Bernoulli **24**(4A), 2018, 3013–3038
https://doi.org/10.3150/17-BEJ952

Simultaneous nonparametric regression analysis of sparse longitudinal data

HONGYUAN CAO[1], WEIDONG LIU[2] and ZHOU ZHOU[3]

[1]*Department of Statistics, University of Missouri-Columbia, Columbia, MO 65211, USA.*
E-mail: caohong@missouri.edu
[2]*Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, China.*
E-mail: weidongl@stju.edu.cn
[3]*Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada.*
E-mail: zhou@utstat.toronto.edu

Longitudinal data arise frequently in many scientific inquiries. To capture the dynamic relationship between longitudinal covariates and response, varying coefficient models have been proposed with point-wise inference procedures. This paper considers the challenging problem of asymptotically accurate simultaneous inference of varying coefficient models for sparse and irregularly observed longitudinal data via the local linear kernel method. The error and covariate processes are modeled as very general classes of non-Gaussian and non-stationary processes and are allowed to be statistically dependent. Simultaneous confidence bands (SCBs) with asymptotically correct coverage probabilities are constructed to assess the overall pattern and magnitude of the dynamic association between the response and covariates. A simulation based method is proposed to overcome the problem of slow convergence of the asymptotic results. Simulation studies demonstrate that the proposed inference procedure performs well in realistic settings and is favored over the existing point-wise and Bonferroni methods. A longitudinal dataset from the Chicago Health and Aging Project is used to illustrate our methodology.

Keywords: local polynomial estimation; maximum deviation; nonparametric regression; simultaneous confidence band; sparse longitudinal data

1. Introduction

Sparse and irregularly spaced longitudinal data frequently occur in biomedicine, epidemiology, psychiatry, education, and other fields of natural and social sciences. The sparsity refers to the availability of only a few observations per subject and the irregularity means that measurement times vary across subjects. In regression analysis of such data, oftentimes scientists and practitioners are interested in investigating the overall pattern and magnitude of the association between the response and predictors across the whole period of observation time with accurate statistical guidance. For example, given a response process $Y(t)$ of time t and p vector processes of covariates $X(t) = \{X^{(1)}(t), \dots, X^{(p)}(t)\}^T$, consider the following varying coefficient model [13]

$$Y(t) = X(t)^T \beta(t) + \varepsilon(t), \tag{1.1}$$

where $\beta(t) = \{\beta_1(t), \dots, \beta_p(t)\}^T$ is a p vector of smooth functions of t and $\varepsilon(t)$ is a mean zero error process satisfying $E\{X(t)\varepsilon(t)\} = 0$ for all t. In many applications, one is interested in examining the overall shape of $\beta(t)$ or testing whether certain parametric functions are adequate in describing the overall trend of the regression relationship over time. In such cases, it is desirable to perform accurate simultaneous statistical inference of $\beta(t)$ as a function of time. To illustrate the use of such inference, we consider a longitudinal study from the Chicago Health and Aging Project [2]. The dataset contains 2846 persons initially free of Alzheimer's disease but who are at risk of developing it. Their demographics are recorded at baseline, and they are longitudinally followed for clinical evaluation of Alzheimer's disease. We are interested in investigating how global cognition is dynamically associated with covariates as a person ages. In our analyses, we use a composite measure of global cognition constructed with a battery of 18 tests [26] and investigate its time-varying association with three covariates, gender, race (white and African American) and years in education using model (1.1). Four panels in Figure 4 illustrate the fitted $\beta(t)$. It can be seen that the gender effect is slightly fluctuating around 0 with age. But is this fluctuation statistically significant? On the other hand, after observing the fitted race effect, the practitioner may ask whether it can be represented by a downward linear function or whether the downward trend is statistically significantly accelerated as age increases. For an asymptotically accurate and visually friendly answer to such questions, it is desirable to construct simultaneous confidence bands (SCBs) for the regression functions. Specifically, for a pre-specified confidence level $1-\alpha$ and a given p-dimensional smooth function $a(t)$, we aim to find smooth random functions $L(t)$ and $U(t)$, such that

$$P\left\{L(t) \le a(t)^T\beta(t) \le U(t), \forall t \in [a,b]\right\} \to 1-\alpha$$

as number of subjects $n \to \infty$ for a pre-specified closed interval $[a, b]$.

On the other hand, however, the construction of smooth SCBs for nonparametric regression analysis of sparse longitudinal data is known to be a difficult problem. The major difficulty lies in the irregularity of the observation points and the dependence among observations from the same subject. As a result, one has to perform uniform investigation into a dependent empirical process where stochastic variations must be controlled in both time and the response and covariate processes. This is drastically different from and significantly more difficult than the dense longitudinal data case where observation times are abundant for each subject and nonparametric regression estimators are typically stochastic equi-continuous. Recently, some progresses have been made toward this problem. For instance, Ma, Yang and Carroll [20] constructed SCBs for the mean functions via piece-wise constant spline fitting. Gu et al. [12] constructed piece-wise constant SCBs for B-spline nonparametric regression of sparse longitudinal data. As pointed out in [34], the piece-wise constant spline method suffers from consisting of discontinuous step functions. Meanwhile, Zheng, Yang and Härdle [34] established smooth SCBs for the mean functions of sparse longitudinal data. The theory and methodology in [34] seem to be tailored to the mean inference problem and it seems to be difficult to generalize to the regression settings where both the covariates and the errors are stochastic. To our knowledge, the aforementioned problem of smooth SCBs construction for nonparametric regression analysis of sparse longitudinal data remains open at the moment.

In this paper, we shall construct smooth and asymptotically accurate SCBs for the regression functions in model (1.1) via establishing an asymptotic theory for the maximum deviations of the local polynomial estimates of $a(t)^T \beta(t)$. We show that, for a very general class of non-Gaussian and non-stationary error and covariate processes, the appropriately centered and normalized maximum deviation of $a(t)^T \hat{\beta}(t)$ converges to a Gumbel distribution. In particular, our theory allows a mixture of both time-variant and time-invariant covariates. This is flexible and realistic as in practice longitudinal studies often include time-invariant covariates such as gender and race and time-variant covariates such as heart rate and blood pressure. Additionally, we permit $X(t)$ and $\varepsilon(t)$ to be statistically dependent in (1.1) which allows the error process to be heteroscedastic with respect to the covariates. Furthermore, we allow the number of observations for each subject to diverge to infinity with a sufficiently slow rate depending on the smoothing bandwidth (see assumption (A1)). This significantly generalizes most of the previous settings in sparse longitudinal studies where the number of observations per subject is assumed to be bounded or random with certain bounded moments.

Our theoretical investigation depends heavily on a highly non-trivial chaining argument for dependent and double indexed empirical processes which transfers the problem of maximum deviation on a continuous time interval to a corresponding problem on a dense discrete grid. We then utilize a deep Gaussian approximation result established in [32] to further connect the current problem with that of maximum deviations of Gaussian random vectors. The Gaussian approximation results also directly suggest a finite sample simulation based bootstrapping method which improves coverage accuracy in practical implementations. Finally, the above mentioned theoretical results are of general interest and can be used for a wide class of simultaneous inference problems for sparse longitudinal data.

Extensive investigation on the estimation and inference for varying coefficient models based on a number of different smoothing methods and sampling schemes has been carried out in the literature. It is impossible to have a complete reference here and we only list some representative works. For sparse longitudinal data, Hoover et al. [14] suggested kernel-type local polynomial estimator, whose theoretical properties and inference procedures were rigorously studied by Wu, Chiang and Hoover [28]. Chiang, Rice and Wu [5] studied smoothing spline estimation, Huang, Wu and Zhou [15] used basis approximations approach, Fan and Zhang [10] adopted a two-stage estimation strategy, Cao, Zeng and Fine [4] developed a counting process approach on the observation time and Yao, Müller and Wang [31] developed functional analytical approaches. For independent samples, in an influential paper, Bickel and Rosenblatt [1] pioneered the maximum deviations of density function estimates. This was followed by Johnston [16], Eubank and Speckman [8], Xia [30] and Fan and Zhang [11] for cross-sectional data. Wu and Zhao [29], Zhao and Wu [33], Zhou and Wu [35] and Liu and Wu [19] investigated the simultaneous inference problem for time series data. Wang and Yang[25], Degras [6] and Cao, Yang and Todem [3] performed simultaneous inference of the mean function of univariate dense functional data where observations within subject approach infinity sufficiently fast as number of subjects increases. In this paper, we are interested in constructing smooth and asymptotically correct SCBs for $a(t)^T \beta(t)$ based on sparse and irregularly observed longitudinal data.

The rest of the paper is organized as follows. In Section 2, we discuss estimation for model (1.1) using the local linear kernel method and provide the corresponding theoretical findings. In Section 3, we propose a simulation based implementation method to overcome the problem of slow convergence of the theoretical results and an automatic bandwidth selection procedure. Section 4 reports some simulation studies and applies our method to a longitudinal dataset from the Chicago Health and Aging Project. Concluding remarks are given in Section 5. Proofs of results from Section 2 are relegated in the Appendix.

2. Theory and methods

Suppose that a random sample from model (1.1) consists of n subjects. The jth measurement of $\{t, Y(t), X(t)\}$ for the ith subject is $\{t_{ij}, Y_{ij}, X_{ij}\}$, where $1 \le i \le n$, $1 \le j \le m_i$, m_i is the number of measurements for the ith subject, t_{ij} is the measurement time, Y_{ij} is the measurement of the response process $Y_i(t)$ at t_{ij} and X_{ij} is the observation of $X_i(t)$ at t_{ij}. The total number of observations across all subjects is $N = \sum_{i=1}^{n} m_i$. We consider the local linear estimator [9]:

$$\{\hat{\beta}(t), \hat{\beta}'(t)\} = \arg\min_{\beta_0, \beta_1 \in R^p} \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} \{Y_{ij} - X_{ij}^T \beta_0 - X_{ij}^T \beta_1 (t_{ij} - t)\}^2 K_{h_N}(t_{ij} - t) \right],$$

where $K(\cdot)$ is an even kernel function with support $[-A, A]$, $K(\cdot) \ge 0$, $\int_{-A}^{A} K(t)\,dt = 1$ and $K_h(t) = K(t/h)$. The bandwidth $h_N \to 0$ and $nh_N \to \infty$. Define

$$S_{n,l}(t) = (nh_N)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} X_{ij} X_{ij}^T \{(t_{ij} - t)/h_N\}^l K_{h_N}(t_{ij} - t)$$

and

$$R_{n,l}(t) = (nh_N)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} X_{ij} Y_{ij} \{(t_{ij} - t)/h_N\}^l K_{h_N}(t_{ij} - t).$$

Let $\hat{\eta}_{h_N}(t) = \{\hat{\beta}^T(t), h_N(\hat{\beta}'(t))^T\}^T$. Then

$$\hat{\eta}_{h_N}(t) = \begin{pmatrix} S_{n,0}(t) & S_{n,1}^T(t) \\ S_{n,1}(t) & S_{n,2}(t) \end{pmatrix}^{-1} \begin{pmatrix} R_{n,0}(t) \\ R_{n,1}(t) \end{pmatrix} =: S_n^{-1}(t) R_n(t). \tag{2.1}$$

Let $\gamma_i(t)$, $i = 1, 2, \dots, n$ be i.i.d. centered non-stationary stochastic processes. Before we establish the simultaneous asymptotic theory for $\hat{\beta}(t)$, we shall first propose the following important theorem on the maximum deviation of kernel estimates of $\{\gamma_i(t)\}$ sampled at sparse and irregular time points. We need the following assumptions:

(A1) $\max_{1 \le i \le n} m_i \le C \min\{(nh_N)^{\delta/2}, h_N^{-\delta}\}$ for some $0 < \delta < 1$.
(A2) The design time t_{ij}, $1 \le i \le n$, $1 \le j \le m_i$ are i.i.d. random variables with density function $f(t)$ and are independent of $\{\gamma_i(t)\}$, $i = 1, 2, \dots, n$. The density function $f(t) > 0$ for $t \in [l, u]$, where $l < u$ are pre-determined constants.
(A3) There exist $0 < \delta_2 \le \delta_1 < 1$ such that $n^{-\delta_1} = O(h_N)$ and $h_N = O(n^{-\delta_2})$.
(A4) $K(x)$ is differentiable over $(-A, A)$. The right [resp., left] derivative $K'(-A)$ [resp., $K'(A)$] exists, and $\sup_{|x| \le A} K'(x) < \infty$. The Lebesgue measure of the set $\{x \in [-A, A] : K(x) = 0\}$ is zero.

