

# Semiparametric Estimation of Covariance Matrixes for Longitudinal Data

Jianqing FAN and Yichao WU

---

Estimation of longitudinal data covariance structure poses significant challenges because the data usually are collected at irregular time points. A viable semiparametric model for covariance matrixes has been proposed that allows one to estimate the variance function non-parametrically and to estimate the correlation function parametrically by aggregating information from irregular and sparse data points within each subject. But the asymptotic properties of the quasi-maximum likelihood estimator (QMLE) of parameters in the covariance model are largely unknown. We address this problem in the context of more general models for the conditional mean function, including parametric, nonparametric, or semiparametric. We also consider the possibility of rough mean regression function and introduce the difference-based method to reduce biases in the context of varying-coefficient partially linear mean regression models. This provides a more robust estimator of the covariance function under a wider range of situations. Under some technical conditions, consistency and asymptotic normality are obtained for the QMLE of the parameters in the correlation function. Simulation studies and a real data example are used to illustrate the proposed approach.

KEY WORDS: Correlation structure; Difference-based estimation; Quasi-maximum likelihood; Varying-coefficient partially linear model.

---

## 1. INTRODUCTION

Longitudinal data (Diggle, Heagerty, Liang, and Zeger 2002) are characterized by repeated observations over time on the same set of individuals. Observations on the same subject tend to be correlated. As a result, a core issue for analyzing longitudinal data is the estimation of its covariance structure. Good estimation of the covariance structure improves the efficiency of model estimation and results in better predictions of individual trajectories over time. However, the challenge of covariance matrix estimation comes from the fact that measurements are often taken at sparse and subject-dependent irregular time points, as illustrated in the following typical example.

Progesterone, a reproductive hormone, is responsible for normal fertility and menstrual cycling. A longitudinal hormone study on progesterone (Sowers et al. 1998) collected urine samples from 34 healthy women in a menstrual cycle on alternative days. Zhang, Lin, Raz, and Sowers (1998) analyzed the data using semiparametric stochastic mixed models. A total of 492 observations were made on the 34 subjects in the study, with 11–28 observations per subject. The subjects' menstrual cycle lengths ranged from 23 to 56 days (average, 29.6 days). Biologically, it makes sense to assume that a woman's change in progesterone level depends on the time during a menstrual cycle relative to her cycle length; thus, the menstrual cycle length of each woman was standardized (Sowers et al. 1998). A typical logarithmic transformation was applied on the progesterone level to make the data more homoscedastic. The progesterone data were unbalanced in that different subjects have different numbers of observations, and observation times were not regular and differed among subjects.

To address these challenges in covariance matrix estimation, Fan, Huang, and Li (2007) modeled the variance function non-parametrically and correlation structure parametrically. They

focused mainly on the improvement in the estimation of the mean regression function using a possibly misspecified covariance structure. In this work we focus on semiparametric modeling of the covariance matrix itself, with emphasis on the asymptotic properties of the quasi-maximum likelihood estimator (QMLE) of parameters in correlation function. Thus, we study the problem under a general mean-regression model,

$$y(t) = m(\mathbf{x}(t)) + \varepsilon(t), \quad t \in \mathcal{T}, \quad (1)$$

where  $t$  indexes time in the longitudinal data and the conditional mean function  $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$  can be parametric, nonparametric, or semiparametric. The semiparametric covariance structure is specified as

$$\text{var}(\varepsilon(t)) = \sigma^2(t), \quad \text{corr}(\varepsilon(s), \varepsilon(t)) = \rho(s, t, \boldsymbol{\theta}),$$

where  $\rho(\cdot, \cdot, \boldsymbol{\theta})$  is a positive definite function for any  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ . The model is flexible, especially when the number of parameters in  $\boldsymbol{\theta}$  is large. On the other hand, the model is estimable even when individuals have only a few data points observed sparsely in time. The variance function can be estimated using the marginal information of the data, as long as the aggregated time points of all subjects are dense in time. The parameters  $\boldsymbol{\theta}$  can be estimated by aggregating information from all individuals whose responses are observed at two or more time points. The novelty of this family of models is that it has time sparsity and irregularity of longitudinal data at its heart.

This semiparametric covariance structure is very flexible and basically covers any possibility by allowing more parameters in the correlation structure. For example, one may consider a convex combination of different parametric correlation structures, such as ARMA or random-effect models. But correct specification of the correlation structure generally requires relatively few parameters. For the progesterone data, the response is taken as the change of progesterone level from an individual's average level. Biologically, we can imagine that for two observations on the same subject, the closer their observation times, the greater the correlation in the response. Thus we use an ARMA(1, 1) correlation structure when analyzing the progesterone data in Section 6.

---

Jianqing Fan is Frederick L. Moore'18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, and Honorary Professor, Department of Statistics, Shanghai University of Finance and Economics, Shanghai, China (E-mail: [jqfan@princeton.edu](mailto:jqfan@princeton.edu)). Yichao Wu is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh NC 27695 (E-mail: [wu@stat.ncsu.edu](mailto:wu@stat.ncsu.edu)). This research is supported in part by National Science Foundation grant DMS-03-54223 and National Institutes of Health grant R01-GM07261. The authors thank the editor the associate editor, and two referees for their helpful comments that led to improvements to the manuscript.

