ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t^2}\end{pmatrix}$$
$$+o_p(ht_{2,N}^2+hz_{2,N}^2)\Big]\rightsquigarrow\mathcal{N}(0,\mathbf{\Sigma}^*),$$

where $\mathbf{\Sigma}^*=\{f(t,z)\}^{-1}\sigma_e^2\mathfrak{F}$ *and other symbols are same as those defined in Theorem 1.*

Remark 2. Due to the fact that $\mathrm{var}(e_{i1})=\mathrm{var}(\varepsilon_{i1})$ and $\mathrm{var}(e_{ij})\le\mathrm{var}(\varepsilon_{ij})$ for all $j\in\{2,\ldots,m_i\}$, one has $\sigma_e^2\le\omega^2$ and $\mathbf{\Sigma}^*\le\mathbf{\Sigma}$. As a result, the improved local linear estimators $\widehat{\mu}^I(t,z)$, $\partial\widehat{\mu}^I(t,z)/\partial t$ and $\partial\widehat{\mu}^I(t,z)/\partial z$ are asymptotically more efficient than the local linear estimators $\widehat{\mu}(t,z)$, $\partial\widehat{\mu}(t,z)/\partial t$ and $\partial\widehat{\mu}(t,z)/\partial z$, respectively and the latter neglect the informative correlation structure.

3 Model identification

Obviously, modeling the time-varying effect increases the flexibility of modeling and tends to reduce the modeling bias greatly. However, it adds unnecessary complexity when the covariate effect is actually not time-varying. Therefore, an interesting question is whether we could identify the situation when $\mu(t,z)$ is actually only a function of z. If so, a simpler model with mean function equal to $\mu(z)$ could be adopted. Recently, Vogt [23] proposed a kernel-based L_2-test statistic to tackle a similar problem. It should be noted that his procedure is developed only for the scenario that the observations are made on a equally spaced grid densely over time. In addition, his method is based on the bootstrap technique and has a heavy calculation burden. We propose here a novel approach using the idea of shrinkage estimation [22] and show that the shrinkage method can identify the true model structure (time-varying or non-time-varying) with probability approaching 1.

Without loss of generality, we assume that $t_{11}\le\cdots\le t_{1m_1}\le\cdots\le t_{nm_n}$. For any $i\in\{1,\ldots,n\}$ and $j\in\{1,\ldots,m_i\}$, let

$$\mathbf{b}_{ij}=\mathbf{D}\mathbf{v}_{ij}=-(\mu(t_{12},Z_{ij})-\mu(t_{11},Z_{ij}),\ldots,\mu(t_{nm_n},Z_{ij})-\mu(t_{n,m_n-1},Z_{ij}))^\top$$

and $||\mathbf{b}_{ij}||=\sqrt{\mathbf{v}_{ij}^\top\mathbf{D}^\top\mathbf{D}\mathbf{v}_{ij}}$, where $\mathbf{D}_{(N-1)\times N}=(\mathbf{I}_{N-1},\mathbf{0}_{(N-1)\times1})-(\mathbf{0}_{(N-1)\times1},\mathbf{I}_{N-1})$ and $\mathbf{v}_{ij}=(\mu(t_{11},Z_{ij}),\ldots,\mu(t_{nm_n},Z_{ij}))^\top$ is an $N\times1$ vector. It is easy to see that if the marginal function

does not change over time, then

$$||\overline{\mathbf{b}}|| = \frac{1}{N}\sqrt{\mathbf{v}^\top \mathbf{D}^{*\top}\mathbf{D}^*\mathbf{v}} = \frac{1}{N}\left(\sum_{i=1}^{n}\sum_{j=1}^{m_i}\|\mathbf{b}_{ij}\|^2\right)^{1/2} = 0,$$

otherwise $||\overline{\mathbf{b}}|| > 0$, where $\mathbf{D}^*$ is a N^2 diagonal block matrix with $\mathbf{D}$ in each diagonal element and $\mathbf{v} = (\mathbf{v}_{11}^\top, \ldots, \mathbf{v}_{nm_n}^\top)^\top$ is a $N^2 \times 1$ vector. Based on the above fact, following the grouped LASSO idea of Yuan and Lin [25] we propose the penalized estimate

$$\widehat{\mathbf{B}}_\lambda = (\widehat{\boldsymbol{\mu}}_\lambda(t_{11}, Z), \ldots, \widehat{\boldsymbol{\mu}}_\lambda(t_{1m_1}, Z), \ldots, \widehat{\boldsymbol{\mu}}_\lambda(t_{nm_n}, Z))^\top = \underset{\mathbf{B}\in\mathbb{R}^{N\times N}}{\arg\min}\, Q_\lambda(\mathbf{B}),$$

where λ is the tuning parameter, $\widehat{\boldsymbol{\mu}}_\lambda(t_{ij}, Z) = (\widehat{\mu}_\lambda(t_{ij}, Z_{11}), \ldots, \widehat{\mu}_\lambda(t_{ij}, Z_{1m_1}), \ldots, \widehat{\mu}_\lambda(t_{ij}, Z_{nm_n}))^\top$ and

$$Q_\lambda(\mathbf{B}) = \sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij} - \mu(t_{i_1j_1}, Z_{i_2j_2})\}^2 K_{ht_{3,N}}(t_{ij} - t_{i_1j_1})K_{hz_{3,N}}(Z_{ij} - Z_{i_2j_2}) + \lambda\|\overline{\mathbf{b}}\|. \tag{6}$$

In a typical least squares regression, computational algorithms for the LASSO-type problems have been very well developed. These algorithms include the shooting algorithm [6], local quadratic approximation [5], the least angle regression [2], and many others. We describe here an easy implementation based on the idea of the local quadratic approximation [5]. Specifically, our implementation is based on an iteration algorithm with the unpenalized Nadaraya–Watson estimator as the initial estimator. Next, we define

$$\widehat{\mathbf{B}}_\lambda^{(m)} = (\widehat{\boldsymbol{\mu}}_\lambda^{(m)}(t_{11}, Z), \ldots, \widehat{\boldsymbol{\mu}}_\lambda^{(m)}(t_{1m_1}, Z), \ldots, \widehat{\boldsymbol{\mu}}_\lambda^{(m)}(t_{nm_n}, Z))^\top,$$

to be the estimate obtained in the mth iteration. Let

$$\widehat{\overline{\mathbf{b}}}_\lambda^m = \frac{1}{N}(\widehat{\mu}_\lambda^m(t_{12}, Z_{11}) - \widehat{\mu}_\lambda^m(t_{11}, Z_{11}), \ldots, \widehat{\mu}_\lambda^m(t_{nm_n}, Z_{nm_n}) - \widehat{\mu}_\lambda^m(t_{n,m_n-1}, Z_{nm_n}))^\top.$$

Then the loss function in (6) can be locally approximated by

$$\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij} - \mu(t_{i_1j_1}, Z_{i_2j_2})\}^2 K_{ht_{3,N}}(t_{ij} - t_{i_1j_1})K_{hz_{3,N}}(Z_{ij} - Z_{i_2j_2}) + \lambda\frac{\|\overline{\mathbf{b}}\|^2}{\|\widehat{\overline{\mathbf{b}}}_\lambda^m\|}. \tag{7}$$

We can update the estimator by the solution of $\widehat{\mu}_\lambda^{(m)}(t, z)$ that minimizes (7), denote by $\widehat{\mathbf{B}}_\lambda^{(m+1)}$. It is

easy to see that the minimizer has a closed form, viz.

$$\mathrm{vec}(\widehat{\mathbf{B}}_\lambda^{(m+1)}) = (\mathcal{M} + \mathcal{D}^{(m)})^{-1}\mathcal{N},$$

where $\mathrm{vec}(\mathbf{A})$ denotes the vectorization of matrix $\mathbf{A}$,

$$\mathcal{M} = \mathrm{diag}\Big\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{11})K_{hz_{3,N}}(Z_{ij}-Z_{11}), \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{12})K_{hz_{3,N}}(Z_{ij}-Z_{11}), \ldots,$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{1m_1})K_{hz_{3,N}}(Z_{ij}-Z_{11}), \ldots, \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{nm_n})K_{hz_{3,N}}(Z_{ij}-Z_{nm_n})\Big\},$$

and

$$\mathcal{N} = \Big(\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{11})K_{hz_{3,N}}(Z_{ij}-Z_{11})Y_{ij}, \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{12})K_{hz_{3,N}}(Z_{ij}-Z_{11})Y_{ij}, \ldots,$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{1m_1})K_{hz_{3,N}}(Z_{ij}-Z_{11})Y_{ij}, \ldots, \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ht_{3,N}}(t_{ij}-t_{nm_n})K_{hz_{3,N}}(Z_{ij}-Z_{nm_n})Y_{ij}\Big)^\top,$$

and $\mathcal{D}^{(m)}$ is a block diagonal matrix with each diagonal block is given by $\lambda/\|\widehat{\widehat{\mathbf{b}}}_\lambda^{m}\|\mathbf{D}^*\mathbf{D}^{*\top}$. So as $m \to \infty$, denote the limiting values of $\widehat{\mathbf{B}}_\lambda^{(m+1)}$ and $\widehat{\mu}_\lambda^{(m)}(t,z)$ respectively by $\widehat{\mathbf{B}}_\lambda$ and $\widehat{\mu}_\lambda(t,z)$, which are our final estimators.

When N is very large, the above calculation might be time consuming as $\mathcal{M}$ is an $N^2 \times N^2$ matrix. One way to simplify the calculation is to use sparser grids as suggested in [8]. In addition, the tuning parameter λ should be selected. Denote by df_λ the efficient number of degrees of freedom. If the selected model is bivariate, then $df_\lambda = 1$, otherwise 0. Also let

$$\mathrm{RSS}_\lambda = \frac{1}{N^3}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij} - \widehat{\mu}_\lambda(t_{i_1j_1}, Z_{i_2j_2})\}^2 K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2}).$$

Then λ is selected according to the following BIC-type criterion

$$\mathrm{BIC}_\lambda = \ln(\mathrm{RSS}_\lambda) + df_\lambda \times \frac{\ln(Nht_{3,N}hz_{3,N})}{Nht_{3,N}hz_{3,N}}. \tag{8}$$

It should be noted that the efficient sample size $Nht_{3,N}hz_{3,N}$ is used instead of the original sample size N. Then the tuning parameter can be obtained as $\widehat{\lambda} = \arg\min_\lambda \mathrm{BIC}_\lambda$.

Based on the regular conditions listed in Section 2, the sparsity, oracle efficiency and selection consistency for the estimator can be established as following.

Theorem 4. *(Estimation Sparsity) Suppose that Assumption 1 to 6 hold,* $Nht_{3,N}^{-1/2}\ hz_{3,N}^{-1/2}\lambda \to \infty$ *as* $n \to \infty$*. In addition, suppose that the true model is* $\mu(t,z) \equiv \mu(z)$*. Then there exists a univariate function of* z*,* $\widehat{\mu}(z)$*, such that*

$$\sup_{t\in\mathcal{T},z\in\mathcal{Z}} \Pr\{\widehat{\mu}_\lambda(t,z) = \widehat{\mu}(z)\} = 1.$$

By Theorem 4 we know that the true model can be ideally specified, then (1) becomes a standard nonparametric regression model for longitudinal data, which has been investigated by many author. Since this specification is not always available in practice, we call the estimator under the ideal specification the oracle estimator. Specifically, for any given z

$$\widehat{\mu}_{ora}(z) = \left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{hz_{3,N}}(Z_{ij}-z)\right\}^{-1}\left\{\sum_{i=1}^{n}\sum_{j=1}^{m_i} Y_{ij}K_{hz_{3,N}}(Z_{ij}-z)\right\}.$$

Then the following theorem establishes the oracle property.