(A5) $\sigma^2(t) := \mathrm{var}\{\gamma_i(t)\}$ and $f(t)$ are positive, bounded Lipschitz continuous functions.
(A6) $\sup_t \mathsf{E}|\gamma_1(t)|^q < \infty$ for some $q > 2/(1-\delta_1)$.

Theorem 1. *Under conditions* (A1)–(A6), *let*

$$M_n(t) = \frac{1}{\sqrt{\lambda_K N h_N \sigma^2(h_N t) f(h_N t)}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \gamma_i(t_{ij}) K\left(\frac{t_{ij}}{h_N} - t\right), \tag{2.2}$$

where $\lambda_K = \int_{-\infty}^{\infty} K^2(t)\,dt$. *We have, as* $n \to \infty$, *for every* $z \in R$,

$$\mathsf{P}\Big[\big(2\log \bar{h}^{-1}\big)^{1/2}\Big\{\sup_{l \le t \le u} \big|M_n(t/h_N)\big| - d_n \le z\Big\}\Big] \to e^{-2e^{-z}}, \tag{2.3}$$

where $\bar{h} = h_N/(u-l)$,

$$d_n = \big(2\log \bar{h}^{-1}\big)^{1/2} + \frac{1}{(2\log \bar{h}^{-1})^{1/2}}\left\{\log \frac{K_1}{\sqrt{\pi}} + \frac{1}{2}\log\log \bar{h}^{-1}\right\}$$

if $K_1 := \{K^2(-A) + K^2(A)\}/(2\lambda_K) > 0$; *otherwise*

$$d_n = \big(2\log \bar{h}^{-1}\big)^{1/2} + \frac{1}{(2\log \bar{h}^{-1})^{1/2}} \log \frac{K_2^{1/2}}{2^{1/2}\pi}$$

with $K_2 = \int_{-A}^{A}\{K'(t)\}^2\,dt/(2\lambda_K)$.

A few comments on the regularity conditions are in order. Condition (A1) posits a certain level of sparsity for each subject. This assumption is significantly weaker than most of the sparsity conditions imposed in the literature of longitudinal data analysis where m_i is required to be non-stochastic and bounded or random with bounded moments. Condition (A2) stipulates that the observation time is random across subjects. This is critical for our proposed method to work. Similar assumptions have been made in longitudinal data analysis literature [34]. Condition (A3) specifies the allowable range of the bandwidths. Condition (A4) states some mild restrictions on the kernel function. Condition (A5) places requirements on the variance of the error term and the measurement time density function, which would typically be satisfied in practice. Condition (A6) ensures that $\gamma_i(t)$ has finite moment greater than two across the time domain, which is fairly mild.

Theorem 1, which is established in the Appendix, is a general and important result leading to the theory of normalized maximum deviation of local polynomial estimators of sparse longitudinal data. Note that $\gamma_i(t)$ can be a wide class of non-Gaussian and non-stationary processes and we do not post any restrictions on the temporal dependence structure of those processes. Furthermore, note that the normalized asymptotic limits in the theorem are the same for all choices of $\gamma_i(t)$. Hence, one can use Theorem 1 to construct critical values of the SCBs based on a Monte Carlo method with such simple choices as $\gamma_i(t_{ij}) \sim N(0,1)$, $1 \le i \le n$, $1 \le j \le m_i$, $\sigma^2(t) = 1$

and $f(t) = 1, \forall t \in [l, u]$. A Monte Carlo based method for implementation is illustrated in detail in Section 3.

Let $a(t)$ be a pre-specified p dimensional vector function. We are interested in constructing a SCB for the function $a(t)^T \beta(t)$. For example, $a(t) = 1_{p,k}$, where $1_{p,k}$ denotes the p dimensional vector with the kth entry 1 and all other entries 0. In this case $a(t)^T \beta(t)$ is the kth component function of $\beta(t)$. We first list a few more regularity conditions:

(A7) Both $\beta(t)$ and $a(t)$ are twice continuously differentiable on $[l, u]$.

(A8) Let $\Sigma_p(t) = \mathsf{E}[\{X_i^{(1)}(t), \dots, X_i^{(p)}(t)\}^T \{X_i^{(1)}(t), \dots, X_i^{(p)}(t)\}]$ and $\Xi_p(t) = \mathsf{Cov}[\{\varepsilon_i(t) X_i^{(1)}(t), \dots, \varepsilon_i(t) X_i^{(p)}(t)\}]$. Assume that both $\Sigma_p(t)$ and $\Xi_p(t)$ are uniformly positive definite and Lipschitz continuous on $[l, u]$.

Define

$$\Delta_n = \sup_{t \in [l,u]} \sqrt{\frac{N h_N f(t)}{a(t)^T \Sigma_p^{-1}(t) \Xi_p(t) \Sigma_p^{-1}(t) a(t) \lambda_K}} \bigg| a(t)^T \hat{\beta}(t) - a(t)^T \beta(t) - \frac{h_N^2}{2} \big[a(t)^T \{\beta(t)\}''\big] \int_{-A}^{A} x^2 K(x)\, dx \bigg|.$$

Theorem 2. *Suppose* (A1)–(A5) *hold with* $\gamma_i(t)$ *therein replaced by* $\varepsilon_i(t)$. *Assume* $\mathsf{E}|\varepsilon_i(t)|^q < \infty$ *for some* $q > 4/(1 - \delta_1)$ *and the design points* $\{t_{ij}\}$ *are independent of* $\{\varepsilon_i(t), X_i(t)\}_{i=1}^n$. *Assume that* (A7) *and* (A8) *hold. Furthermore*, $\sup_t \mathsf{E}\|X_i(t)\|^q < \infty$ *for some* $q > 4/(1 - \delta_1)$ *and*

$$N h_N^7 = o\Big(1/\sqrt{\log h_N^{-1}}\Big). \tag{2.4}$$

Then we have, as $n \to \infty$, *for every* $z \in R$

$$\mathsf{P}\big\{\big(2 \log \bar{h}^{-1}\big)^{1/2} (\Delta_n - d_n) \le z\big\} \to e^{-2e^{-z}}. \tag{2.5}$$

Theorem 2 establishes the asymptotic maximum deviation of $a^T(t)\hat{\beta}(t)$ on $[l, u]$. In particular, if one chooses an undersmoothing bandwidth $h_N \ll N^{-1/5}$ and reduces bias of the kernel estimation to the second order, then the theorem implies that one can construct an asymptotic $100(1 - \alpha)\%$ SCB for $a(t)^T \hat{\beta}(t)$ as

$$a(t)^T \hat{\beta}(t) \pm l_1 \sqrt{\frac{a(t)^T \Sigma_p^{-1}(t) \Xi_p(t) \Sigma_p^{-1}(t) a(t) \lambda_K}{N h_N f(t)}}, \qquad l \le t \le u, \tag{2.6}$$

where $l_1 = \frac{z_\alpha}{(2 \log \bar{h}^{-1})^{1/2}} + d_n$ and $z_\alpha = -\log\log\{(1 - \alpha)^{-1/2}\}$.

In point-wise inference of time-varying coefficient models for sparse longitudinal data, it is well known that the asymptotic behavior of localized nonparametric estimators does not depend on the temporal dependence structure of the covariate and error processes [14]. Theorem 2 extends the aforementioned results and establishes that the latter property holds true for simultaneous nonparametric regression analysis of sparse longitudinal data. This is a nice property in exploratory analyses for longitudinal data since generally it is difficult to accurately estimate the covariance structure of the error and covariate processes.

3. Practical implementation

Due to the slow rate of convergence of the Gumbel distribution, in practice, the SCB in (2.6) may not have good finite-sample performances. To circumvent this problem, we shall adopt a simulation assisted bootstrapping approach. We use a special case under Theorem 1. Specifically, Let T_{ij} be i.i.d. $\mathcal{U}[0, 1]$ random variables and η_{ij} be i.i.d. standard normal distribution, and T_{ij} and η_{ij} are independent, $1 \le i \le n$, $1 \le j \le m_i$. Denote

$$\Pi_n = \sup_{l \le t \le u} \left| \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \eta_{ij} K(\frac{T_{ij}-t}{h_N})}{\sqrt{\lambda_K N h_N}} \right|.$$

By Theorem 1 and Theorem 2, with proper centering and scaling, Π_n and Δ_n have the same asymptotic Gumbel distribution. So the cutoff value $\gamma_{1-\alpha}$, the $(1-\alpha)$th quantile of Δ_n, can be estimated by the sample $(1-\alpha)$th quantile of Π_n based on a large number of replications. Thus, the SCB for $a(t)^T \beta(t)$ can be constructed as

$$a(t)^T \hat{\beta}(t) \pm \gamma_{1-\alpha} \sqrt{\frac{a(t)^T \Sigma_p^{-1}(t) \Xi_p(t) \Sigma_p^{-1}(t) a(t) \lambda_K}{N h_N f(t)}} \tag{3.1}$$

if the bandwidth is selected to satisfy (A3) and $h_N \ll N^{-1/5}$. The rationale behind this approach is that the simulated distribution of Π_n is likely to be closer to Δ_n than the Gumbel distribution in moderate samples.

To implement (3.1) in practice, we need to estimate $f(t)$, $\Sigma_p(t)$ and $\Xi_p(t)$. The estimate of $f(t)$ can be achieved through a kernel density estimation as

$$\hat{f}(t) = \frac{1}{N b_N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} K\left(\frac{t_{ij}-t}{b_N}\right), \tag{3.2}$$

The following proposition establishes the uniform consistency of $\hat{f}(t)$.