Our approach of flexible covariance structure estimation provides insight into solving a long-standing problem of improving the efficiency of parameter estimation ideally using the unknown true covariance structure. In a seminal article, Liang and Zeger (1986) introduced generalized estimating equations (GEEs), extending generalized linear models to longitudinal data, and proposed using a working correlation matrix to improve efficiency. Misspecification of the working correlation matrix is possible, however. To improve efficiency under misspecification, Qu, Lindsay, and Li (2000) represented the inverse of the working correlation matrix by a linear combination of basis matrixes and proposed a method using quadratic inference functions. Their theoretical and simulated results showed better efficiency than GEE when misspecification occurs. In a nonparametric setting, Lin and Carroll (2000) extended GEE to kernel GEE and reported a rather unexpected result that higher efficiency was obtained by assuming independence than by using the true correlation structure. In their later work, they showed that the true covariance function can be used to improve the variance of a nonparametric estimator. Wang (2003) provided a deep understanding for this result, proposed an alternative kernel smoothing method, and established the asymptotic result that the new estimator achieves the minimum variance when the correlation is correctly specified. Wang, Carroll, and Lin (2005) extended it to the semiparametric, partially linear models. All of these works required specifying a true correlation matrix, but did not provide a systematic estimation scheme. Our method provides a flexible approach to this important endeavor.

There are several approaches to estimating a covariance matrix. Most are nonparametric. Wu and Pourahmadi (2003) used nonparametric smoothing to regularize the estimation of large covariance matrix based on the method of two-step estimation studied by Fan and Zhang (2000). Huang, Liu, and Liu (2007) used the modified Cholesky decomposition of the covariance matrix, proposed a more direct approach of smoothing, and claimed their estimation is more efficient than that of Wu and Pourahmadi (2003). Bickel and Levina (2006) investigated the regularization of covariance matrix through the idea of banding and obtained many insightful results (see also Rothman, Bickel, Levina, and Zhu 2007). All of the aforementioned estimation methods have the same limitation that observations are assumed to be made over a grid, that is, balanced or nearly balanced longitudinal data. But this assumption often is not tenable, especially for longitudinal data that are commonly collected at irregular and possibly subject-specific time points. This feature of longitudinal data makes studying its covariance structure a challenge. Yao, Müller, and Wang (2005a,b) took a different approach based on functional data analysis.

The remainder of the article is organized as follows. In Section 2 we discuss estimation of the conditional mean function  $m(\mathbf{x})$ . We focus on semiparametric, varying-coefficient, partially linear models, for which we propose a more robust estimation scheme. In Section 3 we present the estimation method for the semiparametric covariance structure. We give the asymptotic properties of the estimated variance function and correlation structure parameter in Section 4, and present extensive simulation studies and an application to the progesterone data in Sections 5 and 6. We provide technical proofs in the Appendix.

## 2. SEMIPARAMETRIC VARYING-COEFFICIENT PARTIALLY LINEAR MODEL

Our data consist of a series of observations made on a random sample of  $n$  subjects from model (1). We denote a generic subject with  $J$  pairs of observation  $(\mathbf{x}(t_j), y(t_j))$  at times  $\{t_j\}$  from model (1) by  $\mathbb{X} = \{J, (t_j, \mathbf{x}(t_j), y(t_j)), j = 1, 2, \dots, J\}$ , with data from subject  $i$  denoted by  $\mathbb{X}_i = \{J_i, (t_{ij}, \mathbf{x}_i(t_{ij}), y_i(t_{ij})), j = 1, 2, \dots, J_i\}$ . Thus the complete data set is represented as  $\{\mathbb{X}_1, \mathbb{X}_2, \dots, \mathbb{X}_n\}$ .

Note that our semiparametric specification of the covariance structure has few requirements for  $m(\mathbf{x})$  in (1); it can be of any form as long as it is consistently estimated. In this section we focus on the semiparametric, varying-coefficient, partially linear model considered by Fan et al. (2007),

$$m(\mathbf{x}(t)) = m(\mathbf{x}_1(t), \mathbf{x}_2(t)) = \mathbf{x}_1(t)^T \boldsymbol{\alpha}(t) + \mathbf{x}_2(t)^T \boldsymbol{\beta}, \quad (2)$$

where  $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ ,  $\mathbf{x}_1 \in \mathbb{R}^{d_1}$ , and  $\mathbf{x}_2 \in \mathbb{R}^{d_2}$ . It includes parametric and nonparametric models as special cases. We denote the true coefficients by  $\boldsymbol{\alpha}_0(\cdot)$  and  $\boldsymbol{\beta}_0$ . In this case we denote the covariates at the  $j$ th observation time of subject  $i$  by  $\mathbf{x}_{1,i}(t_{ij})$  and  $\mathbf{x}_{2,i}(t_{ij})$ .