Theorem 5. *(Estimation Oracle) Suppose that Assumption 1 to 6 hold,* $Nht_{3,N}^{-1/2}\ hz_{3,N}^{-1/2}\lambda \to \infty$ *as* $n \to \infty$*. In addition, suppose that the true model is* $\mu(t,z) \equiv \mu(z)$*. Then we have that*

$$\sup_{t\in\mathcal{T},z\in\mathcal{Z}} \|\widehat{\mu}_\lambda(t,z) - \widehat{\mu}_{ora}(z)\| = o_p(hz_{3,N}^2 + 1/\sqrt{Nhz_{3,N}}).$$

Define $\mathcal{S}$ as an arbitrary model. Then $\mathcal{S}_T$ denotes the true model, and $\mathcal{S}_\lambda$ represents the model identified by the proposed estimate $\widehat{\mu}_\lambda(t,z)$. Consequently, $\mathcal{S}_{\widehat{\lambda}}$ represents the model identified by $\widehat{\mu}_{\widehat{\lambda}}(t,z)$. The following theorem indicates that the tuning parameter selected by the BIC criterion in (8) can indeed identify the true model consistently.

Theorem 6. *(Selection consistency) Suppose that Assumption 1 to 6 hold,* $Nht_{3,N}^{-1/2}\ hz_{3,N}^{-1/2}\lambda \to \infty$ *as* $n \to \infty$*. Then the tuning parameter* λ *selected by the BIC criterion can indeed identify the true model consistently, i.e.,* $\Pr(\mathcal{S}_{\widehat{\lambda}} = \mathcal{S}_T) \to 1$ *as* $n \to \infty$*.*

4 Simulation studies

In this section, we conduct a set of Monte Carlo simulations to demonstrate the finite-sample performance of the proposed methods in the above sections.

Example 4.1 In this example, we focus on evaluating the improvement of the estimation of the mean function by taking the within-subject correlation among repeated measurements over time into account. The data are generated by the longitudinal data nonparametric regression model defined, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$, by

$$Y_i(t_{ij}) = \mu(t_{ij}, Z_{ij}) = 2\cos(\pi t_{ij})(1 + Z_{ij}^2) + \varepsilon_{ij},$$

where $t_{ij} \sim \mathcal{U}(0, 2)$, $Z_{ij} = 0.05 \times t_{ij} + 0.9 \times \eta_{ij}$ with $\eta_{ij} \sim \mathcal{U}(0, 1)$. Furthermore, ε_{ij} satisfy (3) and (4) with $\mathbf{W}_{j,\ell}^{(i)} = (1, t_{il} - t_{ij})^\top$ and $e_{ij} \sim \mathcal{N}(0, 0.2)$.

The sample size takes $n = 50, 100, 200$ with $m_i = m = 4, 6, 8$ for each individual, i.e., the balanced longitudinal data structure, and 500 simulation replications are run to draw summary statistics. In order to describe different correlation structure, we use four different $\boldsymbol{\theta}$, namely $\boldsymbol{\theta} = (0, 0)^\top$, $\boldsymbol{\theta} = (0.2, 0.3)^\top$, $\boldsymbol{\theta} = (0.3, 0.2)^\top$, $\boldsymbol{\theta} = (0.2, 0.5)^\top$ and $\boldsymbol{\theta} = (0.5, 0.2)^\top$ in our simulation.

It is well known that bandwidth selection has a much larger effect than the choice of kernel function. As a result, the Gaussian kernel is used in our simulation study and 10-fold cross-validation is used to select bandwidth. For the two-stage estimator, 80% of the optimal bandwidth is used as the bandwidth in the first stage and the optimal bandwidth is used in the second stage. Based on our simulation study experience, the two-stage estimator is not too sensitive to the choice of the bandwidth of the first-stage estimator.

We compare the finite empirical performance of the initial estimator $\widehat{\mu}(t, z)$, the improved estimator $\widehat{\mu}^I(t, z)$ and the benchmark estimator $\widehat{\mu}^B(t, z)$. The benchmark estimator is the same as the proposed two-stage estimator except that the within-subject correlation among repeated measurements over time is assumed to be known exactly. The finite-sample performance of these three estimators is assessed via the root of average squared errors (RASE):

$$\text{RASE}\{\check{\mu}(t, z)\} = \left[\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\check{\mu}(t_{ij}, Z_{ij}) - \mu(t_{ij}, Z_{ij})\}^2\right]^{1/2}.$$

The mean and standard deviation (Std) of the RASE over 500 simulated samples are presented in Table 1. From this table we can see that

(i) The finite-sample performance of $\widehat{\mu}(t, z)$, $\widehat{\mu}^I(t, z)$ and $\widehat{\mu}^B(t, z)$ improves with sample size n and the number of repeated measurements.

(ii) When there is no within-subject correlation among repeated measurements over time, namely $\boldsymbol{\theta} = 0$, the finite-sample performance of the improved estimator $\widehat{\mu}^I(t, z)$ is almost the same as

that of the initial estimator $\widehat{\mu}(t,z)$. When there exists within-subject correlation among repeated measurements over time, the improved estimator $\widehat{\mu}^I(t,z)$ outperforms the initial estimator $\widehat{\mu}(t,z)$. The improvement is much greater when the within-subject correlation is strong.

(iii) In most of the scenarios, the finite-sample performance of the improved estimator $\widehat{\mu}^I(t,z)$ is almost the same as the benchmark estimator $\widehat{\mu}^B(t,z)$.

Table 1 also reports the finite-sample performance of the estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)^\top$ of the autoregressive coefficients $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$. From Table 1 we can see that $\widehat{\boldsymbol{\theta}}$ is almost unbiased and its standard deviation decreases with the increasing of the sample size n or the number of repeated measurements.

It is easy to see that the results mentioned above are consistent with the theory developed in Section 2.

Example 4.2 In this example, we check the robustness of linear model assumption for the correlation structure. The setting is the same as in Example 4.1 except that $\boldsymbol{\theta} = (0.3, 0.2, 0.5)^\top$ and $\mathbf{W}_{j,\ell}^{(i)} = (1, t_{i\ell} - t_{ij}, (t_{i\ell} - t_{ij})^2)^\top$. We compare the finite-sample performance of the initial estimator $\widehat{\mu}(t,z)$ and the improved estimator $\widehat{\mu}^I(t,z)$ which takes the correlation structure with $\mathbf{W}_{j,\ell}^{(i)} = (1, t_{i\ell} - t_{ij})^\top$. The results are summarized in Table 2. From this table, we can see that the results mentioned above show that our method is robust for the misspecification of the correlation structure.

In Section 3, we developed a method to check whether the mean function is time-independent. The following example is used to evaluate the finite sample performance of the proposed model identification method.

Example 4.3 The data are generated by the longitudinal data nonparametric regression model defined, for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m_i\}$ by

$$Y_i(t_{ij}) = \mu(t_{ij}, Z_{ij}) = \{1 - c\cos(\pi t_{ij})\}Z_{ij}^2 + \varepsilon_{ij},$$

where the definitions of t_{ij}, Z_{ij} are same as those in the Example 4.1 and c is a constant which is used to evaluate the degree of time-independent of the mean function. Obviously, $c = 0$ implies that the mean function is time-independent. When the value of c increases, the degree of time-dependent of the mean function becomes strong. The results are summarized in Table 3, which shows that the proposed method works very well.

Table 1: Finite-sample performance of the estimators of the unknown nonparametric function and autoregressive coefficients in the error process.

	m	4			6			8		
(θ_1,θ_2)	n	50	100	200	50	100	200	50	100	200
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.0865	0.0508	0.0413	0.0602	0.0443	0.0363	0.0521	0.0415	0.0308
	Std(RASE)	0.0163	0.0083	0.0057	0.0101	0.0066	0.0053	0.0080	0.0056	0.0041
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0867	0.0508	0.0417	0.0602	0.0443	0.0363	0.0522	0.0416	0.0308
	Std(RASE)	0.0164	0.0083	0.0057	0.0101	0.0066	0.0053	0.0079	0.0056	0.0041
$(0.0,0.0)$ $\widehat{\mu}^B(\cdot,\cdot)$	Mean(RASE)	0.0865	0.0508	0.0413	0.0602	0.0443	0.0363	0.0521	0.0415	0.0308
	Std(RASE)	0.0163	0.0083	0.0057	0.0101	0.0066	0.0053	0.0080	0.0056	0.0041
$\widehat{\theta}_1$	mean	0.0009	-0.0112	-0.0065	-0.0107	-0.0112	-0.0029	-0.0120	-0.0049	-0.0046
	std	0.1061	0.0797	0.0543	0.0647	0.0468	0.0315	0.0478	0.0326	0.0247
$\widehat{\theta}_2$	mean	0.0018	0.0152	0.0086	0.0076	0.0165	0.0019	0.0177	0.0076	0.0074
	std	0.1637	0.1200	0.0852	0.1182	0.0890	0.0576	0.0967	0.0630	0.0478
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.0912	0.0568	0.0457	0.0717	0.0544	0.0406	0.0717	0.0544	0.0423
	Std(RASE)	0.0208	0.0103	0.0073	0.0165	0.0100	0.0071	0.0153	0.0100	0.0068
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0901	0.0531	0.0432	0.0652	0.0477	0.0378	0.0610	0.0477	0.0347
	Std(RASE)	0.0208	0.0097	0.0062	0.0159	0.0094	0.0076	0.0150	0.0094	0.0075
$(0.2,0.3)$ $\widehat{\mu}^B(\cdot,\cdot)$	Mean(RASE)	0.0900	0.0530	0.0432	0.0650	0.0476	0.0378	0.0609	0.0476	0.0347
	Std(RASE)	0.0208	0.0098	0.0062	0.0159	0.0094	0.0075	0.0151	0.0094	0.0075
$\widehat{\theta}_1$	mean	0.1942	0.1893	0.1901	0.1956	0.1977	0.1984	0.1944	0.1977	0.1976
	std	0.1052	0.0812	0.0552	0.0619	0.0437	0.0290	0.0426	0.0437	0.0202
$\widehat{\theta}_2$	mean	0.2288	0.2654	0.2871	0.2407	0.2773	0.2841	0.2743	0.2773	0.2927
	std	0.1729	0.1321	0.0941	0.1282	0.0932	0.0598	0.1056	0.0932	0.0478
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.0930	0.0588	0.0473	0.0784	0.0586	0.0437	0.0862	0.0640	0.0530
	Std(RASE)	0.0219	0.0109	0.0080	0.0192	0.0109	0.0077	0.0204	0.0150	0.0091
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0917	0.0539	0.0437	0.0685	0.0496	0.0390	0.0684	0.0518	0.0381
	Std(RASE)	0.0222	0.0103	0.0064	0.0190	0.0111	0.0089	0.0212	0.0134	0.0105
$(0.3,0.2)$ $\widehat{\mu}^B(\cdot,\cdot)$	Mean(RASE)	0.0917	0.0538	0.0437	0.0683	0.0495	0.0390	0.0684	0.0518	0.0381
	Std(RASE)	0.0223	0.0103	0.0064	0.0190	0.0111	0.0089	0.0212	0.0134	0.0105
$\widehat{\theta}_1$	mean	0.2794	0.2811	0.2840	0.2892	0.2936	0.2956	0.2920	0.2994	0.2957
	std	0.1033	0.0800	0.0545	0.0593	0.0414	0.0271	0.0398	0.0276	0.0179
$\widehat{\theta}_2$	mean	0.1495	0.1766	0.1961	0.1522	0.1854	0.1899	0.1780	0.1835	0.1965
	std	0.1708	0.1309	0.0933	0.1254	0.0896	0.0569	0.1032	0.0674	0.0441
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.0950	0.0608	0.0495	0.0797	0.0594	0.0447	0.0845	0.0630	0.0519
	Std(RASE)	0.0226	0.0114	0.0093	0.0196	0.0109	0.0078	0.0196	0.0144	0.0088
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0912	0.0543	0.0440	0.0679	0.0496	0.0389	0.0665	0.0508	0.0376
	Std(RASE)	0.0227	0.0106	0.0066	0.0186	0.0112	0.0088	0.0195	0.0125	0.0100
$(0.2,0.5)$ $\widehat{\mu}^B(\cdot,\cdot)$	Mean(RASE)	0.0907	0.0542	0.0440	0.0675	0.0496	0.0389	0.0665	0.0507	0.0376
	Std(RASE)	0.0227	0.0106	0.0066	0.0186	0.0112	0.0088	0.0195	0.0125	0.0100
$\widehat{\theta}_1$	mean	0.2089	0.2032	0.1981	0.2065	0.2076	0.2039	0.2056	0.2101	0.2021
	std	0.1060	0.0817	0.0552	0.0627	0.0437	0.0282	0.0420	0.0295	0.0190
$\widehat{\theta}_2$	mean	0.3801	0.4287	0.4672	0.4042	0.4511	0.4695	0.4394	0.4538	0.4788
	std	0.1812	0.1377	0.0972	0.1381	0.0977	0.0612	0.1125	0.0747	0.0484
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.1047	0.0696	0.0578	0.1144	0.0844	0.0657	0.1535	0.1178	0.0946
	Std(RASE)	0.0266	0.0149	0.0125	0.0319	0.0181	0.0124	0.0509	0.0321	0.0267
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0969	0.0583	0.0464	0.0869	0.0608	0.0460	0.1194	0.0843	0.0643
	Std(RASE)	0.0283	0.0135	0.0082	0.0350	0.0212	0.0153	0.0582	0.0414	0.0330
$(0.5,0.2)$ $\widehat{\mu}^B(\cdot,\cdot)$	Mean(RASE)	0.0970	0.0581	0.0463	0.0867	0.0608	0.0461	0.1199	0.0843	0.0644
	Std(RASE)	0.0287	0.0135	0.0082	0.0355	0.0214	0.0154	0.0587	0.0413	0.0331
$\widehat{\theta}_1$	mean	0.4596	0.4793	0.4806	0.4781	0.4951	0.4911	0.4868	0.5014	0.4965
	std	0.1020	0.0779	0.0524	0.0577	0.0382	0.0237	0.0403	0.0267	0.0146
$\widehat{\theta}_2$	mean	0.1515	0.1632	0.1940	0.1573	0.1699	0.1922	0.1713	0.1652	0.1886
	std	0.1781	0.1329	0.0931	0.1299	0.0892	0.0539	0.1162	0.0759	0.0417