Proposition 1. *Under conditions* (A1)–(A5) *with h_N therein replaced by b_N, we have*

$$\sup_{l \le t \le u} \left| \frac{1}{N b_N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} K\left(\frac{t_{ij}-t}{b_N}\right) - f(t) \right| = O_{\mathrm{P}}\left(\sqrt{\frac{\log 1/b_N}{N b_N}} + b_N \right) \tag{3.3}$$

and

$$\sup_{l \le t \le u} \left| \frac{1}{N b_N} \sum_{i=1}^{n} \sum_{j=1}^{m_i} K_1\left(\frac{t_{ij}-t}{b_N}\right) \right| = O_{\mathrm{P}}\left(\sqrt{\frac{\log 1/b_N}{N b_N}} + b_N \right), \tag{3.4}$$

where $K_1(x) = x K(x)$.

For the covariance functions $\Sigma_p(t)$ and $\Xi_p(t)$, they can be estimated via the local kernel method as

$$\hat{\Sigma}_p(t) = \frac{1}{Nc_N}\sum_{i=1}^{n}\sum_{j=1}^{m_i} X_{ij}X_{ij}^T K\left(\frac{t_{ij}-t}{c_N}\right) \tag{3.5}$$

and

$$\hat{\Xi}_p(t) = \frac{1}{Nd_N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\hat{\varepsilon}_i(t_{ij})X_{ij}\right]\left[\hat{\varepsilon}_i(t_{ij})X_{ij}\right]^T K\left(\frac{t_{ij}-t}{d_N}\right), \tag{3.6}$$

where $\hat{\varepsilon}_i(t_{ij}) = Y_{ij} - X_{ij}^T\hat{\beta}(t_{ij})$ are the residuals of the regression. Following similar arguments as those in the proofs of Theorems 1, 2 and Proposition 1, it can be shown that both $\hat{\Sigma}_p(t)$ and $\hat{\Xi}_p(t)$ are uniformly consistent on $[l, u]$. As kernel estimation is local and we are dealing with sparse longitudinal data, the dependence caused by data coming from the same subject in a local neighborhood is asymptotically negligible. Thus the bandwidths b_N, c_N and d_N used in $\hat{f}(t)$, $\hat{\Sigma}_p(t)$ and $\hat{\Xi}_p(t)$ can be chosen based on classic bandwidth selectors of kernel density and kernel nonparametric estimations for independent data [22].

We now discuss the choice of the bandwidth h_N in (2.1). Theorem 1 specifies the theoretical range of allowable bandwidths. However, an automatic bandwidth selection procedure is of practical interest and is usually needed to provide a preliminary idea of a suitable bandwidth that is suggested by data. We adopt the leave-one-subject-out cross-validation procedure for bandwidth selection suggested by Rice and Silverman [21]:

$$\mathrm{CV}(h_N) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{Y_{ij} - \hat{Y}^{(-i)}(t_{ij})\right\}^2, \tag{3.7}$$

where $\hat{Y}^{(-i)}(t_{ij})$ is the local linear estimator of $Y(t_{ij})$ computed with all measurements of the ith subject deleted. A cross-validation bandwidth h_{CV} is then obtained by minimizing $\mathrm{CV}(h_N)$ with respect to h_N; that is, $h_{\mathrm{CV}} = \inf_{h_N\in\mathcal{H}}\mathrm{CV}(h_N)$, where $\mathcal{H}$ is the allowable range of h_N specified by (A3) and (2.4). This cross-validation idea can be easily extended to other smoothing estimators, such as smoothing splines. As a heuristic justification for the above cross-validation method, we adopt the arguments in [15]. Define the average squared error for the local linear estimators

$$\mathrm{ASE}(h) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[X_{ij}^T\left(\hat{\beta}(t_{ij}) - \beta(t_{ij})\right)\right]^2.$$

Then h_{CV} should minimize $\mathrm{ASE}(h)$ asymptotically. Indeed, note that

$$\begin{aligned}\mathrm{CV}(h) = {}& \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\varepsilon_i^2(t_{ij}) - \frac{2}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\varepsilon_i(t_{ij})X_{ij}^T\right]\left(\hat{\beta}^{-i}(t_{ij}) - \beta(t_{ij})\right)\\ &+ \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[X_{ij}^T\left(\hat{\beta}^{-i}(t_{ij}) - \beta(t_{ij})\right)\right]^2.\end{aligned} \tag{3.8}$$

Since $\hat{\beta}^{-i}(t_{ij}) - \beta(t_{ij})$ is independent of $\varepsilon_i(t_{ij})X_{ij}^T$ and the latter has 0 mean, it can be shown that the second term in (3.8) is stochastically dominated by the third. Note that the first term in (3.8) is independent of h and the third term is approximately equal to ASE(h) for large samples. Hence, intuitively the bandwidth h which minimizes CV(h) should minimize ASE(h) asymptotically.

4. Simulation studies

We investigate the finite sample performance of our proposed methodology through Monte Carlo simulations. We first consider model (1.1) with one covariate:

$$Y(t) = \beta_0(t) + \beta_1(t)X(t) + \varepsilon(t), \tag{4.1}$$

where $\beta_0(t) = 5(t-0.6)^2$, $\beta_1(t) = 0.5 + 0.4\sin\{2\pi(t-0.5)\}, 0.5\cos\{2\pi(t-0.3)\}, 4(t-0.4)^2$ or $4(t-0.5)^3$ and we are interested in constructing SCBs for $\beta_1(t)$. We generate 1000 datasets, each consisting of $n = 200$ or 400 subjects. Each subject has 4 observations and the observation times are generated from the uniform distribution $\mathcal{U}(0, 1)$. For the ith subject, the covariate process $X_i(t)$ is generated from a mean zero Gaussian process with variance $\exp(t)$ and correlation $\exp(-|t_{ij} - t_{ik}|/4)$ for observations taking place at t_{ij} and t_{ik}. The error process $\varepsilon(t) = \zeta X(t)$, where ζ is a standard normal random variable, independent of $X(t)$. This induces the cross dependence between the covariate process and the error process. The results for other choices of the model parameters are very similar and thus omitted.

We use 191 equally spaced grid points in $[0.025, 0.975]$ for the calculation of coverage probabilities. We use the simulation-assisted approach described in Section 3 to find critical values $\gamma_{1-\alpha}$ at significance levels $\alpha = 0.1$ and 0.05. These are the 90 and 95 percentile of $\sup_{0.025\le t\le 0.975}\left|\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\eta_{ij}K(\frac{T_{ij}-t}{h_N})}{\sqrt{\lambda_k h_N \sum_{j=1}^{n} m_j}}\right|$, where $m_i = 4, T_{ij}$ are i.i.d. with density $f(t) = 1$ and η_{ij} are i.i.d. standard normal random variables, $1 \le i \le n, 1 \le j \le m_i$. For each simulated data, local linear kernel estimator $\hat{\beta}_1(t)$ based on (2.1) is computed using the Epanechnikov kernel, which is $K(x) = 0.75(1-x^2)_+$. As a consequence, $\lambda_K = 0.6$. Results based on other commonly used kernels, such as the Gaussian kernel and the uniform kernel, are similar and thus omitted. To select a data-adaptive bandwidth, we follow the leave-one-subject-out cross validation approach described in Section 3 in the range $(0.06, 0.09)$. Specifically, one subject is reserved as a test subject and the other subjects are used to calculate $\hat{\beta}_1(t)$ at time points that the test subject is observed. Do this for all subjects and the bandwidth which minimizes the mean squared error is the one used in the inference stage and the estimation of $\Sigma_p(t)$ and $\Xi_p(t)$ in (3.1).

We summarize the average ASE based on 1000 replications in Table 1. Furthermore, we plot the ASE based on one realization for the four different forms of $\beta_1(t)$ in Figure 1. The bandwidth that has the smallest ASE is selected in the construction of SCB.

Figure 2 shows a typical plot of $\beta_1(t), \hat{\beta}_1(t)$ and its 95% SCB for $n = 200, 400$ and $\beta_1(t) = 0.5\cos\{2\pi(t-0.3)\}, 4(t-0.4)^2$. Table 2 summarizes the uniform coverage probabilities over 1000 simulations. We observe that as the sample size increases, the coverage probabilities based on our approach are close to the nominal ones, point-wise confidence interval is not valid for

Table 1. Average ASE based on 1000 simulations

n	sin	cos	quad	cubic
200	1.049	1.052	1.045	1.035
400	1.025	1.026	1.019	1.019

Note: "sin" represents $\beta_1(t) = 0.5 + 0.4\sin\{2\pi(t-0.5)\}$, "cos" represents $\beta_1(t) = 0.5\cos\{2\pi(t-0.3)\}$, "quad" represents $\beta_1 = 4(t-0.4)^2$ and "cubic" represents $\beta_1(t) = 4(t-0.5)^3$.

simultaneous inference and the Bonferroni method is overly conservative with larger coverage probabilities.

We next study (1.1) with two covariates. The model is

$$Y(t) = \beta_0(t) + \beta_1(t)X^{(1)}(t) + \beta_2(t)X^{(2)}(t) + \varepsilon(t).$$