To estimate the varying-coefficient partially linear model, Fan et al. (2007) assumed the existence of a second-order derivative of  $\boldsymbol{\alpha}_0(\cdot)$ . But this assumption is not always desirable; for example, it does not hold for continuous piecewise linear  $\boldsymbol{\alpha}_0(\cdot)$ . When  $\boldsymbol{\alpha}_0(\cdot)$  is rough, as in Example 3 in Section 5, estimation can be adversely affected. In reality,  $\boldsymbol{\alpha}_0(\cdot)$  can be rough as well; for example, it occurs when there is some structure break or technical innovation that causes the change point in time.

Departing from Fan et al. (2007), we consider (2) under a much weaker smoothness assumption on  $\boldsymbol{\alpha}_0(\cdot)$  as stated in Condition (L) and propose a more robust estimation scheme for  $m(\cdot)$  than their profiling scheme:

- (L) There are constants  $a_0 > 0$  and  $\kappa > 0$  such that  $\|\boldsymbol{\alpha}_0(s) - \boldsymbol{\alpha}_0(t)\| \leq a_0|t - s|^\kappa$  when  $|t - s|$  is small for  $0 \leq s, t \leq T$ .

### 2.1 Difference-Based Estimator of $\boldsymbol{\beta}$

In this section we use a difference-based technique to estimate the parametric regression coefficient  $\boldsymbol{\beta}$  under the Lipschitz condition (L) on  $\boldsymbol{\alpha}_0(\cdot)$ . A comparison study in Section 5 shows its significant improvement when  $\boldsymbol{\alpha}_0(\cdot)$  is rough. This technique has been used to remove the nonparametric component in the partial linear model or nonparametric heteroscedastic model by various authors (e.g., Yatchew 1997; Fan and Huang 2001, 2005; Brown and Levine 2007), but all in the univariate nonparametric setup. In our setting, multiple nonparametric functions need to be removed, and new ideas are required.

To apply the difference-based technique, we sort our data in the increasing order of the observation time  $t_{ij}$  and denote them as  $\{(t_{(j)}, \mathbf{x}_{1(j)}, \mathbf{x}_{2(j)}, y_{(j)}), j = 1, 2, \dots, N\}$ , where  $N = \sum_{i=1}^n J_i$ . Under mild conditions, the spacing  $t_{(i+j)} - t_{(i)}$  can be shown to be of order  $O_p(1/N) = O_p(1/n)$  for each given  $j$ . Condition (L) implies that

$$\begin{aligned} \|\boldsymbol{\alpha}_0(t_{(i+j)}) - \boldsymbol{\alpha}_0(t_{(i)})\| &= O_p((t_{(i+j)} - t_{(i)})^\kappa) \\ &= O_p(1/n^\kappa), \quad \text{for } j = 1, 2, \dots, d_1. \end{aligned} \quad (3)$$

For any  $i$  between 1 and  $N - d_1$ , choose weights  $w_{i,j}, j = 1, 2, \dots, N$ , and define the following weighted variables:

$$y_{(i)}^* = \sum_{j=i}^{i+d_1} w_{i,j} y_{(j)}, \quad \mathbf{x}_{2(i)}^* = \sum_{j=i}^{i+d_1} w_{i,j} \mathbf{x}_{2(j)}$$

and

$$\varepsilon_{(i)}^* = \sum_{j=i}^{i+d_1} w_{i,j} \varepsilon_{(j)}.$$

The weights  $w_{i,j}$ 's are selected such that

$$\sum_{j=i}^{i+d_1} w_{i,j} \mathbf{x}_{1(j)} = \mathbf{0} \quad \text{for all } i, 1 \leq i \leq N - d_1, \quad (4)$$

to remove the nonparametric component  $\mathbf{x}_1(t)^T \boldsymbol{\alpha}(t)$ . The weights are further normalized as done by Hall, Kay, and Titterton (1990) such that  $\sum_{j=1}^N w_{i,j}^2 = 1$ . Note that if  $\mathbf{x}_{1(i)}, \mathbf{x}_{1(i+1)}, \dots, \mathbf{x}_{1(i+d_1-1)}$  are linearly independent (which holds with probability 1 under some mild conditions for continuously distributed  $\mathbf{x}_1$ ), then the weights are uniquely determined up to a sign change; more explicitly,

$$\begin{aligned} & (w_{i,i}, w_{i,i+1}, \dots, w_{i,i+d_1}) \\ &= \pm \frac{(((\mathbf{x}_{1(i)}, \mathbf{x}_{1(i+1)}, \dots, \mathbf{x}_{1(i+d_1-1)})^{-1} \mathbf{x}_{1(i+d_1)})^T, -1)}{\|((\mathbf{x}_{1(i)}, \mathbf{x}_{1(i+1)}, \dots, \mathbf{x}_{1(i+d_1-1)})^{-1} \mathbf{x}_{1(i+d_1)})^T, -1\|}. \end{aligned} \quad (5)$$

Without loss of generality, we can take a positive sign in (5) and denote the corresponding  $(N - d_1) \times N$  weight matrix by  $\mathbf{W}$ , with  $(i, j)$  element  $w_{i,j}$ . Writing  $\tilde{\mathbf{X}}_2 = (\mathbf{x}_{2(1)}, \mathbf{x}_{2(2)}, \dots, \mathbf{x}_{2(N)})^T$ , we have  $\mathbf{X}_2^* = (\mathbf{x}_{2(1)}^*, \mathbf{x}_{2(2)}^*, \dots, \mathbf{x}_{2(N-d_1)}^*)^T = \mathbf{W} \tilde{\mathbf{X}}_2$ .