Table 2: Finite-sample performance of the estimators of the unknown nonparametric function when the error structure is misspecified.

	m	4			6			8		
$(\theta_1, \theta_2, \theta_3)$	n	50	100	200	50	100	200	50	100	200
$\widehat{\mu}(\cdot,\cdot)$	Mean(RASE)	0.1026	0.0673	0.0560	0.0943	0.0693	0.0536	0.1053	0.0798	0.0632
	Std(RASE)	0.0252	0.0140	0.0119	0.0239	0.0140	0.0094	0.0276	0.0191	0.0128
$\widehat{\mu}^I(\cdot,\cdot)$	Mean(RASE)	0.0934	0.0567	0.0456	0.0733	0.0533	0.0413	0.0782	0.0590	0.0436
	Std(RASE)	0.0256	0.0121	0.0075	0.0235	0.0146	0.0110	0.0285	0.0196	0.0156

Table 3: Correct model identification probability.

		m	4			6			8		
(θ_1, θ_2)	c	n	50	100	200	50	100	200	50	100	200
	$c = 0$		0.98	0.97	0.99	0.99	0.97	0.99	0.99	0.98	0.99
	$c = 0.25$		0.06	0.14	0.78	0.21	0.36	0.90	0.20	0.75	0.88
$(0.0, 0.0)$	$c = 0.50$		0.71	1.00	1.00	0.92	1.00	1.00	0.97	1.00	1.00
	$c = 0.75$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.00$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.50$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 0$		0.94	0.97	0.96	0.95	0.94	0.97	0.89	0.92	0.93
	$c = 0.25$		0.10	0.14	0.62	0.19	0.57	0.67	0.30	0.48	0.73
$(0.2, 0.3)$	$c = 0.50$		0.66	1.00	1.00	0.80	1.00	1.00	0.95	0.99	1.00
	$c = 0.75$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.00$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.50$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 0$		0.93	0.97	0.96	0.92	0.92	0.96	0.82	0.88	0.91
	$c = 0.25$		0.12	0.15	0.55	0.19	0.53	0.58	0.31	0.34	0.74
$(0.3, 0.2)$	$c = 0.50$		0.62	1.00	1.00	0.80	1.00	1.00	0.89	0.92	1.00
	$c = 0.75$		0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.00$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 1.50$		1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$c = 0$		0.93	0.94	0.93	0.78	0.83	0.86	0.70	0.63	0.63
	$c = 0.25$		0.19	0.24	0.36	0.41	0.35	0.57	0.40	0.50	0.52
$(0.5, 0.3)$	$c = 0.50$		0.40	0.97	0.99	0.61	0.72	1.00	0.54	0.68	0.79
	$c = 0.75$		0.90	1.00	1.00	0.89	0.98	1.00	0.69	0.89	0.97
	$c = 1.00$		1.00	1.00	1.00	0.98	1.00	1.00	0.89	0.97	1.00
	$c = 1.50$		1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00

5 Real data study

We now apply the proposed method to analyze a data set from a panel hormone study [26]. The study involved 34 women whose urine samples were collected in one menstrual cycle and whose urinary progesterone was assayed on alternate days. A total of 492 observations were obtained, with each woman contributing from 11 to 28 observations over time. Each woman's cycle length was standardized uniformly to a reference 28-day cycle since the change of the progesterone level for each woman depends on the time during a menstrual cycle. Zhang et al. [26] and He et al. [7] used the longitudinal partially linear regression model to fit this data set. See Zhang et al. [26] for a detailed description of this data set. To fit these data, we propose the nonparametric regression model

$$y_{ij} = \mu(t_{ij}, Z_{ij}) + \varepsilon_{ij},$$

where y, t and Z stand for *progesterone value*, *day* and *bmi*, respectively. In the above model, we furthermore suppose that the covariates for the autoregressive components take the form $\mathbf{W}_{j,\ell}^{(i)} = (1, t_{il} - t_{ik})^\top$. Then applying our methods to this data set we find $\widehat{\theta}_1 = 0.2183$ (0.0190), $\widehat{\theta}_2 = -0.3736$ (0.0464). The nonparametric estimates of the functions of *day* and *bmi* are plotted in Figure 1.

Figure 1 demonstrates that *day* and *bmi* affect the *progesterone value* nonlinearly, and our proposed test method shows that they are really nonlinear, where Figures 2 and 3 give the biases of the two methods and 95% pointwise confidence band for $\mu(t, Z)$ over time with difference Z with and without considering the within correlation.

In addition, we also calculated the sum of squared errors (SSE) for the partially linear regression model used in He et al. [7], which is 270.6651. The SSE for our bivariate regression model is 241.6877. We also compared the prediction accuracy of our bivariate nonparametric model (BNPM) with that of partially linear model (PLM) in Zhu and Fung [7]. We randomly split the 34 subjects into a training set and a test set with corresponding proportions π and $1-\pi$, respectively. We used the training set to estimate both of the PLM and BNPM, and then predict the response of the test set. 500 replications were used to calculate the prediction errors corresponding to $\pi = 0.3$, $\pi = 0.5$ and $\pi = 0.7$. Table 4 reports the root mean square error (RMSE) of the prediction errors and indicates that our model seems suitable for these data.

6 Concluding remarks

In this paper, we have investigated the efficient estimation and model identification for a nonparametric model with a time-varying regression function and longitudinal covariate. In order to improve

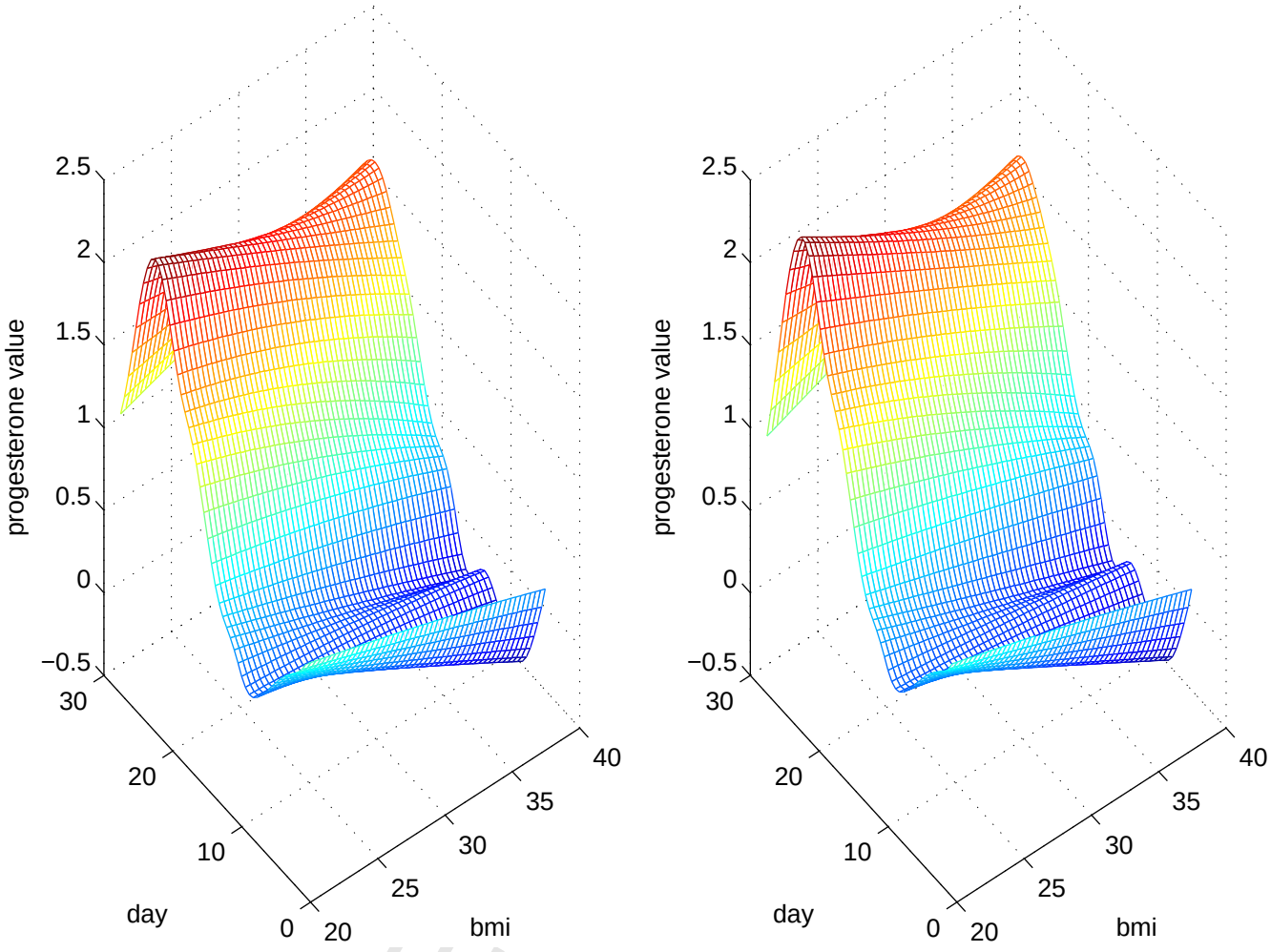


Figure 1: Estimation of the bivariate nonparametric of *day* and *bmi*. The left panel exhibits the estimated function which does not consider the within-cluster correlation. The right panel depicts the function based on our proposed method.