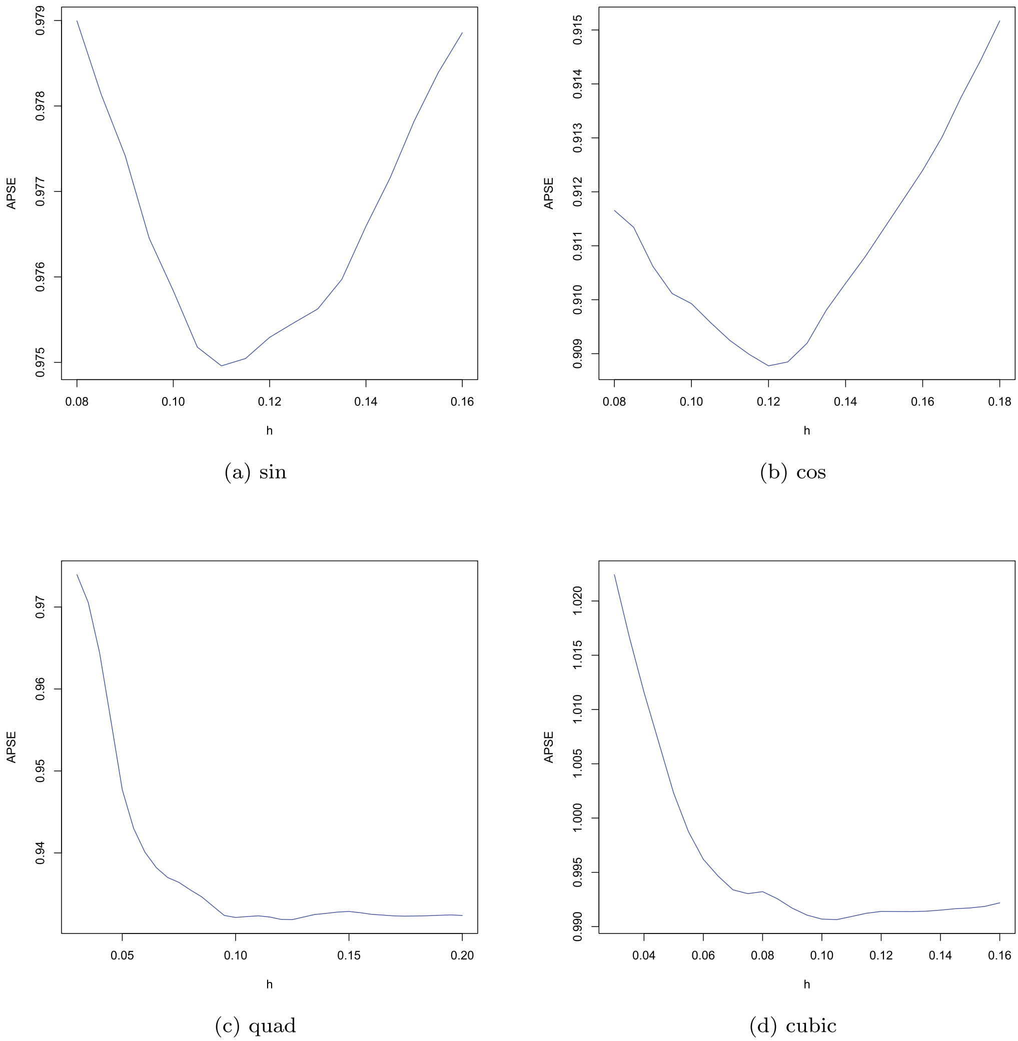
We set $\beta_0(t) = \sqrt{t}$, $\beta_1(t) = 0.4(t-0.6)^2$ and $\beta_2(t) = 0.5\cos\{2\pi(t-0.5)\}$. For the ith subject, the covariate process $X_i^{(1)}(t)$ (resp. $X_i^{(2)}(t)$) follows a mean zero Gaussian process with variance $\exp(t)$ (resp. 2^t) and correlation $\exp(-|t_{ij} - t_{ik}|/4)$ (resp. $2^{-|t_{ij}-t_{ik}|/4}$) for observations taking place at t_{ij} and t_{ik}. Furthermore, $X^{(1)}(t)$ is independent of $X^{(2)}(t)$. The error process $\varepsilon(t) = \xi X^{(1)}(t)$, where ξ is a standard normal random variable, independent of $X^{(1)}(t)$ and $X^{(2)}(t)$. The rest of the simulation set up is the same as in one covariate case.

We are interested in constructing SCBs for $\beta_1(t)$, $\beta_2(t)$, $\beta_1(t)+\beta_2(t)$ and $\beta_1(t)-\beta_2(t)$ which corresponds to $a_i(t)^T\beta(t)$, $i = 1, 2, 3, 4$ with $a_1(t) = (0, 1, 0)$, $a_2(t) = (0, 0, 1)$, $a_3(t) = (0, 1, 1)$ and $a_4(t) = (0, 1, -1)$. The same simulation-assisted critical value method and leave-one-

Table 2. Results of 1000 simulations with one covariate

		90%			95%		
n	Function	cp (ours)	cp (pointwise)	cp (bonferroni)	cp (ours)	cp (pointwise)	cp (bonferroni)
200	1	89.5	7.7	99.6	95.6	28.0	99.8
400	1	93.3	8.4	99.9	98.1	28.3	100
200	2	89.1	7.2	99.8	96.2	27.5	99.8
400	2	90.5	9.0	99.8	96.7	28.1	100
200	3	85.8	10.7	99.4	93.9	31.5	99.7
400	3	89.0	13.3	99.7	95.3	34.9	99.7
200	4	87.9	11.0	99.0	92.9	32.3	99.4
400	4	90.0	11.1	99.3	96.2	33.0	99.9

Note: "function" represents the functional format of $\beta_1(t)$, where 1 represents $\beta_1(t) = 0.5 + 0.4\sin\{2\pi(t-0.5)\}$, 2 represents $\beta_1(t) = 0.5\cos\{2\pi(t-0.3)\}$, 3 represents $\beta_1(t) = 4(t-0.4)^2$ and 4 represents $\beta_1(t) = 4(t-0.5)^3$; "cp (ours)" represents the coverage probability based on our approach; "cp (pointwse)" represents the coverage probability based on point wise inference and "cp (bonferroni)" represents the coverage probability based on Bonferroni method. All numbers are presented in percentage forms.



(a) sin (b) cos

(c) quad (d) cubic

Figure 1. ASE as a function of bandwidth. Different functional forms of $\beta_1(t)$ in (a), (b), (c) and (d).

subject-out cross validation approach are used as in one covariate case. Specifically, the $1-\alpha$ SCB of $\beta_1(t)+\beta_2(t)$ is of the form

$$\hat{\beta}_1(t)+\hat{\beta}_2(t)\pm\gamma_{1-\alpha}\sqrt{a_3(t)^T\hat{\Sigma}_3^{-1}(t)\hat{\Xi}_3(t)\hat{\Sigma}_3^{-1}(t)a_3(t)\lambda_K/\left\{Nh_N\hat{f}(t)\right\}}. \tag{4.2}$$

The $1-\alpha$ SCBs of $\beta_1(t)$, $\beta_2(t)$ and $\beta_1(t)-\beta_2(t)$ can be obtained similarly. We also calculate the coverage probabilities of the point-wise and Bonferroni confidence intervals.

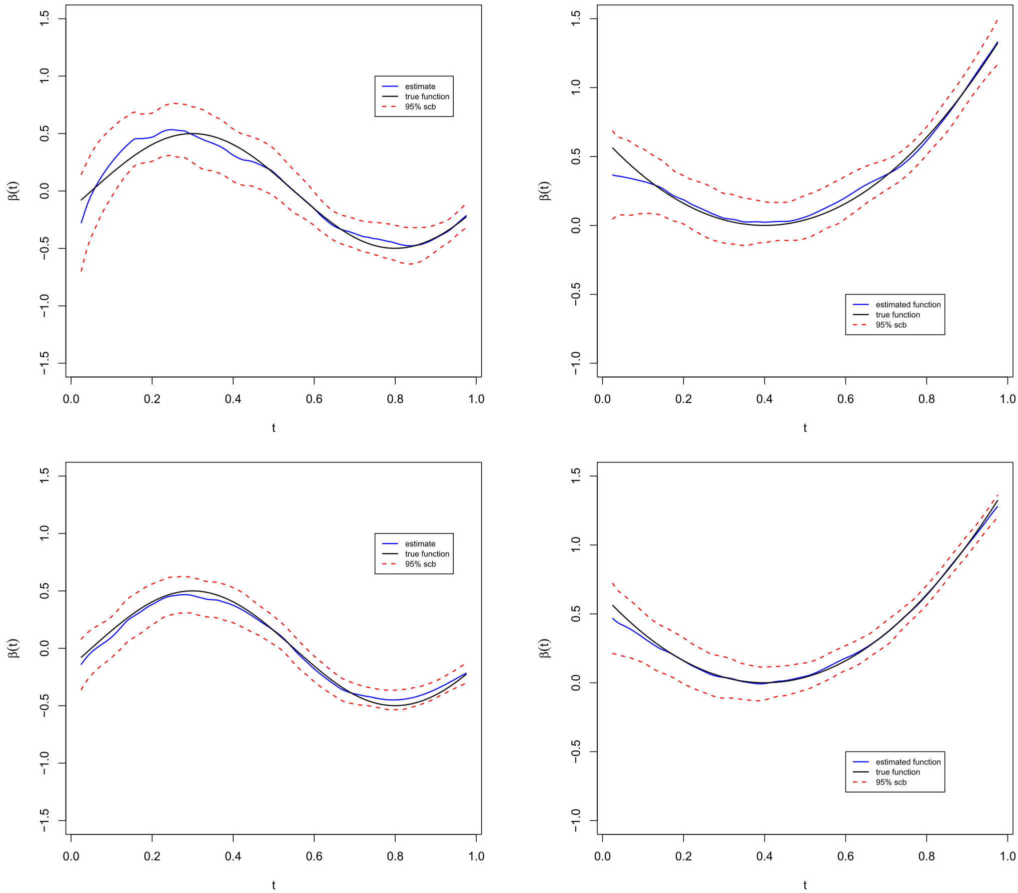


Figure 2. Typical plots of $\beta_1(t)$, $\hat{\beta}_1(t)$ and its 95% SCB. Top panel: $n = 200$, $\beta_1(t) = 0.5\cos\{2\pi(t - 0.3)\}$ and $4(t - 0.4)^2$; bottom panel: $n = 400$, $\beta_1(t) = 0.5\cos\{2\pi(t - 0.3)\}$ and $4(t - 0.4)^2$.

The results in Table 3 demonstrate that our method continues to provide coverage probabilities close to the nominal ones in the two predictor case. On the other hand, point-wise confidence intervals cannot provide correct simultaneous coverage percentages and the Bonferroni approach is too conservative, with too wide confidence bands.

5. Real example

We shall analyze a longitudinal dataset from the Chicago Health and Aging Project [2]. This is a longitudinal population study of common chronic health problems of older persons, especially of risk factors for Alzheimer's disease, in a biracial neighborhood of the south side of Chicago from 1993 through 2006. The dataset contains 2846 persons initially free of Alzheimer's disease

Table 3. Results of 1000 simulations with two covariates

		90%			95%		
n	Function	cp (ours)	cp (pointwise)	cp (bonferroni)	cp (ours)	cp (pointwise)	cp (bonferroni)
200	1	87.3	23.7	99.2	94.0	25.4	99.2
400	1	88.4	29.6	99.7	95.4	31.4	99.7
200	2	88.1	6.7	98.8	93.1	6.7	98.8
400	2	90.1	10.2	99.2	93.1	10.2	99.2
200	3	86.2	5.4	99.1	93.2	17.5	99.3
400	3	88.6	7.6	99.9	93.8	23.0	99.9
200	4	88.3	6.0	99.8	93.7	18.7	99.9
400	4	88.6	8.5	99.3	94.0	24.1	99.6

Note: "function" represents the functional format of $a_i(t)^T\beta(t)$, $i = 1, 2, 3, 4$, where 1 represents $\beta_1(t) = 0.4(t - 0.6)^2$, 2 represents $\beta_2(t) = 0.5\cos\{2\pi(t - 0.5)\}$, 3 represents $\beta_1(t) + \beta_2(t) = 0.4(t - 0.6)^2 + 0.5\cos\{2\pi(t - 0.5)\}$ and 4 represents $\beta_1(t) - \beta_2(t) = 0.4(t - 0.6)^2 - 0.5\cos\{2\pi(t - 0.5)\}$; the rest have the same meaning as in Table 1.

but who are at risk of developing it. Their demographics are recorded at baseline and they are longitudinally followed for clinical evaluation of Alzheimer's disease. Under missing at random assumption [18], 2821 persons were used for analysis. Their ages range from 60 to 100 and are rescaled to the interval [0, 1]. The left panel in Figure 3 shows the histogram of their ages. It can be seen that they are dense in the interval [0, 1] with relative few observations at the beginning and end of the time interval.