A combination of (2)–(4) leads to

$$y_{(i)}^* \approx (\mathbf{x}_{2(i)}^*)^T \boldsymbol{\beta} + \varepsilon_{(i)}^*, \quad i = 1, 2, \dots, N - d_1, \quad (6)$$

where the approximation error is of order  $O_p(1/n^\kappa)$ .

Model (6) is a standard multivariate linear regression problem. Applying the ordinary least squares (OLS) technique on (6) with data  $\{(\mathbf{x}_{2(i)}^*, y_{(i)}^*), i = 1, 2, \dots, N - p\}$ , we get the difference-based estimator (DBE) of  $\boldsymbol{\beta}$ . Because the approximation error in (6) is of order  $O_p(1/n^\kappa)$ , a standard result for the OLS implies that our DBE  $\hat{\boldsymbol{\beta}}$  is consistent with order  $n^{-\kappa \wedge .5}$ , where  $a \wedge b = \min(a, b)$ .

In general, we can use  $d_1 + k$  ( $k > 1$ ) neighboring observations to remove the nonparametric terms. In such a case the weights are not uniquely determined, and finding an optimal weighting scheme is complicated.

### 2.2 Kernel Smoothing Estimator of $\boldsymbol{\alpha}(\cdot)$

Plug the consistent DBE  $\hat{\boldsymbol{\beta}}$  into model (2) and define  $\tilde{y}(t) = y(t) - \mathbf{x}_2(t)^T \hat{\boldsymbol{\beta}}$ . Model (2) becomes

$$\tilde{y}(t) \approx \mathbf{x}_1(t)^T \boldsymbol{\alpha}(t) + \varepsilon(t), \quad (7)$$

with approximation error of order  $O_p(n^{-\kappa \wedge .5})$ . Model (7) is exactly a varying-coefficient model; it was studied by Hastie and Tibshirani (1993) for the case of iid observations and also by Fan and Zhang (2000) in the context of longitudinal data.

A local smoothing technique can be used to estimate  $\boldsymbol{\alpha}(\cdot)$ . Because of the weak Lipschitz Condition (L) on true  $\boldsymbol{\alpha}_0(\cdot)$ , we use the local constant regression (Nadaraya–Watson estimator) to estimate  $\boldsymbol{\alpha}(\cdot)$  based on data  $\{(t_{ij}, \mathbf{x}_{1,i}(t_{ij}), \tilde{y}_i(t_{ij})), j = 1, 2, \dots, J_i; i = 1, 2, \dots, n\}$ .

Note that for any  $t$  in a neighborhood of  $t_0$ , due to Condition (L), we have

$$\alpha_{0l}(t) \approx \alpha_{0l}(t_0) + O(|t - t_0|^k) \quad \text{for } l = 1, 2, \dots, q,$$

where  $\alpha_{0l}(\cdot)$  is the  $l$ th component of  $\boldsymbol{\alpha}_0(\cdot)$ . Local constant regression estimates  $\boldsymbol{\alpha}(t_0)$  by minimizing

$$\sum_{i=1}^n \sum_{j=1}^{J_i} (\tilde{y}_i(t_{ij}) - \mathbf{a}^T \mathbf{x}_{1,i}(t_{ij}))^2 K_b(t_{ij} - t_0), \quad (8)$$

with respect to  $\mathbf{a} = (a_1, a_2, \dots, a_q)^T$ . Denote  $\tilde{\mathbf{y}} = (\tilde{y}_1(t_{11}), \tilde{y}_1(t_{12}), \dots, \tilde{y}_n(t_{nJ_n}))^T$  and  $\mathbf{K}_{t_0} = \text{diag}(K_b(t_{11} - t_0), K_b(t_{12} - t_0), \dots, K_b(t_{nJ_n} - t_0))$ . The local constant regression estimator of  $\boldsymbol{\alpha}(t_0)$  is

$$\hat{\boldsymbol{\alpha}}(t_0) \equiv \hat{\boldsymbol{\alpha}}(t_0, \hat{\boldsymbol{\beta}}) = \hat{\mathbf{a}} = (\mathbf{X}_1^T \mathbf{K}_{t_0} \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{K}_{t_0} \tilde{\mathbf{y}},$$

where  $\mathbf{X}_1 = (\mathbf{x}_{1,1}(t_{11}), \mathbf{x}_{1,1}(t_{12}), \dots, \mathbf{x}_{1,n}(t_{nJ_n}))^T$ . A typical asymptotic nonparametric convergence rate applies to the local constant regression estimator  $\hat{\boldsymbol{\alpha}}(\cdot)$ , and the rate is of order  $O_p(b^\kappa + 1/\sqrt{nb}) = O_p(n^{-\kappa/(2\kappa+1)})$  when  $b = O(n^{-1/(2\kappa+1)})$ .