Table 4: Prediction results for the hormone study.

		$\pi = 0.3$	$\pi = 0.5$	$\pi = 0.7$
NPM	Mean(RMSE)	0.1207	0.1007	0.0985
	Std(RMSE)	0.1184	0.0882	0.0791
PLM	Mean(RMSE)	0.1919	0.1548	0.1806
	Std(RMSE)	0.1418	0.1192	0.1379

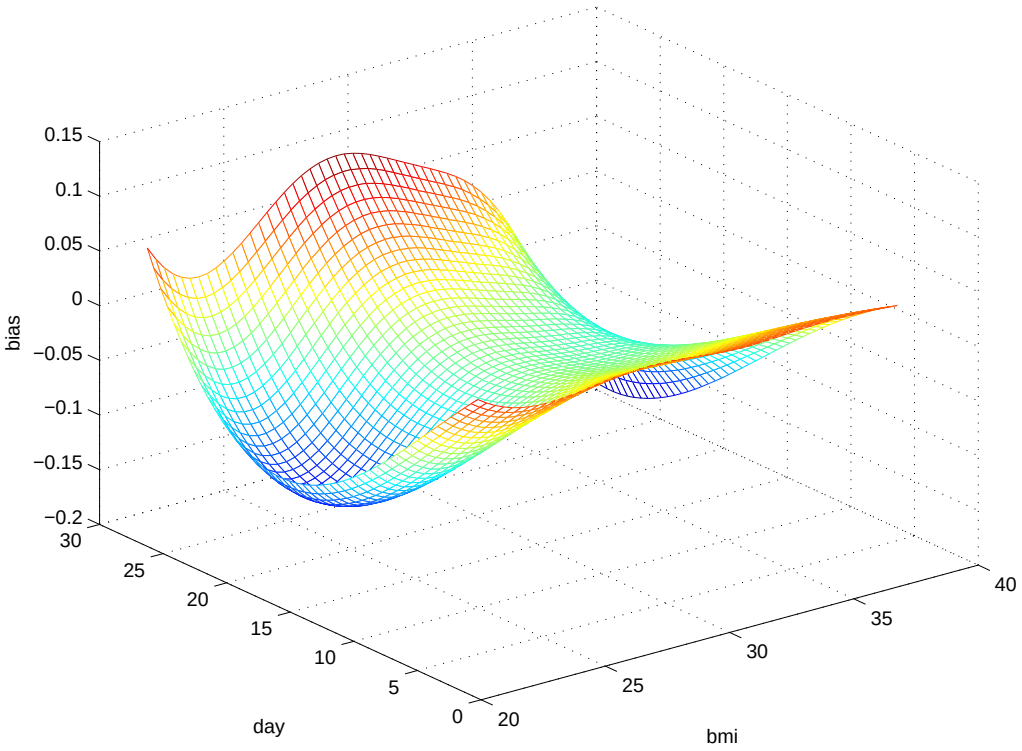


Figure 2: Biases of the two method.

estimation efficiency by considering within-subject correlation among repeated measures over time, as in [24], we used a modified Cholesky decomposition to decompose the within-subject covariance matrix into a unit triangular matrix involving generalized autoregressive coefficients and a diagonal matrix involving innovation variances. A local polynomial smoothing method was then used to estimate the unknown nonparametric functions of the marginal mean. The asymptotic theory of the resultant estimators was developed as well. In addition, we proposed a shrink-based method to tackle the model identification problem, which could identify the model consistently and estimate the model simultaneously.

In closing, we mention possible future directions. Throughout this paper, we assumed that the covariate Z_{ij} is univariate. It is easy to see that our proposed methods could be extended to the scenario of the covariate Z_{ij} being two or three-dimensional. However, when the dimension of the covariate Z_{ij} is greater than 3, our proposed methods will be subject to the so-called "curse of dimensionality." As in Jiang andWang [9] and Zhang et al. [28], one could overcome this challenge and achieve both dimension reduction and sensible model interpretation through single index or additive nonparametric

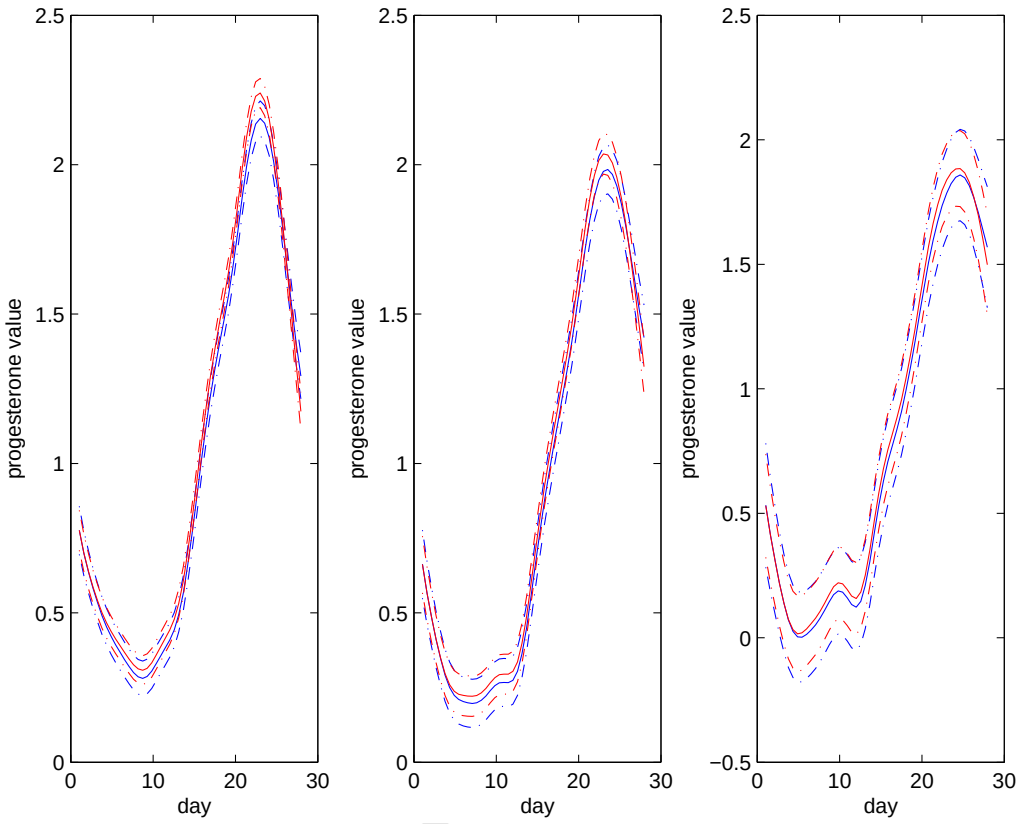


Figure 3: 95% pointwise confidence band for $\mu(t, Z)$ over time with $Z = 24.5450$, $Z = 28.8979$ and $Z = 33.6136$, which are the lower, middle and upper quartiles of the *bmi* range, respectively. The red solid lines represent the estimated nonparametric function considering the within-subject correlation, and the red broken lines represent the 95% pointwise confidence band for it. Finally, the blue lines represent the estimated nonparametric function without considering the within-subject correlation, and the blue broken lines delineate the 95% pointwise confidence band for it.

regression modeling. The methods developed in this paper were mainly built upon (local) least squares. It is well known that least squares technique is very sensitive to outliers and for many commonly used non-normal errors, such as a mixture normal distribution, Laplace distribution, Cauchy distribution and so on, its efficiency may be significantly ruined. An effective way to cope with this issue is to develop robust estimation and method of model identification.

Acknowledgments

The authors are grateful to the Editor-in-Chief for substantial help in improving the readability of the manuscript. Liu's research was supported by grants from the National Natural Science Foundation of China (NSFC) (No. 11626154). You's research was supported by grants from the National Natural Science Foundation of China (NSFC) (No. 11471203) and Program for Innovative Research Team of Shanghai University of Finance and Economics (IRTSHUFE). This work was also partially supported by the Program for Changjiang Scholars and Innovative Research Team in University (IRT13077).

Appendix. Proof of the main results

For easy reference, we start by introducing a lemma that is needed to prove the main results.

Lemma 1. *Under Assumptions 1, 3, 4 and 6, for* $k_1 + k_2 = k \in \{0, 1, 2, 4\}$,

$$\sup_{t\in\mathcal{T}} \left| \frac{1}{Nhthz} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} K\left(\frac{t_{ij}-t}{ht}\right) K\left(\frac{Z_{ij}-z}{hz}\right) \left(\frac{t_{ij}-t}{ht}\right)^{k_1} \left(\frac{Z_{ij}-z}{hz}\right)^{k_2} - f(t,z)\mu_k \right|$$
$$= O_p\left\{ h_t^2 + h_z^2 + \left(\frac{\ln N}{Nhthz}\right)^{1/2} \right\},$$

and

$$\sup_{t\in\mathcal{T}} \frac{1}{Nhthz} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} K\left(\frac{t_{ij}-t}{ht}\right) K\left(\frac{Z_{ij}-z}{hz}\right) \left(\frac{t_{ij}-t}{ht}\right)^{k_1} \left(\frac{Z_{ij}-z}{hz}\right)^{k_2} \varepsilon_{ij} = O_p\left\{ \left(\frac{\ln N}{Nhthz}\right)^{1/2} \right\},$$

as $n \to \infty$, *as long as* ht *and* hz *satisfy Assumption 6.*

Proof. Lemma 1 follows immediately from results of Mack and Silverman [16] and Masry [17]. □

Proof of Theorem 1. The proof of Theorem 1 can be found in Masry [17]. We omit the details. □

Proof of Theorem 2. By the definition of $\widehat{\boldsymbol{\zeta}}_{ij}$, we have

$$\widehat{\boldsymbol{\Pi}}_{ij} - \boldsymbol{\Pi}_{ij} = \left(\sum_{k=1}^{j-1} \{\mu(t_{ik}, Z_{ik}) - \widehat{\mu}(t_{ik}, Z_{ik})\} W_{j,k,1}^{(i)}, \ldots, \sum_{k=1}^{j-1} \{\mu(t_{ik}, Z_{ik}) - \widehat{\mu}(t_{ik}, Z_{ik})\} W_{j,k,q}^{(i)} \right)^{\top}.$$

Thus an application of Theorem 1 yields

$$\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})^\top = O_p\left[\left\{ht_{1,N}^2+hz_{1,N}^2+\left(\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\}^2\right]\mathbf{I}_q.$$