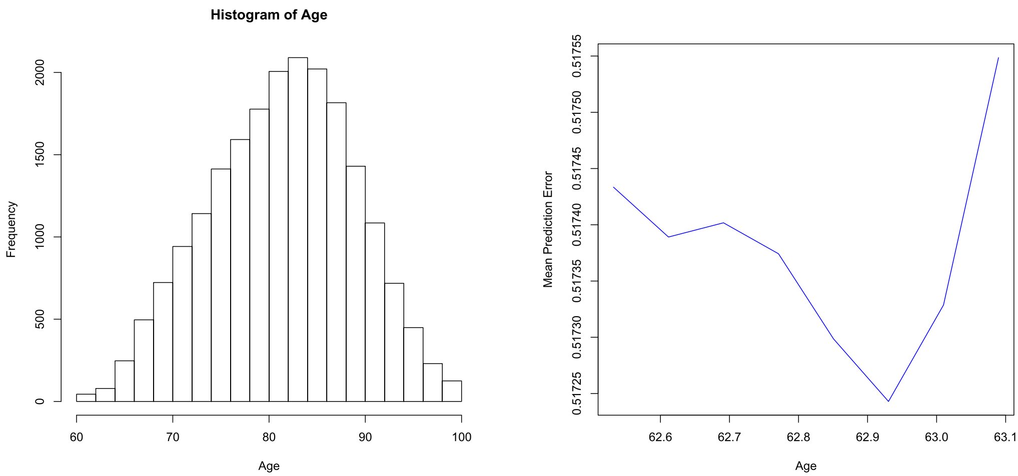


Figure 3. Age distribution and ASE as a function of age in Chicago Health and Aging Project.

We are interested in a composite measure of global cognition (Cognition) constructed with a battery of 18 tests [26]. We use gender, race and education years as covariates and study their time-varying associations with longitudinally measured global cognition. The covariate education years is positive and skewed. We first do a log transformation to this covariate per customary.

Consider the following model

$$\text{Cognition}(t) = \beta_0(t) + \beta_1(t)\text{Gender} + \beta_2(t)\text{Race} + \beta_3(t)\log(\text{Education}) + e(t). \tag{5.1}$$

The cross validation procedure selects age bandwidth 62.93 with corresponding ASE 0.517 as shown in the right panel of Figure 3. Figure 4 shows the 95% SCBs for the regression coefficient functions. We also fitted a constant and a linear function to the regression coefficients to check their adequacy in explaining the dynamic associations.

As we can see from Figure 4, all the SCBs tend to become wider at the beginning and end of the time interval. This is due to the fact that there are relatively fewer observations at those periods of time. The SCB for gender fully contains the horizontal line $y = 0$. This implies that

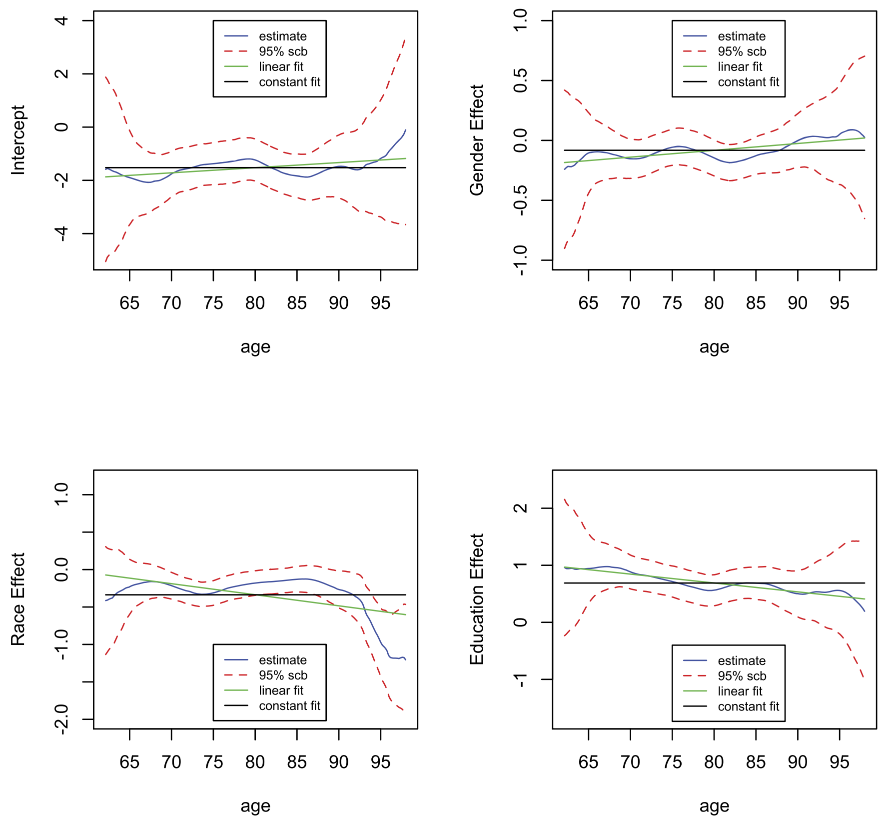


Figure 4. SCB with local linear kernel estimate, fitted constant and linear trends for model (5.1).

there is no evidence against the claim that $\beta_1(t) = 0$ in model (5.1). In other words, the effect of gender is not statistically significant in this study. Similarly, the effect of education and intercept term should be modeled as constants. From Figure 4, we observe that the SCB for race does not contain a straight line or a constant line. This implies that there does exist a statistically significant nonlinearly time-varying effect of race on global cognition. In summary, our analysis with the methodology proposed in this paper suggests that the data can be fitted with the following model

$$\text{Cognition}(t) = a_0 + a_1(t)\text{Race} + a_2 \log(\text{Education}) + e(t). \tag{5.2}$$

Such results are consistent with those in the literature. It has been shown that education attainment is a significant predictor of global cognition [27]. Studies have shown racial disparities among older adults in cognitive decline [23].

6. Concluding remarks

We have developed smooth SCBs for regression coefficient functions in varying coefficient models with sparse and irregularly spaced longitudinal data. We use the local linear estimator and the methodology proposed in this paper can be easily adapted for other nonparametric estimators such as the local polynomial estimators. In constructing the local linear estimator, each subject could have different weight proportional to m_i, the number of observations for this subject. We expect similar results to hold in this case.

Sparsity in this paper means that the observations for each subject are sparse in time. Note that at the same time our asymptotic results allow m_i to diverge to infinity at a sufficiently slow rate (see condition (A1)). On the other hand, however, it is well known that if m_i diverges to infinity sufficiently fast, then we are in the dense longitudinal data domain and the local linear kernel estimates $\hat{\beta}(t)$ will be tight and hence their asymptotic behavior will be totally different from that established in this paper. Therefore it remains an interesting question to establish the divergence rate at which the asymptotic distribution changes from one to another. Extensions to other nonparametric regression models, such as single index models and additive models are possible and of great interest for future research.

In this paper, we require that the observation times t_{ij} are independent of the covariates and errors, which is a frequently made assumption in longitudinal data analysis [7]. However, it is well known that if such an assumption is violated, then the local linear estimators may be biased. In this situation, it is necessary to model the joint distribution between the observation times and the covariates and errors. There have been some discussions on informative observation times in the literature; see, for instance, [24] for a conditional approach and [17] for a joint approach. We shall leave the problem of simultaneous inference for sparse longitudinal data with informative observation times to a future research.

Appendix: Proofs of theorems

For briefness, we denote h_N by h throughout the proofs. Before stating the proof for the main results, we first prove a lemma on the covariance structure of $\{M_n(t)\}$. Let

$$r(s) = 1 - \frac{\int_R (K(x) - K(x+s))^2\,dx}{2\lambda_K}.$$

By Theorems B1 and B2 in [1], we have

$$r(s) = 1 - C_0|s|^\alpha + o\big(|s|^\alpha\big) \qquad \text{as } s \to 0,$$

where $(\alpha, C_0) = (1, K_1)$ if $K_1 > 0$ and $(\alpha, C_0) = (2, K_2)$ if $K_1 = 0$.

Lemma 1. *Under* (A1)–(A6), *we have*

$$\mathsf{E}M_n(t)M_n(s) = r(t-s) + O\Big(h \max_{1\le i\le n} m_i\Big)$$

uniformly in $s, t \in R$.

Proof. Define

$$G_n^*(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \gamma_i(t_{ij}) K\left(\frac{t_{ij}}{h} - t\right).$$

Note that

$$\mathsf{E}G_n^*(t)G_n^*(s) = \sum_{i=1}^{n} \mathsf{E}\left[\sum_{j=1}^{m_i} \gamma_i(t_{ij}) K\left(\frac{t_{ij}}{h} - t\right) \sum_{j=1}^{m_i} \gamma_i(t_{ij}) K\left(\frac{t_{ij}}{h} - s\right)\right].$$

For $j \ne l$, we have

$$\begin{aligned}
&\mathsf{E}\left[\gamma_i(t_{ij})\gamma_i(t_{il}) K\left(\frac{t_{ij}}{h} - t\right) K\left(\frac{t_{il}}{h} - s\right)\right] \\
&\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathsf{E}\big[\gamma_i(u)\gamma_i(v)\big] K\left(\frac{u}{h} - t\right) K\left(\frac{v}{h} - s\right) f(u) f(v)\,du\,dv \\
&\quad \le \sup_t \mathsf{E}\big[\gamma_1^2(t)\big] \left\{\int_{-\infty}^{\infty} K\left(\frac{u}{h} - t\right) f(u)\,du\right\} \left\{\int_{-\infty}^{\infty} K\left(\frac{u}{h} - s\right) f(u)\,du\right\} \\
&\quad = O\big(h^2\big),
\end{aligned} \tag{A.1}$$

where $O(1)$ does not depend on s and t. For $j = l$, we have

$$\begin{aligned}
&\mathsf{E}\left[\gamma_i^2(t_{ij})K\left(\frac{t_{ij}}{h} - t\right)K\left(\frac{t_{ij}}{h} - s\right)\right] \\
&\quad = \int_{-\infty}^{\infty} \sigma^2(v)K\left(\frac{v}{h} - t\right)K\left(\frac{v}{h} - s\right)f(v)\,dv \\
&\quad = h\int_{-\infty}^{\infty} K(s - t + v)K(v)\sigma^2(hv + hs)f(hv + hs)\,dv \\
&\quad = h\sqrt{\sigma^2(ht)\sigma^2(hs)f(ht)f(hs)}\int_{-\infty}^{\infty} K(s - t + v)K(v)\,dv + O\big(h^2\big),
\end{aligned}$$

where the last inequality follows from

$$\left|\sigma^2(hv + hs) - \sqrt{\sigma^2(ht)\sigma^2(hs)}\right| = O(h)$$

and

$$\left|f(hv + hs) - \sqrt{f(ht)f(hs)}\right| = O(h)$$

uniformly for $|s - t| \le 2A$ and $|v| \le 2A$. The above two inequalities can be derived from the Lipschitz continuity of $\sigma^2(t)$ and $f(t)$ and the fact that they are positive and bounded away from zero. The proof is complete. $\square$