### 3. ESTIMATION OF SEMIPARAMETRIC COVARIANCE

We assume that  $m(\cdot)$  can be consistently estimated by some method depending on its particular form. This consistency assumption is formulated as Condition (C) in the Appendix. We next present our estimation scheme of the semiparametric covariance structure.

#### 3.1 Estimation of the Variance Function

Denote the estimated conditional mean function by  $\hat{m}(\cdot)$ . Plug it into (1) and define the estimated realizations of the random errors as

$$r_{ij} \equiv r_{ij}(\hat{m}(\cdot)) = y_i(t_{ij}) - \hat{m}(\mathbf{x}_i(t_{ij})), \quad (9)$$

which consistently estimate the realized random errors  $\varepsilon_i(t_{ij})$  due to the consistency assumption on  $\hat{m}(\cdot)$ .

Based on  $r_{ij}$ , we use kernel smoothing to estimate the variance function  $\sigma^2(t) = E\varepsilon^2(t)$  by

$$\hat{\sigma}^2(t) = \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} r_{ij}^2 K_h(t - t_{ij})}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t - t_{ij})}, \quad (10)$$

where  $h$  is a smoothing parameter and  $K_h(\cdot) = K(\cdot/h)/h$  is a rescaling of the kernel  $K(\cdot)$ . The estimator of  $\sigma^2(t)$  was studied by Fan and Yao (1998) using local linear regression.

### 3.2 Estimation of the Correlation Structure Parameter

For each subject  $i$ , denote  $\text{corr}(\boldsymbol{\varepsilon}_i)$  by  $\mathbf{C}(\boldsymbol{\theta}; i)$ , with  $(j, k)$  element  $\rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$ ,  $\mathbf{r}_i = (r_{i1}, r_{i2}, \dots, r_{iJ_i})^T$ , and  $\hat{\mathbf{V}}_i = \text{diag}\{\hat{\sigma}(t_{i1}), \hat{\sigma}(t_{i2}), \dots, \hat{\sigma}(t_{iJ_i})\}$ . We estimate the correlation structure parameter  $\boldsymbol{\theta}$  by the quasi-maximum likelihood method,

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \left\{ -\frac{1}{2} \log |\mathbf{C}(\boldsymbol{\theta}; i)| - \frac{1}{2} \mathbf{r}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{C}(\boldsymbol{\theta}; i)^{-1} \hat{\mathbf{V}}_i^{-1} \mathbf{r}_i \right\}.$$

Denote by  $\zeta(t) \equiv \varepsilon(t)/\sigma(t)$ . For a generic subject  $\mathbb{X}$  with  $J$  observations, the ‘‘standardized’’ random error vector  $\boldsymbol{\zeta} = (\zeta(t_1), \zeta(t_2), \dots, \zeta(t_J))^T$  is assumed to follow an elliptically contoured distribution with a multivariate density function proportional to  $|\mathbf{C}(\boldsymbol{\theta}_0)|^{-1/2} h_0(\boldsymbol{\zeta}^T \mathbf{C}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\zeta})$ , where  $\mathbf{C}(\boldsymbol{\theta}_0)$  is the correlation matrix with its  $(i, j)$  element  $\rho(t_i, t_j, \boldsymbol{\theta}_0)$  for  $1 \leq i, j \leq J$  and  $h_0(\cdot)$  is an arbitrary univariate density function defined on  $[0, \infty)$ . In the next section we show that the QMLE  $\hat{\boldsymbol{\theta}}$  is consistent and also enjoys asymptotic normality when the correlation structure  $\rho(\cdot, \cdot, \boldsymbol{\theta})$  is correctly specified.

When the gaps between some observation times are too close (i.e., below a threshold) for some individuals, the matrix  $\mathbf{C}(\boldsymbol{\theta}; i)$  can be ill-conditioned. In this case we can either delete some of their observations or remove those cases, thereby reducing the influence of those individuals. Under Condition (B), such cases are rare, because individuals have no more than  $(\log n)$  observations.

## 4. SAMPLING PROPERTIES

In this section we study large-sample properties of the estimators presented in Section 3 for the semiparametric covariance structure in our model (1). To derive asymptotic properties, we assume that the data are a random sample collected from the population process  $\{y(t), \mathbf{x}(t)\}$  as described by model (1) with true conditional mean function  $m_0(\cdot)$ , variance function  $\sigma_0^2(\cdot)$ , and correlation structure parameter  $\boldsymbol{\theta}_0$  over a bounded time domain,  $t \in [0, T]$ , for some  $T > 0$ . To ease our presentation, we further assume that observation numbers  $J_i, i = 1, 2, \dots, n$ , are iid and that Condition (B) is satisfied. For each subject  $i$ , given the number of observations  $J_i$ , observation times  $t_{ij}, j = 1, 2, \dots, J_i$ , are iid with density function  $f(t)$ . Large-sample properties of our estimators are stated in Theorems 1–3 in the next section. Technical conditions and proofs are relegated to the Appendix.