Similarly,

$$\begin{aligned}\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\widehat{\mathbf{\Pi}}_{ij}^\top &= \frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\mathbf{\Pi}_{ij}^\top+\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})^\top\\ &= \frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\mathbf{\Pi}_{ij}^\top+O_p\left[\left\{ht_{1,N}^2+hz_{1,N}^2+\left(\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\}^2\right]\mathbf{I}_q,\end{aligned}$$

and

$$\left|\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\mathbf{\Pi}_{ij}^\top\right| = O_p\left[\left\{ht_{1,N}^2+hz_{1,N}^2+\left(\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\}^2\right]\mathbf{I}_q.$$

These imply

$$\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\widehat{\mathbf{\Pi}}_{ij}^\top = O_p\left[\left\{ht_{1,N}^2+hz_{1,N}^2+\left(\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\}^2\right]\mathbf{I}_q.$$

Due to the fact that

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=2}^{m_i}\widehat{\mathbf{\Pi}}_{ij}\widehat{\mathbf{\Pi}}_{ij}^\top=\sum_{i=1}^{n}\sum_{j=2}^{m_i}\{(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})+\mathbf{\Pi}_{ij}\}\{(\widehat{\mathbf{\Pi}}_{ij}^\top-\mathbf{\Pi}_{ij}^\top)+\mathbf{\Pi}_{ij}^\top\}\\ &=\sum_{i=1}^{n}\sum_{j=2}^{m_i}\{(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})^\top+\ (\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\widehat{\mathbf{\Pi}}_{ij}^\top+\mathbf{\Pi}_{ij}(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})^\top+\mathbf{\Pi}_{ij}\mathbf{\Pi}_{ij}^\top\},\end{aligned}$$

we see that, as $n\to\infty$,

$$\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\widehat{\mathbf{\Pi}}_{ij}\widehat{\mathbf{\Pi}}_{ij}^\top=\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\mathbf{\Pi}_{ij}\mathbf{\Pi}_{ij}^\top+o_p(1)\mathbf{I}_q\to_p\mathbf{\Lambda}.$$

Write

$$\sum_{i=1}^{n}\sum_{j=2}^{m_i}\widehat{\mathbf{\Pi}}_{ij}\widehat{\varepsilon}_{ij} = \sum_{i=1}^{n}\sum_{j=2}^{m_i}\{(\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})(\widehat{\varepsilon}_{ij}-\varepsilon_{ij}) + (\widehat{\mathbf{\Pi}}_{ij}-\mathbf{\Pi}_{ij})\varepsilon_{ij} + \mathbf{\Pi}_{ij}(\widehat{\varepsilon}_{ij}-\varepsilon_{ij}) + \mathbf{\Pi}_{ij}\varepsilon_{ij}\}$$
$$\equiv \sqrt{N-n}\,(J_1+J_2+J_3+J_4).$$

Based on Theorem 1, we have

$$J_1 = O_p\left[\left\{ht_{1,N}^2 + hz_{1,N}^2 + \left(\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\}^2\right]\mathbf{1}_q = o_p(1)\mathbf{1}_q.$$

By the definition of $\widehat{\mu}(t,z)$,

$$\widehat{\mu}(t,z)-\mu(t,z) = \mathrm{diag}(1,0,0)(\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}\mathbf{D}_{t,z})^{-1}\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}\boldsymbol{\varepsilon} + \mathrm{diag}(1,0,0)(\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}\mathbf{D}_{t,z})^{-1}\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}$$
$$\times(\mu(t_{11},Z_{11}),\ldots,\mu(t_{1m_1},Z_{1m_1}),\ldots,\mu(t_{nm_n},Z_{nm_n}))^\top - \mu(t,z).$$

Note that each element of $\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}\mathbf{D}_{t,z}$ has the form of kernel regression, hence by Lemma 1, we have

$$\frac{1}{N}\mathbf{D}_{t,z}^\top\mathbf{W}_{t,z}\mathbf{D}_{t,z} = f(t,z)\mathrm{diag}(1,\mu_2,\mu_2)O_p\left[1+\left\{\frac{\ln N}{Nht_{1,N}hz_{1,N}}\right\}^{1/2}\right]. \tag{9}$$

Therefore, combining the proofs of Lemma A.5 and A.6 of Liang et al. [13], we get

$$\frac{1}{\sqrt{N-n}}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\sum_{k=1}^{j-1}\mathbf{W}_{j,k}^{(i)}\mathrm{diag}(1,0,0)(\mathbf{D}_{t_{ik},z_{ik}}^\top\mathbf{W}_{t_{ik},z_{ik}}\mathbf{D}_{t_{ik},z_{ik}})^{-1}\mathbf{D}_{t_{ik},z_{ik}}^\top\mathbf{W}_{t_{ik},z_{ik}}\boldsymbol{\varepsilon}\varepsilon_{ij} = o_p(1).$$

In addition, $\mu(t,z)$ are smooth in the neighborhood of $|t_{ij}-t| < ht_{1,N}$ and $|Z_{ij}-z| < hz_{1,N}$, so that

$$\mu(t_{ij},Z_{ij}) = \mu(t,z) + ht_{1,N}\frac{\partial\mu(t,z)}{\partial t}\frac{t_{ij}-t}{ht_{1,N}} + hz_{1,N}\frac{\partial\mu(t,z)}{\partial z}\frac{Z_{ij}-z}{hz_{1,N}} + ht_{1,N}hz_{1,N}\frac{\partial^2\mu(t,z)}{\partial t\partial z}$$
$$+ ht_{1,N}^2\frac{\partial^2\mu(t,z)}{2\partial t^2}\left(\frac{t_{ij}-t}{ht_{1,N}}\right)^2 + hz_{1,N}^2\frac{\partial^2\mu(t,z)}{2\partial z^2}\left(\frac{Z_{ij}-z}{hz_{1,N}}\right)^2 + o_p(ht_{1,N}^2+hz_{1,N}^2), \tag{10}$$

where $\partial\mu(t,z)/\partial t$ and $\partial\mu(t,z)/\partial z$, $\partial^2\mu(t,z)/\partial t^2$ and $\partial^2\mu(t,z)/\partial z^2$ are the first and the second deriva-

tives of $\mu(t,z)$. As a result,

$$\frac{1}{\sqrt{N-n}}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\sum_{k=1}^{j-1}\mathbf{W}_{j,k}^{(i)}\{\mathrm{diag}(1,0,0)(\mathbf{D}_{t_{ik},z_{ik}}^{\top}\mathbf{W}_{t_{ik},z_{ik}}\mathbf{D}_{t_{ik},z_{ik}})^{-1}\mathbf{D}_{t_{ik},z_{ik}}^{\top}\mathbf{W}_{t_{ik},z_{ik}}$$
$$\times(\mu(t_{11},Z_{11}),\ldots,\mu(t_{1m_1},Z_{1m_1}),\ldots,\mu(t_{nm_n},Z_{nm_n}))^{\top}-\mu(t_{ik},z_{ik})\}\varepsilon_{ij}=o_p(1).$$

It follows that $J_2=o_p(1)$ and $J_3=o_p(1)$. Furthermore,

$$J_4=\frac{1}{\sqrt{N-n}}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\mathbf{\Pi}_{ij}\varepsilon_{ij}=\frac{1}{\sqrt{N-n}}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\mathbf{\Pi}_{ij}\mathbf{\Pi}_{ij}^{\top}\boldsymbol{\theta}+\frac{1}{\sqrt{N-n}}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\mathbf{\Pi}_{ij}e_{ij}.$$

For any $q\times 1$ constant vector $\boldsymbol{\kappa}=(\kappa_1,\ldots,\kappa_q)^{\top}$, let $Z_d=\sum_{i=1}^{n}\sum_{j=2}^{m_i}(\boldsymbol{\kappa}^{\top}\mathbf{\Pi}_{ij})e_{ij}/(N-n)$. It is obvious that $\mathrm{E}(Z_d)=0$ and

$$\mathrm{var}(\sqrt{N-n}Z_d)=\frac{1}{N-n}\sum_{i=1}^{n}\mathrm{E}\Big\{\sum_{j=2}^{m_i}(\boldsymbol{\kappa}^{\top}\mathbf{\Pi}_{ij})e_{ij}\Big\}^2=\frac{1}{N-n}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\sigma^2\mathrm{E}\Big(\sum_{k=1}^{q}\kappa_k\pi_{ijk}\Big)^2.$$

Denote $\xi_i=\sum_{j=2}^{m_i}(\boldsymbol{\kappa}^{\top}\mathbf{\Pi}_{ij})e_{ij}$. Then

$$S^2=\sum_{i=1}^{n}\mathrm{var}(\xi_i)=\sum_{i=1}^{n}\sum_{j=2}^{m_i}\mathrm{E}(\boldsymbol{\kappa}^{\top}\mathbf{\Pi}_{ij})^2\sigma^2.$$

This leads to
$$\frac{\sum_{i=1}^{n}\mathrm{E}|\xi_i|^3}{S^3}\le\frac{Cn\mathrm{E}(\sum_{k=1}^{q}|\kappa_k|\cdot|\pi_{ijk}|)^3}{n^{3/2}[\mathrm{E}\{\sum_{k=1}^{q}\kappa_k\pi_{ijk}\}^2]^{3/2}}=O_p\Big(\frac{1}{\sqrt{N}}\Big)=o_p(1).$$

So Theorem 2 follows from the Lyapunov Central Limit Theorem. □

Proof of Theorem 3. From the expression of $(\widehat{\mu}^{TS}(t,z),\partial\widehat{\mu}^{TS}(t,z)/\partial t,\partial\widehat{\mu}^{TS}(t,z)/\partial z)^{\top}$, it follows

that

$$\begin{aligned}
&\mathbf{H}_{2,N}\left[\{\widehat{\mu}^{TS}(t,z),\frac{\partial\widehat{\mu}^{TS}(t,z)}{\partial t},\frac{\partial\widehat{\mu}^{TS}(t,z)}{\partial z}\}^{\top}-\{\mu(t,z),\frac{\partial\mu(t,z)}{\partial t},\frac{\partial\mu(t,z)}{\partial z}\}^{\top}\right]\\
&\quad=\Bigg\{(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)\mu(t_{ij},Z_{ij})-\\
&\qquad+\{\mu(t,z),ht_{2,N}\frac{\partial\mu(t,z)}{\partial t},hz_{2,N}\frac{\partial\mu(t,z)}{\partial z}\}^{\top}\Bigg\}\\
&\qquad+(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)e_{ij}\\
&\qquad+(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)(\mathbf{\Pi}_{ij}^{\top}\boldsymbol{\theta}-\widehat{\mathbf{\Pi}}_{ij}^{\top}\widehat{\boldsymbol{\theta}})\\
&\quad\equiv J_1+J_2+J_3.
\end{aligned}$$

Similar to (9),

$$\frac{1}{N}\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*}=f(t,z)\mathrm{diag}(1,\mu_2,\mu_2)O_p\left[1+\left\{\frac{\ln N}{Nht_{2,N}hz_{2,N}}\right\}^{1/2}\right],$$

with probability approaching 1. Combining with (10), we have

$$J_1=\left(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*}\right)^{-1}\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{B}_{t,z}+o_p(ht_{2,N}^2+hz_{2,N}^2),$$

where

$$\begin{aligned}
\mathbf{B}_{t,z}=\frac{ht_{2,N}^2}{2}\left(\left(\frac{t_{11}-t}{ht_{2,N}}\right)^2,\ldots,\left(\frac{t_{nm_n}-t}{ht_{2,N}}\right)^2\right)^{\top}\times\frac{\partial^2\mu(t,z)}{\partial t^2}\\
+\frac{hz_{2,N}^2}{2}\left(\left(\frac{Z_{11}-t}{ht_{2,N}}\right)^2,\ldots,\left(\frac{Z_{nm_n}-t}{ht_{2,N}}\right)^2\right)^{\top}\times\frac{\partial^2\mu(t,z)}{\partial z^2}.
\end{aligned}$$