A.1. Proof of Theorem 1

Without loss of generality, we assume that $A = 1$ in $K(\cdot)$, $u = 1$ and $l = 0$. Let

$$\tilde{\gamma}_i(t_{ij}) = \gamma_i(t_{ij})I\left\{\left|\gamma_i(t_{ij})\right| \le \Big(\max_{1\le i\le n} m_i\Big)^{-1}\sqrt{nh}/(\log n)^8\right\}, \qquad \hat{\gamma}_i(t_{ij}) = \tilde{\gamma}_i(t_{ij}) - \mathsf{E}\tilde{\gamma}_i(t_{ij}),$$

$$\hat{G}_n(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\hat{\gamma}_i(t_{ij})K\left(\frac{t_{ij}}{h} - t\right), \qquad \breve{G}_n(t) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\big[\gamma(t_{ij}) - \hat{\gamma}_i(t_{ij})\big]K\left(\frac{t_{ij}}{h} - t\right),$$

$$\hat{M}_n(t) = \frac{\hat{G}_n(t)}{\sqrt{\lambda_K Nh\sigma^2(ht)f(ht)}}, \qquad \breve{M}_n(t) = \frac{\breve{G}_n(t)}{\sqrt{\lambda_K Nh\sigma^2(ht)f(ht)}}.$$

Lemma 2. *Under the conditions of Theorem 1, we have*

$$\sup_{0\le t\le 1}\left|\breve{M}_n(t/h)\right| = o_{\mathsf{P}}\big(1/\sqrt{\log h^{-1}}\big).$$

Proof. By (A4) and (A5), we have

$$\mathsf{P}\Big(\max_{1\le i\le n}\max_{1\le j\le m_i} m_i\left|\gamma(t_{ij})\right| \ge \sqrt{nh}/(\log n)^2\Big) \le Cn\Big(\max_{1\le i\le n} m_i\Big)^q (nh)^{-q/2}(\log n)^{2q} = o(1).$$

Also,

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}|\mathsf{E}\tilde{\gamma}_i(t_{ij})|}{\sqrt{Nh}} \le Cn\Big(\max_{1\le i\le n} m_i\Big)^{q-1}(nh)^{-q/2}(\log n)^{2q-2} = o(1).$$

This yields that, for any $\delta > 0$,

$$\mathsf{P}\Big(\sup_{0\le t\le 1}\big|\breve{M}_n(t/h)\big| \ge \delta/\sqrt{\log h^{-1}}\Big) = o(1).$$

This completes the proof of Lemma 2. □

By Lemma 2, it suffices to prove Theorem 1 holds by replacing $M_n(t/h)$ with $\hat{M}_n(t/h)$. To this end, we split the interval $[0, h^{-1}]$ into big and small intervals $W_1, V_1, \dots, W_N, V_N$, where $W_i = [a_i, a_i + w]$, $V_i = [a_i + w, a_{i+1}]$, $a_i = (i-1)(w+v)$. We will let w be fixed and v go to zero. Define $M^+ = \max_{1\le i\le N}\sup_{t\in W_i}\hat{M}_n(t)$, $M^{-1} = \min_{1\le i\le N}\inf_{t\in W_i}\hat{M}_n(t)$. Let

$$R_1 = \mathsf{P}\Big(\max_{1\le k\le N}\sup_{t\in V_k}\hat{M}_n(t) \ge x\Big), \qquad R_2 = \mathsf{P}\Big(\min_{1\le k\le N}\inf_{t\in V_k}\hat{M}_n(t) \le -x\Big).$$

Then we have

$$\Big|\mathsf{P}\Big(\max_{0\le t\le h^{-1}}\big|\hat{M}_n(t)\big| \ge x\Big) - \mathsf{P}\big(\{M^+ \ge x\}\cup\{M^- \le -x\}\big)\Big| \le R_1 + R_2.$$

By Lemma 3 below, we have $\lim_{v\to 0}\limsup_{n\to\infty}[R_1 + R_2] = 0$. It suffices to deal with the probability $\mathsf{P}(\{M^+ \ge x\}\cup\{M^- \le -x\})$. Let

$$\Lambda_k^+ = \max_{1\le j\le \chi}\hat{M}_n\big(a_k + jax^{-2/\alpha}\big), \qquad \Lambda_k^- = \min_{1\le j\le \chi}\hat{M}_n\big(a_k + jax^{-2/\alpha}\big),$$

where $\chi = [wx^{2/\alpha}/a]$, $a > 0$. By some elementary calculations,

$$\begin{aligned}
&\Big|\mathsf{P}\big(\{M^+ \ge x\}\cup\{M^- \le -x\}\big) - \mathsf{P}\Big(\Big\{\max_{1\le k\le N}\Lambda_k^+ \ge x\Big\}\cup\Big\{\min_{1\le k\le N}\Lambda_k^- \le -x\Big\}\Big)\Big| \\
&\quad \le \sum_{k=1}^{N}\Big|\mathsf{P}\Big(\sup_{t\in W_i}\hat{M}_n(t) \ge x\Big) - \mathsf{P}\big(\Lambda_k^+ \ge x\big)\Big| + \sum_{k=1}^{N}\Big|\mathsf{P}\Big(\inf_{t\in W_i}\hat{M}_n(t) \le -x\Big) - \mathsf{P}\big(\Lambda_k^- \le -x\big)\Big| \\
&\quad =: R_3 + R_4.
\end{aligned}$$

We now show that $\lim_{a\to 0}\limsup_{n\to\infty}[R_3 + R_4] = 0$. Define $\phi(x) = e^{-x^2/2}/(x\sqrt{2\pi})$ and $x = d_n + z/(2\log h^{-1})^{1/2}$. We also define $H_\alpha(a)$ and H_α as the Pickands constants (see Theorem A1 and Lemmas A1 and A3 in [1]). By these results, we see that $H_1 = 1$, $H_2 = 1/\sqrt{\pi}$ and $\lim_{a\to 0} H_\alpha(a)/a = H_\alpha$.

Lemma 3. *Suppose the conditions in Theorem 1 hold. Let $t > 0$ such that $\inf\{s^{-\alpha}(1-r(s)) : 0 \le s \le t\} > 0$. We have*

$$\mathsf{P}\left(\bigcup_{j=1}^{[tx^{2/\alpha}/a]} \{\hat{M}_n(v + jax^{-2/\alpha}) \ge x\}\right) = x^{2/\alpha}\phi(x)\frac{H_\alpha(a)}{a}C_0^{1/\alpha}t + o(x^{2/\alpha}\phi(x))$$

uniformly over $0 \le v \le h^{-1}$. Also,

$$\mathsf{P}\left(\bigcup_{0\le s\le t}\{\hat{M}_n(v+s) \ge x\}\right) = x^{2/\alpha}\phi(x)H_\alpha C_0^{1/\alpha}t + o(x^{2/\alpha}\phi(x))$$

uniformly over $0 \le v \le h^{-1}$.

Proof. We use the arguments in the proof of Lemma 4.6 in [19]. Let $s_j = j/(\log n)^6$, $1 \le j < t_n$, where $t_n = 1 + [(\log n)^6 t]$, $s_{t_n} = t$. Write $[s_{j-1}, s_j] = \bigcup_{k=1}^{q_n}[s_{j,k-1}, s_{j,k}]$, where $s_{j,k} - s_{j,k-1} = (s_j - s_{j-1})/q_n$ and $q_n = [(s_j - s_{j-1})n^2]$. We have, for $s_{j,k-1} \le s \le s_{j,k}$,

$$\begin{aligned}
&\left|\hat{M}_n(v+s) - \hat{M}_n(v+s_{j,k-1})\right| \\
&\quad \le \frac{C}{\max_{1\le i\le n} m_i(\log n)^8}\sum_{i=1}^{n}\sum_{l=1}^{m_i}\left|K\left(\frac{t_{il}}{h} - v - s\right) - K\left(\frac{t_{il}}{h} - v - s_{j,k-1}\right)\right| \\
&\quad \le C_1 n^{-2/3} + \frac{C_1\sum_{i=1}^{n}\sum_{l=1}^{m_i} I\{|\frac{t_{il}}{h} - v - s_{j,k-1} \pm 1| \le C_2 n^{-2}\}}{\max_{1\le i\le n} m_i(\log n)^8}.
\end{aligned}$$

Put $I_{il} = I\{|\frac{t_{il}}{h} - v - s_{j,k-1} \pm 1| \le C_2 n^{-2}\}$. Then $\mathsf{E}I_{il} \le Chn^{-2}$. By Bernstein's inequality,

$$\begin{aligned}
&\mathsf{P}\left(\left|\sum_{i=1}^{n}\sum_{l=1}^{m_i}(I_{il} - \mathsf{E}I_{il})\right| \ge \max_{1\le i\le n} m_i(\log n)^2\right) \\
&\quad \le C_1\exp(-C_2 nh^{-1}(\log n)^4) + C_1\exp(-C_2(\log n)^2).
\end{aligned}$$

Hence, we have

$$\mathsf{P}\left(\max_{j,k}\sup_{s_{j,k-1}\le s\le s_{j,k}}\left|\hat{M}_n(v+s) - \hat{M}_n(v+s_{j,k-1})\right| \ge (\log n)^{-2}\right) = O(n^{-M})$$

for any $M > 0$. Note that we can show a similar inequality as that in Lemma 2 holds for $\hat{M}_n(t)$, and so by the property of $r(s)$ we have

$$\mathsf{E}\left(\hat{M}_n(v+s) - \hat{M}_n(v+s_{j,k-1})\right)^2 \le Cnh(\log n)^{-6}.$$

By Bernstein's inequality again, we have

$$\mathsf{P}\left(\max_{j,k}\left|\hat{M}_n(v+s_{j,k-1}) - \hat{M}_n(v+s_{j-1})\right| \ge (\log n)^{-2}\right) \le C_1 t_n n^2 e^{-C_2(\log n)^2}.$$

Combining the above arguments, we can obtain that

$$\mathsf{P}\Big(\max_{j} \sup_{s_{j-1}\le s\le s_j} \big|\hat{M}_n(v+s)-\hat{M}_n(v+s_{j-1})\big| \ge (\log n)^{-2}\Big) = O\big(n^{-M}\big).$$

It suffices to prove Lemma 2 holds for the probability $\mathsf{P}(\max_{1\le j\le t_n} |\hat{M}_n(v+s_j)| \ge x)$. The rest of the proof is similar to that of Lemma 4.6 in [19] and hence is omitted. □

Let $t=w$ in Lemma 3 with w being small enough. It follows that $\lim_{a\to 0}\limsup_{n\to\infty}[R_3+R_4]=0$. To prove Theorem 1, it suffices to show the following lemma holds.