### 4.1 Estimator of the Covariance Structure

Let  $\dot{\sigma}_0^2(\cdot)$  and  $\ddot{\sigma}_0^2(\cdot)$  denote the first and second derivatives of the variance function  $\sigma_0^2(\cdot)$ .

*Theorem 1.* Under Conditions (A)–(E), if  $h \propto n^{-1/5}$ , then, as  $n \rightarrow \infty$ ,

$$\sqrt{nh}\{\hat{\sigma}^2(t) - \sigma_0^2(t) - b(t)\} \xrightarrow{\mathcal{L}} N(0, v(t)), \quad (11)$$

where the bias and variance are given by  $b(t) = h^2 \mu_2 [\ddot{\sigma}_0^2(t) + 2\dot{\sigma}_0^2(t) f'(t)/f(t)]/2$  and  $v(t) = \text{var}(\varepsilon^2(t))v_0/(f(t)E(J_i))$ , with  $\mu_2 = \int u^2 K(u) du$  and  $v_0 = \int K^2(u) du$ .

When the correlation structure  $\rho(\cdot, \cdot, \boldsymbol{\theta})$  is correctly specified, the next two theorems establish the consistency and asymptotic normality of the QMLE  $\hat{\boldsymbol{\theta}}$ .

*Theorem 2 (Consistency).* For model (1) with the elliptical density assumption on the random error trajectory, under Conditions (A)–(J) in the Appendix, and with  $h$  specified in Theorem 1, the QMLE  $\hat{\boldsymbol{\theta}}$  is consistent, that is,

$$\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_0, \quad \text{in probability, as } n \rightarrow \infty. \quad (12)$$

Denote the  $d \times d$  Fisher information-like matrix by  $\mathbf{I}(\boldsymbol{\theta}_0)$  whose  $(i, j)$  element is given by

$$E_{\mathbb{X}} \left\{ \frac{1}{2} \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}_0) \mathbf{C}_i(\boldsymbol{\theta}_0) \mathbf{C}^{-1}(\boldsymbol{\theta}_0) \mathbf{C}_j(\boldsymbol{\theta}_0)] \right\},$$

where  $\mathbf{C}(\boldsymbol{\theta}_0)$  is the correlation matrix of a generic subject  $\mathbb{X}$ ,  $\mathbf{C}_i(\boldsymbol{\theta}) = \frac{\partial \mathbf{C}(\boldsymbol{\theta})}{\partial \theta_i}$ ,  $\mathbf{C}^i(\boldsymbol{\theta}) = \frac{\partial \mathbf{C}^{-1}(\boldsymbol{\theta})}{\partial \theta_i} = -\mathbf{C}^{-1}(\boldsymbol{\theta}) \mathbf{C}_i(\boldsymbol{\theta}) \mathbf{C}^{-1}(\boldsymbol{\theta})$ , and the expectation is due to the randomness of the observation times and is taken with respect to the true underlying population distribution of  $\mathbb{X}$ . The subscript ‘‘ $\mathbb{X}$ ’’ in  $E_{\mathbb{X}}$  is dropped whenever there is no confusion. Similarly, derivatives are defined for  $\mathbf{C}(\boldsymbol{\theta}, m)$ , and  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  denotes a  $d \times d$  matrix whose  $(i, j)$  element is given by

$$E_{\mathbb{X}} \left\{ \frac{1}{4} (-\text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}) \mathbf{C}_i(\boldsymbol{\theta})] \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}) \mathbf{C}_j(\boldsymbol{\theta})] + \boldsymbol{\zeta}^T \mathbf{C}^i(\boldsymbol{\theta}) \boldsymbol{\zeta} \boldsymbol{\zeta}^T \mathbf{C}^j(\boldsymbol{\theta}) \boldsymbol{\zeta}) \right\}.$$

*Theorem 3 (Asymptotic normality).* Under the conditions of Theorem 2, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{I}(\boldsymbol{\theta}_0)^{-1}). \quad (13)$$

*Remark 1.* In Theorem 3 the asymptotic variance–covariance matrix of  $\hat{\boldsymbol{\theta}}$  has the sandwich form  $\boldsymbol{\Delta} = \mathbf{I}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{I}(\boldsymbol{\theta}_0)^{-1}$ , which can be estimated by  $\hat{\boldsymbol{\Delta}} = \hat{\mathbf{I}}^{-1} \hat{\boldsymbol{\Sigma}} \hat{\mathbf{I}}^{-1}$ . More explicitly, the  $(i, j)$  element of  $\hat{\mathbf{I}}$  and  $\hat{\boldsymbol{\Sigma}}$  are given by