By Lemma 1, we have

$$\frac{1}{N}\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{B}_{t,z} = f(t,z)\bigg(\Big(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{2\partial t} + hz_{2,N}^2\frac{\partial^2\mu(t,z)}{2\partial z}\Big)\mu_2, \\ hz_{2,N}^2\frac{\partial^2\mu(t,z)}{2\partial z^2}\mu_2, ht_{2,N}^2\frac{\partial^2\mu(t,z)}{2\partial t^2}\mu_2\bigg)^{\top}.$$

Combining with the fact $(\mathbf{K}+a\mathbf{M})^{-1} = \mathbf{K}^{-1} - a\mathbf{K}^{-1}\mathbf{M}\mathbf{K}^{-1} + O(a^2)$ as $a \to 0$, we get

$$\left(\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*}\right)^{-1}\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{B}_{t,z} = \frac{1}{2}\begin{pmatrix} \Big(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t} + hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z}\Big)\mu_2 \\ hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z^2} \\ ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t^2} \end{pmatrix}\{1+o_p(1)\}.$$

Therefore,

$$\sqrt{Nht_{2,N}hz_{2,N}}\left\{J_1 - \frac{1}{2}\begin{pmatrix} \Big(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t} + hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z}\Big)\mu_2 \\ hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z^2} \\ ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t^2} \end{pmatrix} + o_p(ht_{2,N}^2 + hz_{2,N}^2)\right\} = o_p(1).$$

Next we show that, as $n \to \infty$,

$$\sqrt{Nht_{2,N}hz_{2,N}}J_2 \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma}^*). \tag{11}$$

For any non-zero constant vector $(d_1, d_2, d_3)^{\top}$, let

$$Z_{t,z} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{d_1 + d_2\left(\frac{t_{ij}-t}{ht_{2,N}}\right) + d_3\left(\frac{Z_{ij}-z}{hz_{2,N}}\right)\right\}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)e_{ij}.$$

Obviously, $\mathrm{E}(Z_{t,z}) = 0$ and

$$
\begin{aligned}
&\mathrm{var}(\sqrt{Nht_{2,N}hz_{2,N}}Z_{t,z}) \\
&= \frac{ht_{2,N}hz_{2,N}}{N}\sum_{i=1}^{n}\mathrm{E}\sum_{j=1}^{m_i}\left[\left\{d_1 + d_2\left(\frac{t_{ij}-t}{ht_{2,N}}\right) + d_3\left(\frac{Z_{ij}-z}{hz_{2,N}}\right)\right\}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)e_{ij}\right]^2 \\
&+ \frac{ht_{2,N}hz_{2,N}}{N}\sum_{i=1}^{n}\sum_{j_1\neq j_2}\mathrm{E}\left[\left\{d_1 + d_2\left(\frac{t_{i,j_1}-t}{ht_{2,N}}\right) + d_3\left(\frac{Z_{i,j_1}-z}{hz_{2,N}}\right)\right\}\left\{d_1 + d_2\left(\frac{t_{i,j_2}-t}{ht_{2,N}}\right) + d_3\left(\frac{Z_{i,j_2}-z}{hz_{2,N}}\right)\right\}\right. \\
&\left.\times K_{ht_{2,N}}(t_{i,j_1}-t)K_{hz_{2,N}}(Z_{i,j_1}-z)K_{ht_{2,N}}(t_{i,j_2}-t)K_{hz_{2,N}}(Z_{i,j_2}-z)e_{i,j_1}e_{i,j_2}\right] \\
&\equiv Q_1 + Q_2.
\end{aligned}
$$

As $n \to \infty$, we have

$$
\begin{aligned}
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\right\} \to f(t,z)\nu_0^2, \\
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\left(\frac{t_{ij}-t}{ht_{2,N}}\right)^2\right\} \to f(t,z)\nu_0\nu_2^2, \\
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\left(\frac{Z_{ij}-z}{hz_{2,N}}\right)^2\right\} \to f(t,z)\nu_0\nu_2^2, \\
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\left(\frac{t_{ij}-t}{ht_{2,N}}\right)\right\} \to f(t,z)\nu_0\nu_1, \\
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\left(\frac{Z_{ij}-z}{hz_{2,N}}\right)\right\} \to f(t,z)\nu_0\nu_1, \\
&ht_{2,N}hz_{2,N}\mathrm{E}\left\{K^2_{ht_{2,N}}(t_{ij}-t)K^2_{hz_{2,N}}(Z_{ij}-z)\left(\frac{t_{ij}-t}{ht_{2,N}}\right)\left(\frac{Z_{ij}-z}{hz_{2,N}}\right)\right\} \to f(t,z)\nu_1^2.
\end{aligned}
$$

Hence

$$
\lim_{n\to\infty} Q_1 = f(t,z)\sigma^2\{d_1^2\nu_0^2 + (d_2^2+d_3^2)\nu_0\nu_2^2 + (2d_1d_2+2d_1d_3)\nu_0\nu_1 + 2d_2d_3\nu_1^2\}.
$$

Similarly, for $j_1 \neq j_2$, we can show that $Q_2 \to 0$ as $n \to \infty$. Further, let $S_N^2 = \sum_{i=1}^n \mathrm{var}(\phi_i)$ with

$$
\phi_i = \sqrt{ht_{2,N}hz_{2,N}}\sum_{j=1}^{m_i}\left\{d_1 + d_2\left(\frac{t_{ij}-t}{h_{2,N}}\right) + d_3\left(\frac{Z_{ij}-z}{h_{2,N}}\right)\right\}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)e_{ij}.
$$

Then

$$S_N^2 = f(t,z)\sigma^2\{d_1^2\nu_0^2 + (d_2^2+d_3^2)\nu_0\nu_2^2 + (2d_1d_2+2d_1d_3)\nu_0\nu_1 + 2d_2d_3\nu_1^2\} + o(n),$$

which together with the fact that

$$\sum_{i=1}^{n}\mathrm{E}|\phi_i|^3 \le O(1)\times\sum_{i=1}^{n}(ht_{2,N}hz_{2,N})^{3/2}\mathrm{E}\left\{|d_1|+|d_2|\cdot\left|\frac{t_{ij}-t}{ht_{2,N}}\right|+|d_3|\cdot\left|\frac{Z_{ij}-z}{hz_{2,N}}\right|\right\}^3$$
$$\times K_{ht_{2,N}}^3(t_{ij}-t)K_{hz_{2,N}}^3(Z_{ij}-z) = O(nht_{2,N}^{-1/2}hz_{2,N}^{-1/2}),$$

implies that the Lyapunov condition $\lim_{n\to\infty} S_N^{-3}\sum_{i=1}^{n}\mathrm{E}|\phi_i|^3 = 0$ is satisfied. Hence (11) is proved.

It remains to show that $J_3 = o_p[ht_{2,N}^2 + hz_{2,N}^2 + \{1/(Nht_{2,N}hz_{2,N})\}^{1/2}]$. Express J_3 as $J_3 = J_{31} + J_{32} - J_{33}$ with

$$J_{31} = (\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)(\mathbf{\Pi}_{ij}^{\top}-\widehat{\mathbf{\Pi}}_{ij}^{\top})\boldsymbol{\theta},$$

$$J_{32} = (\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)(\mathbf{\Pi}_{ij}^{\top}-\widehat{\mathbf{\Pi}}_{ij}^{\top})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}),$$

and

$$J_{33} = (\mathbf{D}_{t,z}^{*\top}\mathbf{W}_{t,z}^{*}\mathbf{D}_{t,z}^{*})^{-1}\sum_{i=1}^{n}\sum_{j=2}^{m_i}\begin{pmatrix}1\\ \frac{t_{ij}-t}{ht_{2,N}}\\ \frac{Z_{ij}-z}{hz_{2,N}}\end{pmatrix}K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)\mathbf{\Pi}_{ij}^{\top}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}).$$

By Theorem 2 and following the arguments proving (11), it is easy to see that

$$J_{33} = O_p\left(\frac{1}{\sqrt{N}}\right)\times O_p\left(\frac{1}{\sqrt{Nht_{2,N}hz_{2,N}}}\right) = o_p\left\{ht_{2,N}^2 + hz_{2,N}^2 + \left(\frac{1}{Nht_{2,N}hz_{2,N}}\right)^{1/2}\right\}.$$

Combining Theorem 1 and 2 we can show that

$$J_{32} = O_p\left(\frac{1}{\sqrt{N}}\right)\times O_p\left\{ht_{1,N}^2+hz_{1,N}^2+\left(\frac{1}{Nht_{1,N}hz_{1,N}}\right)^{1/2}\right\} = o_p\left\{ht_{2,N}^2+hz_{2,N}^2+\left(\frac{1}{Nht_{2,N}hz_{2,N}}\right)^{1/2}\right\}.$$

According to the proof of Theorem 1,

$$\begin{aligned}
&\widehat{\mu}(t_{ij}, Z_{ij}) - \mu(t_{ij}, Z_{ij}) \\
&= \mathrm{diag}(1,0,0)\left\{f(t_{ij}, Z_{ij})\mathbf{S}\right\}^{-1}\begin{pmatrix} \frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_i} K_{ht_{1,N}}(t_{i_1,j_1}-t_{ij})K_{hz_{1,N}}(Z_{i_1,j_1}-z_{ij})\varepsilon_{i_1,j_1} \\ \frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_i}\left(\frac{t_{i_1,j_1}-t_{ij}}{h_{1,N}}\right) K_{ht_{1,N}}(t_{i_1,j_1}-t_{ij})K_{hz_{1,N}}(Z_{i_1,j_1}-z_{ij})\varepsilon_{i_1,j_1} \\ \frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_i}\left(\frac{Z_{i_1,j_1}-z_{ij}}{h_{1,N}}\right) K_{ht_{1,N}}(t_{i_1,j_1}-t_{ij})K_{hz_{1,N}}(Z_{i_1,j_1}-z_{ij})\varepsilon_{i_1,j_1} \end{pmatrix} \\
&\quad + \mathrm{diag}(1,0,0)\left\{f(t_{ij}, Z_{ij})\mathbf{S}\right\}^{-1}\begin{pmatrix} f(t_{ij}, Z_{ij})\Big(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{2\partial t}+hz_{2,N}^2\frac{\partial^2\mu(t_{ij},Z_{ij})}{2\partial z}\Big)\mu_2 \\ f(t_{ij}, Z_{ij})hz_{2,N}^2\frac{\partial^2\mu(t_{ij},Z_{ij})}{2\partial z^2}\mu_2 \\ f(t_{ij}, Z_{ij})ht_{2,N}^2\frac{\partial^2\mu(t_{ij},Z_{ij})}{2\partial t^2}\mu_2 \end{pmatrix} + o_p(ht_{1,N}^2+hz_{1,N}^2) \\
&= \mathbf{G}_1(t_{ij}, Z_{ij}) + \mathbf{G}_2(t_{ij}, Z_{ij}) + o_p(ht_{1,N}^2+hz_{1,N}^2),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{G}_1(t_{ij}, Z_{ij}) &= \frac{1}{f(t_{ij}, Z_{ij})}\frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_i} K_{ht_{1,N}}(t_{i_1,j_1}-t_{ij})K_{hz_{1,N}}(Z_{i_1,j_1}-z_{ij})\varepsilon_{i_1,j_1}, \\
\mathbf{G}_2(t_{ij}, Z_{ij}) &= \frac{1}{2}\Big(ht_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial t}+hz_{2,N}^2\frac{\partial^2\mu(t,z)}{\partial z}\Big)\mu_2 + o_p(ht_{1,N}^2+hz_{1,N}^2)
\end{aligned}$$

and $\mathbf{S} = \mathrm{diag}(1, \mu_2, \mu_2)$. Then we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{n}\sum_{j=2}^{m_i} K_{ht_{2,N}}(t_{ij}-t)K_{hz_{2,N}}(Z_{ij}-z)\sum_{k=1}^{q}\Big\{\sum_{d=1}^{j-1}\mathbf{G}_1(t_{ij}, Z_{ij})\mathbf{W}_{j-1,d,k}^{(i)}\Big\}\theta_k \\
&= \frac{1}{N}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_i}\varepsilon_{i_1,j_1}\Phi_{i_1,j_1},
\end{aligned}$$

where

$$\Phi_{i_1,j_1} = \frac{1}{Nf(t_{ij}, Z_{ij})} \sum_{i=1}^{n} \sum_{j=2}^{m_i} K_{ht_{2,N}}(t_{ij} - t) K_{hz_{2,N}}(Z_{ij} - z) \times \sum_{k=1}^{q} \Big\{ \sum_{d=1}^{j-1} K_{ht_{1,N}}(t_{i_1,j_1} - t_{ij}) K_{hz_{1,N}}(Z_{i_1,j_1} - z_{ij}) \mathbf{W}^{(i)}_{j-1,d,k} \Big\} \theta_k.$$