Lemma 4. *Under the conditions of Theorem 1, for all $z\in R$, we have*

$$\lim_{a\to 0}\limsup_{v\to 0}\limsup_{n\to\infty}\Big|\mathsf{P}\Big(\Big\{\max_{1\le k\le N}\Lambda_k^+\ge x\Big\}\cup\Big\{\min_{1\le k\le N}\Lambda_k^-\le -x\Big\}\Big)-\big(1-e^{-2e^{-z}}\big)\Big|=0.$$

Proof. For $d\ge 1$, set

$$\hat{\mathbf{B}}_{k,j}=\big\{\hat{M}_n\big(a_k+jax^{-2/\alpha}\big)\ge x\big\}\cup\big\{\hat{M}_n\big(a_k+jax^{-2/\alpha}\big)\le -x\big\},$$
$$\hat{\mathbf{D}}_{k,j}^{\pm}=\big\{\hat{Y}_n\big(a_k+jax^{-2/\alpha}\big)\ge x\pm(\log n)^{-2d}\big\}\cup\big\{\hat{Y}_n\big(a_k+jax^{-2/\alpha}\big)\le -x\mp(\log n)^{-2d}\big\},$$

where $\hat{Y}_n(\cdot)$ is a centered Gaussian processes with covariance function satisfying

$$\mathsf{Cov}\big(\hat{Y}_n(s_1),\hat{Y}_n(s_2)\big)=\mathsf{Cov}\big(\hat{M}_n(s_1),\hat{M}_n(s_2)\big)$$

for $s_1\le s_2$. Let $\hat{\mathbf{A}}_k=\bigcup_{j=1}^{\chi}\hat{\mathbf{B}}_{k,j}$ and $\hat{\mathbf{C}}_k=\bigcup_{j=1}^{\chi}\hat{\mathbf{D}}_{k,j}$. Then

$$\mathsf{P}\Big(\Big\{\max_{1\le k\le N}\Lambda_k^+\ge x\Big\}\cup\Big\{\min_{1\le k\le N}\Lambda_k^-\le -x\Big\}\Big)=\mathsf{P}\Big(\bigcup_{k=1}^{N}\hat{\mathbf{A}}_k\Big).$$

Using the Bonferronis inequality, we have for any $l<[N/2]$,

$$\begin{aligned}&\sum_{d=1}^{2l}(-1)^{d-1}\sum_{1\le i_1<\cdots<i_d\le N}\mathsf{P}\Big(\bigcap_{j=1}^{d}\hat{\mathbf{A}}_{i_j}\Big)\\&\quad\le\mathsf{P}\Big(\bigcup_{k=1}^{N}\hat{\mathbf{A}}_k\Big)\le\sum_{d=1}^{2l-1}(-1)^{d-1}\sum_{1\le i_1<\cdots<i_d\le N}\mathsf{P}\Big(\bigcap_{j=1}^{d}\hat{\mathbf{A}}_{i_j}\Big).\end{aligned}\tag{A.2}$$

Write $\hat{\mathbf{C}}_k^{\pm}=\bigcup_{j=1}^{\chi}\hat{\mathbf{D}}_{k,j}^{\pm}$ and $\hat{\mathbf{C}}_k=\bigcup_{j=1}^{\chi}\hat{\mathbf{D}}_{k,j}^{\pm}$. By Theorem 1.1 in [32], we can obtain that for any $M>0$,

$$\mathsf{P}\Big(\bigcap_{j=1}^{d}\hat{\mathbf{C}}_{i_j}^{+}\Big)-Cn^{-M}\le\mathsf{P}\Big(\bigcap_{j=1}^{d}\hat{\mathbf{A}}_{i_j}\Big)\le\mathsf{P}\Big(\bigcap_{j=1}^{d}\hat{\mathbf{C}}_{i_j}^{-}\Big)+Cn^{-M}.\tag{A.3}$$

We now consider the probability $\mathsf{P}(\bigcap_{j=1}^{d} \hat{\mathbf{C}}_{i_j}^{\pm})$. Set $\hat{Z}_{k,j} = \hat{Y}_n(a_k + jax^{-2/\alpha})$ and denote the vector $\hat{\mathbf{Z}}_n = (\hat{Z}_{i_k,j}, 1 \le k \le d, 1 \le j \le \chi)$. Let $\hat{\Sigma}_n = \mathsf{Cov}(\hat{\mathbf{Z}}_n)$. Then we have, for some $\gamma > 0$,

$$\|\hat{\Sigma}_n - \tilde{\Sigma}_n\| = O\big(\chi h^{\gamma}\big),$$

where $\tilde{\Sigma}_n$ is the covariance matrix of $(Z_{i_k,j}, 1 \le k \le d, 1 \le j \le \chi)$, $Z_{k,j} = Y_n(a_k + jax^{-2/\alpha})$ and $Y_n(\cdot)$ is a centered Gaussian processes with covariance function $r(\cdot)$. Let $V_{k,j}$, $k \ge 1$, $j \ge 1$ be i.i.d. $N(0,1)$ random variables and $\delta_n = h^{\delta}$ for some $0 < \delta < \gamma/4$. Define

$$\begin{aligned}
\hat{\mathbf{D}}_{k,j,\delta}^{\pm} &= \big\{\hat{Z}_{k,j} + \delta_n V_{k,j} \ge x \pm 2(\log n)^{-2d}\big\} \cup \big\{\hat{Z}_{k,j} + \delta_n V_{k,j} \le -x \mp 2(\log n)^{-2d}\big\},\\
\mathbf{D}_{k,j,\delta}^{\pm} &= \big\{Z_{k,j} + \delta_n V_{k,j} \ge x \pm 2(\log n)^{-2d}\big\} \cup \big\{Z_{k,j} + \delta_n V_{k,j} \le -x \mp 2(\log n)^{-2d}\big\},\\
\mathbf{D}_{k,j}^{\pm} &= \big\{Z_{k,j} \ge x \pm 3(\log n)^{-2d}\big\} \cup \big\{Z_{k,j} \le -x \mp 3(\log n)^{-2d}\big\},\\
\hat{\mathbf{C}}_{k,\delta}^{\pm} &= \bigcup_{j=1}^{\chi} \hat{\mathbf{D}}_{k,j,\delta}^{\pm}, \qquad \mathbf{C}_{k,\delta}^{\pm} = \bigcup_{j=1}^{\chi} \mathbf{D}_{k,j,\delta}^{\pm}, \qquad \mathbf{C}_{k}^{\pm} = \bigcup_{j=1}^{\chi} \mathbf{D}_{k,j}^{\pm}.
\end{aligned}$$

By the tail probability of the normal distribution, we can show that for any $M > 0$,

$$\mathsf{P}\left(\bigcap_{j=1}^{d} \hat{\mathbf{C}}_{i_j,\delta}^{+}\right) - Cn^{-M} \le \mathsf{P}\left(\bigcap_{j=1}^{d} \hat{\mathbf{C}}_{i_j}^{\pm}\right) \le \mathsf{P}\left(\bigcap_{j=1}^{d} \hat{\mathbf{C}}_{i_j,\delta}^{-}\right) + Cn^{-M}. \tag{A.4}$$

Let $\hat{\Sigma}_n^{\delta}$ and $\tilde{\Sigma}_n^{\delta}$ be the covariance matrices of $(\hat{Z}_{i_k,j} + \delta_n V_{k,j}, 1 \le k \le d, 1 \le j \le \chi)$ and $(Z_{i_k,j} + \delta_n V_{k,j}, 1 \le k \le d, 1 \le j \le \chi)$, respectively. We have

$$\|\hat{\Sigma}_n^{\delta} - \tilde{\Sigma}_n^{\delta}\| = O\big(\chi h^{\gamma}\big)$$

for some $\gamma > 0$. Note that $\hat{\Sigma}_n^{\delta}$ and $\tilde{\Sigma}_n^{\delta}$ are positive definitive and the smallest eigenvalues are larger than δ_n^2. So we have

$$\big\|\big(\hat{\Sigma}_n^{\delta}\big)^{-1} - \big(\tilde{\Sigma}_n^{\delta}\big)^{-1}\big\| = O\big(\chi h^{\gamma} \delta_n^4\big).$$

By the density function of multivariate normal vector and some tedious calculations, we can prove that

$$\mathsf{P}\left(\bigcap_{j=1}^{d} \hat{\mathbf{C}}_{i_j,\delta}^{\pm}\right) = \big(1 + O\big(n^{-\tau}\big)\big)\mathsf{P}\left(\bigcap_{j=1}^{d} \mathbf{C}_{i_j,\delta}^{\pm}\right) + O\big(n^{-M}\big) \tag{A.5}$$

for some $\tau > 0$ and any $M > 0$. Also, by the tail probability of the normal distribution,

$$\mathsf{P}\left(\bigcap_{j=1}^{d} \mathbf{C}_{i_j}^{+}\right) - O\big(n^{-M}\big) \le \mathsf{P}\left(\bigcap_{j=1}^{d} \mathbf{C}_{i_j,\delta}^{\pm}\right) \le \mathsf{P}\left(\bigcap_{j=1}^{d} \mathbf{C}_{i_j}^{-}\right) + O\big(n^{-M}\big). \tag{A.6}$$

We now only need to consider $\mathsf{P}(\bigcap_{j=1}^{d}\mathbf{C}_{i_j}^{\pm})$. Define $q_j = i_{j+1} - i_j$, $1 \le j \le d-1$, and

$$\mathcal{I} = \Big\{1 \le i_1 < \cdots < i_d \le N : \min_{1\le j\le d-1} q_j \le \big[2w^{-1}+2\big]\Big\}.$$

As the proof of Lemma 4.10 in [19], we have

$$\sum_{(i_1,\ldots,i_d)\in\mathcal{I}} \mathsf{P}\Bigg(\bigcap_{j=1}^{d}\mathbf{C}_{i_j}^{\pm}\Bigg) \le Cb^{\tau} \tag{A.7}$$

for some $\tau > 0$. Note that $r(t) = 0$ for all $t \ge 2$. Hence, for $(i_1,\ldots,i_d) \notin \mathcal{I}$, $\mathbf{C}_{i_j}^{\pm}$, $1 \le i_1 < \cdots < i_d \le N$ are independent. Also, $\mathrm{Card}(\mathcal{I}) = O(b^{-d+1})$. So we have

$$\begin{aligned}\Bigg(\sum_{1\le i_1<\cdots<i_d\le N} - \sum_{\mathcal{I}}\Bigg)\mathsf{P}\Bigg(\bigcap_{j=1}^{d}\mathbf{C}_{i_j}^{\pm}\Bigg) &= \Bigg(\sum_{1\le i_1<\cdots<i_d\le N} - \sum_{\mathcal{I}}\Bigg)\prod_{j=1}^{d}\mathsf{P}\big(\mathbf{C}_{i_j}^{\pm}\big) \\ &= \big(1+o(1)\big)\frac{N^d}{d!}\bigg(x^{2/\alpha}\phi(x)\frac{H_\alpha(a)}{a}C_0^{1/\alpha}w\bigg)^d.\end{aligned} \tag{A.8}$$

Submitting (A.3)–(A.8) into (A.2) and using some elementary calculations, we prove Lemma 3. □