$$\frac{1}{2n} \sum_{m=1}^n \text{tr}[\mathbf{C}^{-1}(\hat{\boldsymbol{\theta}}; m) \mathbf{C}_i(\hat{\boldsymbol{\theta}}; m) \mathbf{C}^{-1}(\hat{\boldsymbol{\theta}}; m) \mathbf{C}_j(\hat{\boldsymbol{\theta}}; m)]$$

and

$$\frac{1}{4n} \sum_{m=1}^n \left\{ -\text{tr}[\mathbf{C}^{-1}(\hat{\boldsymbol{\theta}}; m) \mathbf{C}_i(\hat{\boldsymbol{\theta}}; m)] \text{tr}[\mathbf{C}^{-1}(\hat{\boldsymbol{\theta}}; m) \mathbf{C}_j(\hat{\boldsymbol{\theta}}; m)] + \hat{\boldsymbol{\zeta}}_m^T \mathbf{C}^i(\hat{\boldsymbol{\theta}}; m) \hat{\boldsymbol{\zeta}}_m \hat{\boldsymbol{\zeta}}_m^T \mathbf{C}^j(\hat{\boldsymbol{\theta}}; m) \hat{\boldsymbol{\zeta}}_m \right\},$$

where  $\hat{\boldsymbol{\zeta}}_m$  is the estimated standardized random errors for the  $m$ th subject as defined in the Appendix.

### 4.2 Verification of Spectrum Condition on Correlation Matrixes

Note that the structural Condition (H) in the Appendix is very common when studying covariance matrixes. In this section we consider several parametric correlation structures, AR(1), ARMA(1, 1), and, more generally, CARMA( $p, q$ ) (continuous-time ARMA of orders  $p$  and  $q, q < p$ ), for which we show that Condition (H) holds.

For AR(1) and ARMA(1, 1), the parametric correlation structure  $\rho(s, t, \boldsymbol{\theta})$  can be parameterized as  $\rho(s, t, \varphi) = \exp(-|s - t|/\varphi)$  and  $\rho(s, t, (\gamma, \varphi)^T) = \gamma \exp(-|s - t|/\varphi)$ , with

$\varphi \geq 0$  and  $0 \leq \gamma \leq 1$ . According to the autocovariance formula (A.2) of Phadke and Wu (1974), the correlation structure of CARMA( $p, q$ ) is a convex combination of  $p$  AR(1) correlation structures, that is,  $\rho(s, t, \theta) = \sum_{i=1}^p \gamma_i \exp(-|s - t|/\varphi_i)$  with  $\varphi_i \geq 0, 0 \leq \gamma_i \leq 1$  and  $\sum_{i=1}^p \gamma_i = 1$ .

*Proposition 1.* Let  $\mathbf{A}$  be a  $K \times K$  matrix with  $(i, j)$  element  $(\mathbf{A})_{ij} = \exp(-a|s_i - s_j|)$  where  $s_1 < s_2 < \dots < s_K$  and  $a > 0$ . The following results then hold:

- a. The  $(i, j)$  element of  $\mathbf{A}^{-1}$  is given by  $(\mathbf{A}^{-1})_{11} = (1 + \coth(a(s_2 - s_1)))/2$ ;  $(\mathbf{A}^{-1})_{ii} = (\coth(a(s_i - s_{i-1})) + \coth(a(s_{i+1} - s_i)))/2$  for  $i = 2, 3, \dots, K - 1$ ;  $(\mathbf{A}^{-1})_{KK} = (1 + \coth(a(s_K - s_{K-1}))/2$ ;  $(\mathbf{A}^{-1})_{ij} = -(\operatorname{csch}(a|s_i - s_j|))/2$  for  $|i - j| = 1$ ; and  $(\mathbf{A}^{-1})_{ij} = 0$  when  $|i - j| > 1$ , where  $\coth(\cdot)$  and  $\operatorname{csch}(\cdot)$  are the hyperbolic cotangent and cosecant functions.
- b. If  $\min_{i=2,3,\dots,K} (s_i - s_{i-1}) = s_0 > 0$ , then the eigenvalues of  $\mathbf{A}^{-1}$  are bounded between  $\delta_0(s_0, a) = \tanh(as_0/2)$  and  $\delta_1(s_0, a) = 2 \coth(as_0)$ , where  $\tanh(\cdot)$  is the hyperbolic tangent function. Moreover, neither  $\delta_0$  and  $\delta_1$  depends on  $K$ .