Since ε_{i_1,j_1} and Φ_{i_1,j_1} are independent and Assumption 4.2 implies that Φ_{i_1,j_1} is bounded,

$$\frac{1}{N} \sum_{i=1}^{n} \sum_{j=2}^{m_i} K_{ht_{2,N}}(t_{ij} - t) K_{hz_{2,N}}(Z_{ij} - z) \sum_{k=1}^{q} \Big\{ \sum_{d=1}^{j-1} \mathbf{Z}_{ij}^{\top} \mathbf{G}_1(t_{ij}, Z_{ij}) \mathbf{W}^{(i)}_{j-1,d,k} \Big\} \theta_k = O_p(N^{-1/2}).$$

Moreover, invoking Lemma 1, we get

$$\frac{1}{N} \sum_{i=1}^{n} \sum_{j=2}^{m_i} K_{ht_{2,N}}(t_{ij} - t) K_{hz_{2,N}}(Z_{ij} - z) \sum_{k=1}^{q} \Big[\sum_{d=1}^{j-1} \{ \mathbf{G}_2(t_{i,j-k}, Z_{i,j-k}) + o_p(ht_{1,N}^2 + hz_{1,N}^2) \} \mathbf{W}^{(i)}_{j-1,d,k} \Big] \theta_k = O_p(ht_{1,N}^2 + hz_{1,N}^2) = o_p(ht_{2,N}^2 + hz_{2,N}^2).$$

Thus J_{31} is of order $o_p[ht_{2,N}^2 + hz_{2,N}^2 + \{\ln N/(Nht_{2,N}hz_{2,N})\}^{1/2}]$, and so is J_3. Hence the proof of Theorem 3 is complete. □

Lemma 2. *Suppose that Assumption 1 to 6 in the Appendix hold and $\lambda \to 0$ as $n \to \infty$. Then we have*

$$\frac{1}{N^2} \sum_{i_2=1}^{n} \sum_{j_2=1}^{m_{i2}} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i1}} \|\widehat{\mu}_\lambda(t_{i_1 j_1}, Z_{i_2 j_2}) - \mu(t_{i_1 j_1}, Z_{i_2 j_2})\|^2 = O_p\{(Nht_{3,N}hz_{3,N})^{-1}\}.$$

Proof. For an arbitrary matrix $\mathbf{A} = (a_{ij})$, we define its norm as $\|\mathbf{A}\|^2 = \sum a_{ij}^2$. We use $u = (u_{ij}) \in \mathbb{R}^{N \times N}$ to denote an arbitrary $N \times N$ matrix and $vu = \mathrm{vec}(u)$. Let

$$B_0 = (\mu_0(t_{11}, Z), \ldots, \mu_0(t_{1m_1}, Z), \ldots, \mu_0(t_{nm_n}, Z))^{\top} \in \mathbb{R}^{N \times N}.$$

Thus, by Fan and Li [5], it suffices to show that for any small probability $\varepsilon > 0$, we can always find a constant $C > 0$, such that

$$\Pr\Big[\inf_{N^{-2}\|u\|^2 = C^2} Q_\lambda\{B_0 + (Nht_{3,N}hz_{3,N})^{-1/2} u\} > Q_\lambda(B_0) \Big] \geq 1 - \varepsilon. \tag{12}$$

By definition of $Q_\lambda(B)$, we have

$$\begin{aligned}
&\frac{ht_{3,N}hz_{3,N}}{N^2}[Q_\lambda(B_0+\{Nht_{3,N}hz_{3,N})^{-1/2}u\}-Q_\lambda(B_0)]\\
&=\frac{ht_{3,N}hz_{3,N}}{N^2}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij}-\mu(t_{i_1j_1},Z_{i_2j_2})-(Nht_{3,N}hz_{3,N})^{-1/2}u_{i_1j_1,i_2j_2}\}^2\\
&K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2})\\
&-\frac{ht_{3,N}hz_{3,N}}{N^2}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij}-\mu(t_{i_1j_1},Z_{i_2j_2})\}^2K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2})\\
&+\frac{ht_{3,N}hz_{3,N}\lambda}{N^2}\times\frac{\|vu0+(Nht_{3,N}hz_{3,N})^{-1/2}vu\|-\|vu0\|}{N}=R_1.
\end{aligned}$$

By simple algebraic calculations and the fact that $\|vu0\|/N=0$, we have

$$R_1\geq\frac{1}{N^2}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\{u^2_{i_1j_1,i_2j_2}\widehat{f}(t_{i_1j_1},z_{i_2j_2})-2u_{i_1j_1,i_2j_2}\widehat{e}_{i_1j_1,i_2j_2}\}=R_2,$$

where

$$\widehat{e}_{i_1j_1,i_2j_2}=\sqrt{\frac{ht_{3,N}hz_{3,N}}{N}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\mu_{t_{i_1j_1},Z_{i_2j_2}}-\mu_{t_{ij},Z_{ij}}-\varepsilon_{ij})K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2})$$

and

$$\widehat{f}(t_{i_1j_1},z_{i_2j_2})=\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2}).$$

Let $\widehat{f}_{\min}$ be the smallest element of $\widehat{f}(t_{i_1j_1},z_{i_2j_2})$ and $\widehat{e}\in\mathbb{R}^{N\times N}$ with the (i_1j_1,i_2j_2) element equal to $\widehat{e}_{i_1j_1,i_2j_2}$. We have

$$R_2\geq\widehat{f}_{\min}\{N^{-2}\|u\|^2\}-2(N^{-2}\|u\|^2)^{1/2}(N^{-2}\|\widehat{e}\|^2)^{1/2}=R_3.$$

By the condition $N^{-2}\|u\|^2=C^2$, we have

$$R_3=\widehat{f}_{\min}C^2-2C\times(N^{-2}\|\widehat{e}\|^2)^{1/2}. \tag{13}$$

After some algebraic calculations, we have $N^{-2}\|\widehat{e}\|^2=O_p(1)$. By Assumption 2, we have $\Pr(\widehat{f}_{\min}\rightarrow$

$f_{\min}) \to 1$ and $f_{\min} > 0$, where $f_{\min}$ is the smallest value of joint probability density function. Lastly, note that the first term in (13) is a quadratic function in C while the second term is linear in C. As long as C is sufficiently large, the right-hand side of (13) is guaranteed to be positive with probability arbitrarily close to 1. This proves (12). The proof is of Lemma 2 is thus complete. □

Proof of Theorem 4. Suppose the claim is not true, i.e., $\|\overline{\mathbf{b}}\| \neq 0$. Since

$$\overline{\mathbf{b}} = \frac{1}{N}(\mu(t_{12}, Z_{11}) - \mu(t_{11}, Z_{11}), \ldots, \mu(t_{nm_n}, Z_{11}) - \mu(t_{n,m_n-1}, Z_{11}), \ldots, \mu(t_{nm_n}, Z_{nm_n}) - \mu(t_{n,m_n-1}, Z_{nm_n}))^\top,$$

then $\boldsymbol{\nu} = (\mu(t_{11}, Z_{11}), \ldots, \mu(t_{nm_n}, Z_{11}), \ldots, \mu(t_{nm_n}, Z_{nm_n}))^\top / N$ must be the solution of the following normal equation

$$0 = \frac{\partial Q_\lambda(B)}{\partial \boldsymbol{\nu}} = \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2, \tag{14}$$

where $\boldsymbol{\alpha}_1$ is an N^2-dimensional vector with its kth component given by α_{1k}. If $k = 1 + (n1 - 1)N$ for all $n_1 \in \{1, \ldots, N\}$, then for all $i_2 \in \{1, \ldots, n\}$ and $j_2 \in \{1, \ldots, m_{i_2}\}$,

$$\alpha_{1k} = \frac{\lambda\{\mu(t_{11}, Z_{i_2 j_2}) - \mu(t_{12}, Z_{i_2 j_2})\}}{\|\overline{\mathbf{b}}\|}.$$

If $k = k^* + (n_1 - 1)N$ for all $n_1 \in \{, \ldots, N\}$ and $2 < k^* < N$, then for all $i_2 \in \{1, \ldots, n\}$ and $j_2 \in \{1, \ldots, m_{i_2}\}$,

$$\alpha_{1k} = \frac{\lambda\{\mu(t_k, Z_{i_2 j_2}) - \mu(t_{k-1}, Z_{i_2 j_2}) - \mu(t_{k+1}, Z_{i_2 j_2})\}}{\|\overline{\mathbf{b}}\|}.$$

If $k = N + (n_1 - 1)N$ for all $n_1 \in \{1, \ldots, N\}$, then for all $i_2 \in \{1, \ldots, n\}$ and $j_2 \in \{1, \ldots, m_{i_2}\}$,

$$\alpha_{1k} = \frac{\lambda\{\mu(t_{nm_n}, Z_{i_2 j_2}) - \mu(t_{n,m_n-1}, Z_{i_2 j_2})\}}{\|\overline{\mathbf{b}}\|}.$$

Thus α_2 is a N^2-dimensional vector whose kth component is given by

$$-2 \sum_{i_2=1}^{n} \sum_{j_2=1}^{m_{i2}} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i1}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{Y_{ij} - \widehat{\mu}(t_{i_1 j_1}, Z_{i_2 j_2})\} K_{ht_{3,N}}(t_{ij} - t_{i_1 j_1}) K_{hz_{3,N}}(Z_{ij} - Z_{i_2 j_2}).$$

By the standard argument of kernel smoothing when applying Lemma 1, $\|\alpha_2\| = O_p(Nht_{3,N}^{-1/2} hz_{3,N}^{-1/2})$. Furthermore, under the conditions of the theorem, we know that $\Pr(\|\boldsymbol{\alpha}_1\| > \|\boldsymbol{\alpha}_2\|) \to 1$. Consequently,

we know that, with probability tending to 1, the normal equation (14) cannot hold. This implies that $\|\widehat{\widetilde{\mathbf{b}}}_\lambda\|$ must be located at the place where the objective function $Q_\lambda(B)$ is not differentiable. Since the only place where $Q_\lambda(B)$ is not differentiable for $\widehat{\widetilde{\mathbf{b}}}_\lambda$ is the origin, we know immediately that $\Pr(\|\widehat{\widetilde{\mathbf{b}}}_\lambda\| = 0) \to 1$. This completes the proof of Theorem 4. □