Let $\tilde{K}_k(x) = x^k K(x)$ for integers $k \ge 0$. Note that $\tilde{K}_k(x)$ satisfies (A6). We can define $d_{n,k}$ and $\lambda_{\tilde{K}_k}$ as in Section 2 by replacing $K(x)$ with $\tilde{K}_k(x)$. Let

$$\hat{f}_k(t) = \frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\tilde{K}_k\bigg(\frac{t_{ij}-t}{h}\bigg).$$

The above arguments in fact implies that

$$\mathsf{P}\Bigg[\big(2\log h^{-1}\big)^{1/2}\Bigg(\sup_{0\le t\le 1}\Bigg|\sqrt{\frac{Nh}{\lambda_{\tilde{K}}f(t)}}\big[\hat{f}_k(t) - \mathsf{E}\hat{f}_k(t)\big]\Bigg| - d_{n,k}\Bigg) \le z\}\Bigg] \to e^{-2e^{-z}}. \tag{A.9}$$

Hence (3.4) holds by taking $k = 0$. It is easy to prove that

$$\sup_{0\le t\le 1}\bigg|\mathsf{E}\big[\hat{f}_k(t)\big] - f(t)\int_{-A}^{A}\tilde{K}_k(x)\,dx - hf'(t)\int_{-A}^{A}x\tilde{K}_k(x)\,dx\bigg| \le Ch^2.$$

So by (A.9) we have

$$\sup_{0\le t\le 1}\bigg|\hat{f}_k(t) - f(t)\int_{-A}^{A}\tilde{K}_k(x)\,dx - hf'(t)\int_{-A}^{A}x\tilde{K}_k(x)\,dx\bigg| = O_{\mathsf{P}}\Bigg(h^2 + \sqrt{\frac{\log h^{-1}}{Nh}}\Bigg). \tag{A.10}$$

A.2. Proof of Theorem 2

Write

$$\begin{aligned}
&\sum_{i=1}^{n} X_i(t_{ij})^T X_i(t_{ij}) \sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\} \\
&\quad = \sum_{i=1}^{n}\left[X_i(t_{ij})^T X_i(t_{ij}) - \mathsf{E}_* X_i(t_{ij})^T X_i(t_{ij})\right] \sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\} \\
&\qquad + \sum_{i=1}^{n}\left[\mathsf{E}_* X_i(t_{ij})^T X_i(t_{ij})\right] \sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\},
\end{aligned}$$

where $\mathsf{E}_*(\cdot)$ denotes the conditional expectation given $\{t_{ij}\}$. Denote $X_i(t)^T X_i(t) - \mathsf{E}_* X_i(t)^T X_i(t) = (r_{i,kl}(t))_{1\le k,l\le p}$ and

$$Q_{n,kl}(t) = \frac{1}{\sqrt{Nh\,\mathrm{Var}[r_{1,kl}(t)]\lambda_{K_v} f(t)}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} r_{i,kl}(t_{ij}) \tilde{K}_v\{(t_{ij}-t)/h\}.$$

Following exactly the same proof of Theorem 1, we have

$$\mathsf{P}\left[\left(2\log h^{-1}\right)^{1/2}\left(\sup_{0\le t\le 1}\left|Q_{n,kl}(t)\right| - d_n\right) \le z\right] \to e^{-2e^{-z}}.$$

Note that

$$\begin{aligned}
&\frac{1}{Nh}\sum_{i=1}^{n}\left[\mathsf{E}_* X_i(t_{ij})^T X_i(t_{ij})\right] \sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\} \\
&\quad = \Sigma(t)\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\} + O(1)h\frac{1}{Nh}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left|\tilde{K}_{v+1}\{(t_{ij}-t)/h\}\right|.
\end{aligned}$$

This, together with (A.10), implies

$$\begin{aligned}
&\sup_{0\le t\le 1}\left\|\frac{1}{Nh}\sum_{i=1}^{n} X_i(t_{ij})^T X_i(t_{ij}) \sum_{j=1}^{m_i} \tilde{K}_v\{(t_{ij}-t)/h\} - \Sigma_p(t) f(t)\int_{-A}^{A} \tilde{K}_v(x)\,dx\right\| \\
&\quad = O_{\mathsf{P}}\left(\sqrt{\frac{\log h^{-1}}{Nh}} + h\right).
\end{aligned} \tag{A.11}$$

Define $e = (I_{p\times p}, 0_{p\times p})$, where $I_{p\times p}$ is a $p\times p$ identity matrix and $0_{p\times p}$ is a $p\times p$ zero matrix. We have

$$\hat{\beta}(t) - \beta(t) = e\mathbf{S}_n^{-1}(t)\mathbf{V}_n(t) + e\mathbf{S}_n^{-1}(t)\mathbf{Z}_n(t) + e\mathbf{S}_n^{-1}(t)\mathbf{U}_n(t), \tag{A.12}$$

where

$$\mathbf{V}_n(t) = \begin{pmatrix} \mathbf{V}_{n,1}(t) \\ \mathbf{V}_{n,2}(t) \end{pmatrix}, \qquad \mathbf{Z}_n(t) = \begin{pmatrix} \mathbf{Z}_{n,1}(t) \\ \mathbf{Z}_{n,2}(t) \end{pmatrix}, \qquad \mathbf{U}_n(t) = \begin{pmatrix} \mathbf{U}_{n,1}(t) \\ \mathbf{U}_{n,2}(t) \end{pmatrix}.$$

Here

$$\mathbf{V}_{n,l}(t) = (nh)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} X_{ij} \varepsilon_{ij} \tilde{K}_{l-1}\big((t_{ij} - t)/h\big),$$

$$\mathbf{Z}_{n,l}(t) = \frac{h^2}{2} \left[\frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{m_i} X_{ij} X_{ij}^T \beta''(t) \tilde{K}_{l+1}\big((t_{ij} - t)/h\big) \right],$$

$$\big\| \mathbf{U}_{n,l}(t) \big\| \le C h^3 \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \big\| X_{ij} X_{ij}^T \big\| \big| \tilde{K}_{l+2}\big((t_{ij} - t)/h\big) \big|.$$

Define

$$Q_n(t) = \frac{1}{\sqrt{\mathbf{a}(t)^T \Sigma_p^{-1}(t) \Xi_p(t) \Sigma_p^{-1}(t) \mathbf{a}(t) \lambda_{\tilde{K}_l} N h f(t)}} \times \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathbf{a}(t)^T \Sigma_p^{-1}(t) X_{ij} \varepsilon_{ij} \tilde{K}_l \big\{ (t_{ij} - t)/h \big\}.$$

By the same proof of Theorem 1 again, for any $\mathbf{a}(t) \in R^{k+1}$,

$$\mathsf{P}\Big[\big(2 \log h^{-1}\big)^{1/2} \Big(\sup_{0 \le t \le 1} \big| Q_n(t) \big| - d_n \Big) \le z \Big] \to e^{-2e^{-z}}. \tag{A.13}$$

Note that

$$\left\| \mathbf{S}_n(t) - f(t) \operatorname{diag}\left(\Sigma_p(t), \Sigma_p(t) \int_{-A}^{A} x^2 K(x)\, dx \right) \right\| = O_{\mathsf{P}}\left(\sqrt{\frac{\log h^{-1}}{Nh}} + h \right).$$

The theorem is proved by (A.11), (A.12) and (A.13).

Acknowledgements

The authors are grateful to the Rush Alzheimer's Disease Center for making the Chicago Health and Aging Project dataset available for us to use. Hongyuan Cao's research is partially supported by a University of Missouri Research Board grant. Zhou Zhou's research is supported in part by NSERC of Canada.

References

[1] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095.

[2] Bienias, J.L., Beckett, L.A., Bennett, D.A., Wilson, R.S. and Evans, D.A. (2003). Design of the Chicago health and aging project. *J. Alzheimers Dis.* **5** 349–355.

[3] Cao, G., Yang, L. and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *J. Nonparametr. Stat.* **24** 359–377. MR2921141

[4] Cao, H., Zeng, D. and Fine, J.P. (2015). Regression analysis of sparse asynchronous longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **77** 755–776. MR3382596

[5] Chiang, C.-T., Rice, J.A. and Wu, C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96** 605–619. MR1946428

[6] Degras, D.A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica* **21** 1735–1765. MR2895997

[7] Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edition ed. Oxford: Oxford Univ. Press.

[8] Eubank, R.L. and Speckman, P.L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88** 1287–1301. MR1245362

[9] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

[10] Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 303–322. MR1749541

[11] Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Statist.* **27** 715–731. MR1804172

[12] Gu, L., Wang, L., Härdle, W. and Yang, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST* **23** 806–843.

[13] Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **55** 757–796.

[14] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85** 809–822. MR1666699

[15] Huang, J.Z., Wu, C.O. and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89** 111–128. MR1888349

[16] Johnston, G.J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 402–414. MR0666014

[17] Liang, Y., Lu, W. and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* **65** 377–384. MR2751461

[18] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons]. MR1925014

[19] Liu, W. and Wu, W.B. (2010). Simultaneous nonparametric inference of time series. *Ann. Statist.* **38** 2388–2421. MR2676893

[20] Ma, S., Yang, L. and Carroll, R.J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statist. Sinica* **22** 95–122. MR2933169

[21] Rice, J.A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

[22] Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **53** 683–690.

[23] Sloan, F.A. and Wang, J. (2005). Disparities among older adults in measures of cognitive function by race or ethnicity. *J. Gerontol.*, *Ser. B*, *Psychol. Sci. Soc. Sci.* **60** P242–P250.

[24] Sun, J., Park, D., Sun, L. and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *J. Amer. Statist. Assoc.* **100** 882–889.

[25] Wang, J. and Yang, L. (2009). Polynomial spline confidence bands for regression curves. *Statist. Sinica* **19** 325–342.

[26] Wilson, R.S., Aggarwal, N.T., Barnes, L.L., Bienias, J.L., Mendes de Leon, C.F. and Evans, D.A. (2009). Biracial population study of mortality in mild cognitive impairment and Alzheimer disease. *Arch. Neurol.* **66** 767–772.

[27] Wilson, R.S., Hebert, L.E., Scherr, P.A., Barnes, L.L., Mendes de Leon, C.F. and Evans, D.A. (2009). Educational attainment and cognitive decline in old age. *Neurology* **72** 460–465.

[28] Wu, C.O., Chiang, C.T. and Hoover, D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.* **93** 1388–1402.

[29] Wu, W. and Zhao, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 391–410.

[30] Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 797–811.

[31] Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903.

[32] Zaĭtsev, A.Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S. N. Bernstein's inequality conditions. *Probab. Theory Related Fields* **74** 535–566.

[33] Zhao, Z. and Wu, W.B. (2008). Confidence bands in nonparametric time series regression. *Ann. Statist.* **36** 1854–1878.

[34] Zheng, S., Yang, L. and Härdle, W. (2014). A smooth simultaneous confidence corridor for the mean of sparse functional data. *J. Amer. Statist. Assoc.* **109** 661–673.

[35] Zhou, Z. and Wu, W. (2010). Simultaneous inference of linear models with time-varying coefficients. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 513–531.

Received March 2016 and revised January 2017