Proof of Theorem 5. Trivial in view of Theorem 4. We omit the details. □

Proof of Theorem 6. In this paper, we say that an arbitrary model $\mathcal{S}$ is underfitted if the true model is a bivariate nonparametric model but we think that it as a univariate nonparametric model; it is overfitted if the true model is a univariate nonparametric model but we estimate it as a bivariate nonparametric model. Then, according to whether the model $\mathcal{S}_\lambda$ is underfitted, correctly fitted, or overfitted, we can create three mutually exclusive sets, viz.

$$\mathcal{R}_+ = \{\lambda \in R : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}, \quad \mathcal{R}_0 = \{\lambda \in R : \mathcal{S}_\lambda = \mathcal{S}_T\}, \quad \mathcal{R}_- = \{\lambda \in R : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}.$$

Furthermore, following the same idea as in [8], we define a reference tuning parameter λ_n such that $\lambda_n = O_p\{N^{-3/2} h t_{3,N} h z_{3,N} \ln(N)/||\widehat{\widetilde{\mathbf{b}}}||\}$, where $\widehat{\widetilde{\mathbf{b}}}$ has the same definition as $\widehat{\widetilde{\mathbf{b}}}_\lambda^m$ except that $\widehat{\mu}_\lambda^m(t, Z)$ is replaced by the unpenalized estimator. It follows immediately that such a tuning parameter satisfies the technical conditions specified in Theorem 4. Consequently, we know that $\Pr(\mathcal{S}_{\lambda_n} = \mathcal{S}_\mathcal{T}) \to 1$. Then the theorem can be proved by comparing BIC_{λ_n} and BIC_λ. We distinguish two cases.

Case 1 (Underfitted model). Recall that $\widehat{B}_\lambda$ automatically determines a model $\mathcal{S}_\lambda$. Under such a model $\mathcal{S}_\lambda$, we can define another unpenalized estimate $\widehat{B}_{\mathcal{S}_\lambda}$ as

$$\widehat{B}_{\mathcal{S}_\lambda} = \sum_{i_2=1}^{n} \sum_{j_2=1}^{m_{i2}} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i1}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{Y_{ij} - \mu(t_{i_1 j_1}, Z_{i_2 j_2})\}^2 K_{h t_{3,N}}(t_{ij} - t_{i_1 j_1}) K_{h z_{3,N}}(Z_{ij} - Z_{i_2 j_2}).$$

In other words, $\widehat{B}_{\mathcal{S}_\lambda}$ is the unpenalized estimator under the model determined by $\widehat{B}_\lambda$. By definition, we have $\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{\mathcal{S}_\lambda}$. Due to the fact that $\boldsymbol{\beta}_\mathcal{S} \neq \boldsymbol{\beta}_0$ for any $\mathcal{S} \supset \mathcal{S}_T$, we also have

$$\begin{aligned}
\mathrm{RSS}_\lambda - \mathrm{RSS}_{\lambda_n} &> N^{-2} \sum_{i_2=1}^{n} \sum_{j_2=1}^{m_{i2}} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i1}} \{\widehat{\mu}_{\mathcal{S}_\lambda}(t_{i_1 j_1}, Z_{i_2 j_2}) - \widehat{\mu}_\lambda(t_{i_1 j_1}, Z_{i_2 j_2})\}^2 \widehat{f}(t_{i_1 j_1}, Z_{i_2 j_2}) \\
&\geq \widehat{f}_{\min}\{N^{-2} \sum_{i_2=1}^{n} \sum_{j_2=1}^{m_{i2}} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i1}} \{\widehat{\mu}_{\mathcal{S}_\lambda}(t_{i_1 j_1}, Z_{i_2 j_2}) - \widehat{\mu}_\lambda(t_{i_1 j_1}, Z_{i_2 j_2})\}^2\} \\
&= \widehat{f}_{\min}\{\|\widehat{B}_{\mathcal{S}_\lambda} - \widehat{B}_\lambda\|\} \to f_{\min}\|B_{\mathcal{S}_\lambda} - B_\lambda\| > 0,
\end{aligned}$$

where $\widehat{f}_{\min}$ and $f_{\min}$ are defined as in Lemma 2. This together with the definition of BIC_λ suggest that $\Pr(\inf_{\lambda\in\mathcal{R}_-}\mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \xrightarrow{P} 1$.

Case 2 (Overfitted model). Let λ be an arbitrary tuning parameter that produces an overfitted model (i.e., $\lambda \in \mathcal{R}_+$). For the unpenalized estimator $\tilde{B}$, we must have

$$\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij}-\tilde{\mu}(t_{i_1j_1},Z_{i_2j_2})\}K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2})=0.$$

Thus

$$\begin{aligned}\mathrm{RSS}_\lambda &= N^{-3}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij}-\tilde{\mu}(t_{i_1j_1},Z_{i_2j_2})\}^2K_{ht_{3,N}}(t_{ij}-t_{i_1j_1})K_{hz_{3,N}}(Z_{ij}-Z_{i_2j_2})\\ &\quad+N^{-2}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\{\tilde{\mu}(t_{i_1j_1},Z_{i_2j_2})-\widehat{\mu}_\lambda(t_{i_1j_1},Z_{i_2j_2})\}^2\widehat{f}(t_{i_1j_1},Z_{i_2j_2})\\ &=\mathrm{RSS}_F+\mathrm{R}_\lambda.\end{aligned}$$

It follows that

$$\begin{aligned}&\ln(\mathrm{RSS}_\lambda)-\ln(\mathrm{RSS}_F)=\ln\left(\frac{\mathrm{RSS}_\lambda}{\mathrm{RSS}_F}\right)\\ &\geq-\frac{\widehat{f}_{\max}}{N^2\mathrm{RSS}_F}\sum_{i_2=1}^{n}\sum_{j_2=1}^{m_{i2}}\sum_{i_1=1}^{n}\sum_{j_1=1}^{m_{i1}}\|\tilde{\mu}(t_{i_1j_1},Z_{i_2j_2})-\widehat{\mu}_\lambda(t_{i_1j_1},Z_{i_2j_2})\|^2=O_p\{(Nht_{3,N}hz_{3,N})^{-1}\}.\end{aligned}\tag{15}$$

Similarly, we can prove that

$$\ln(\mathrm{RSS}_{\lambda_n})-\ln(\mathrm{RSS}_F)=-|O_p\{(Nht_{3,N}hz_{3,N})^{-1}\}|.\tag{16}$$

Combining (15) and (16) we deduce that

$$\inf_{\lambda\in\mathcal{R}_+}\ln\mathrm{RSS}_\lambda-\ln\mathrm{RSS}_{\lambda_n}\geq-|O_p\{(Nht_{3,N}hz_{3,N})^{-1}\}|.$$

Consequently, it follows that

$$\mathrm{BIC}_\lambda-\mathrm{BIC}_{\lambda_n}\geq-|O_p\{(Nht_{3,N}hz_{3,N})^{-1}\}|+\frac{\ln(Nht_{3,N}hz_{3,N})}{Nht_{3,N}hz_{3,N}}.$$

It is clear that, with probability tending to 1, the right-hand side of the above equations is guaranteed to be positive. Consequently, we have $\Pr(\inf_{\lambda\in\mathcal{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \xrightarrow{P} 1$.

Combining Cases 1 and 2, we have

$$\Pr(\inf_{\lambda\in\mathcal{R}_+\cup\mathcal{R}_-} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \xrightarrow{P} 1.$$

The above equation implies that, with probability tending to 1, the tuning parameters failing to identify the true model cannot be selected by our BIC criterion, because it is at least not as favorable as our reference λ_n. Thus the proof of Theorem 6 is complete. □

References

[1] P. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, Analysis of longitudinal data, Oxford University Press, 2013.

[2] B. Efron, T. Hastie, I. Johnstone, R.J. Tibshirani, Least angle regression, Ann. Statist. 32 (2004) 407–499.

[3] J. Fan, I. Gijbels, Local polynomial modelling and its applications, Chapman & Hall/CRC, 1996.

[4] J. Fan, T. Huang, R. Li, Analysis of longitudinal data with semiparametric estimation of covariance function, J. Amer. Statist. Assoc. 102 (2007) 632–641.

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

[6] W.J. Fu, Penalized regressions: the bridge versus the LASSO, J. Comput. Graph. Statist. 7 (1998) 397–416.

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

[8] T. Hu, Y. Xia, Adaptive semi-varying coefficient model selection, Statist. Sinica 22 (2012) 575–599.

[9] C.-R. Jiang, J.-L. Wang, Covariate adjusted functional principal components analysis for longitudinal data, Ann. Statist. 38 (2010) 1194–1226.

[10] C.-R. Jiang, J.-L. Wang, Functional single index models for longitudinal data, Ann. Statist. 39 (2011) 362–388.

[11] C. Leng, W. Zhang, J. Pan, Semiparametric mean-covariance regression analysis for longitudinal data, J. Amer. Statist. Assoc. 105 (2010) 181–193.

[12] Y. Li, Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation, Biometrika 98 (2011) 355–370.

[13] H. Liang, W. Härdle, R.J. Carroll, Estimation in a semiparametric partially linear errorsin-variables model, Ann. Statist. 27 (1999) 1519–1535.

[14] K.Y. Liang, S.L. Zeger, Longitudinal data analysis using generalized linear models, Biometrika 73 (1986) 13–22.

[15] S. Liu, G. Li, Varying-coefficient mean covariance regression analysis for longitudinal data, J. Statist. Plann. Inference 160 (2015) 89–106.

[16] Y. Mack, B.W. Silverman, Weak and strong uniform consistency of kernel regression estimates, Z. Wahrscheinlichkeitstheor. Verwandte Geb. 61 (1982) 405–415.

[17] E. Masry, Multivariate local polynomial regression for time series: Uniform strong consistency and rates, J. Time Series Anal. 17 (1996) 571–599.

[18] M. Pourahmadi, Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, Biometrika 86 (1999) 677–690.

[19] M. Pourahmadi, Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix, Biometrika 87 (2000) 425–435.

[20] A. Qu, B.G. Lindsay, B. Li, Improving generalised estimating equations using quadratic inference functions, Biometrika 87 (2000) 823–836.

[21] C.Y. Tang, C. Leng, Empirical likelihood and quantile regression in longitudinal data analysis, Biometrika 98 (2011) 1001–1006.

[22] R.J. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.

[23] M. Vogt, Testing for structural change in time-varying nonparametric regression models, Econometric Theory 31 (2015) 811–859.

[24] W. Yao, R. Li, New local estimation procedure for a non-parametric regression function for longitudinal data, J. R. Stat. Soc. Ser. B 75 (2013) 123–138.

[25] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B 68 (2006)49–67.

[26] D. Zhang, X. Lin, J. Raz, M. Sowers, Semiparametric stochastic mixed models for longitudinal data, J. Amer. Statist. Assoc. 93 (1998) 710–719.

[27] W. Zhang, C. Leng, A moving average cholesky factor model in covariance modelling for longitudinal data, Biometrika 99 (2012) 141–150.

[28] X. Zhang, B.U. Park, J.-L. Wang, Time-varying additive models for longitudinal data, J. Amer. Statist. Assoc. 108 (2013) 983–998.

[29] J. Zhou, A. Qu, Informative estimation and selection of correlation structure for longitudinal data, J. Amer. Statist. Assoc. 107 (2012) 701–710.