*Proposition 2.* Consider a generic subject  $\mathbb{X}$ . When  $\min_{1 \leq j \neq k \leq J} |t_j - t_k| \geq t_0 > 0$ , Condition (H) is satisfied for any of the following three cases with some  $\varphi_0 > 0$ :

- a. The AR(1) correlation structure:  $\rho(s, t, \varphi) = \exp(-|s - t|/\varphi)$  with  $\varphi \in [0, \varphi_0]$
- b. The ARMA(1, 1) correlation structure:  $\rho(s, t, (\gamma, \varphi)^T) = \gamma \exp(-|s - t|/\varphi)$  when  $s \neq t$  and 1 otherwise for  $(\gamma, \varphi) \in [0, 1] \times [0, \varphi_0]$
- c. The CARMA( $p, q$ ) correlation structure:  $\rho(s, t, (\gamma_1, \gamma_2, \dots, \gamma_p, \varphi_1, \varphi_2, \dots, \varphi_p)^T) = \sum_{i=1}^p \gamma_i \exp(-|s - t|/\varphi_i)$  with  $q < p, (\gamma_i, \varphi_i) \in [0, 1] \times [0, \varphi_0]$  for  $i = 1, 2, \dots, p$ , and  $\sum_{i=1}^p \gamma_i = 1$

*Remark 2.* In practice, one may face the problem of identifying the order  $(p, q)$  for the CARMA correlation structure. This can be achieved using an Akaike information criterion (AIC)/Bayes information criterion (BIC)-related procedure, because the estimation of the parametric correlation structure is based on maximum likelihood once we have estimated the regression function  $m(\mathbf{x})$  and the variance function  $\sigma^2(t)$ .

### 5. MONTE CARLO STUDY

In this section we study the finite-sample performance of the semiparametric covariance matrix estimator presented in Section 3. While focusing on the varying-coefficient partially linear model in Example 2, we use Example 1 to demonstrate the

efficiency improvement by incorporating the estimated covariance structure for the case of parametric models. A comparison study is provided in Example 3 to illustrate the robustness of our new proposed difference-based estimation scheme.

For all simulation examples, we set  $\mathcal{T} = [0, 13]$ . Each subject has a set of “scheduled” observation times,  $\{0, 1, \dots, 12\}$ , and each scheduled time has a 20% probability of being skipped except the time 0. For each nonskipped scheduled time, the corresponding actual observation time is obtained by adding a standard uniform random variable on  $[0, 1]$ . The true variance function is chosen to be  $\sigma^2(t) = .5 \exp(t/12)$ . Either a true AR(1) (Example 1) or a true ARMA(1, 1) (Examples 2 and 3) correlation structure is assumed; that is,  $\operatorname{corr}(\epsilon(s), \epsilon(t)) = \gamma \rho^{|s-t|}$  when  $s \neq t$  and 1 otherwise. In AR(1),  $\gamma = 1$ . For a particular subject, given the number of observations  $J$  and the observation times  $t_1, t_2, \dots, t_J$ , the standardized random error vector  $(\zeta(t_1), \zeta(t_2), \dots, \zeta(t_J))^T$  has a marginal normal (N) or double-exponential (DE) distribution with mean 0, variance 1, and correlation matrix  $\mathbf{C}(\gamma, \rho; t_1, t_2, \dots, t_J)$  with  $(i, j)$  element  $\gamma \rho^{|t_i - t_j|}$  when  $i \neq j$  and 1 otherwise.

Unless specified otherwise, our simulation result is based on 1,000 independent repetitions, and each training sample is of size 200. The Epanechnikov kernel is used whenever a kernel is needed. For an estimator of a functional component, say  $\sigma^2(\cdot)$ , we report its root average squared error (RASE), which is defined as  $RASE(\hat{\sigma}^2) = \sqrt{\sum_{k=1}^K (\hat{\sigma}^2(t_k) - \sigma^2(t_k))^2 / K}$ , where  $\{t_k : k = 1, 2, \dots, K\}$  form a uniform grid. The number of grid points,  $K$ , is set to 200 in our simulations. After tuning, these necessary smoothing parameters are fixed and used for each independent repetition. In Tables 1–5, the first and the second column blocks designate the marginal distribution type of the “standardized” random error and the correlation structure parameter.

*Example 1* (Parametric model). In this example the covariate is three-dimensional, that is,  $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ , with standard normal marginal distribution and correlation  $\operatorname{corr}(x_i, x_j) = .5^{|i-j|}$ . The true regression parameter vector is set to be  $\beta_0 = (2, 1.5, 3)^T$ . An AR(1) correlation structure is used with three different parameter values,  $\rho = .3, .6, .9$ . After tuning, the best smoothing bandwidth,  $h = 2.53$ , is used for estimating  $\sigma^2(t)$ .

Table 1 reports the mean and standard deviation (in parentheses) of the OLS estimator of  $\beta$  in the third column. The fourth and fifth columns correspond to the kernel estimator of  $\sigma^2(\cdot)$  and the QMLE of  $\rho$ . The last column gives the asymptotic standard errors (ASEs) of QMLE  $\hat{\rho}$  using the formula (13)

Table 1. Finite-sample performance for estimating  $\beta, \sigma^2(\cdot)$  and  $\rho$

Noise	$\rho$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$RASE(\hat{\sigma}^2)$	$\hat{\rho}$	ASE
N	.30	1.9993(.0236)	1.5002(.0266)	2.9991(.0237)	.0493(.0456)	.2997(.0219)	.0263
	.60	1.9995(.0237)	1.5002(.0264)	2.9993(.0235)	.0565(.0509)	.5988(.0172)	.0165
	.90	1.9996(.0238)	1.4998(.0259)	3.0000(.0231)	.0702(.0605)	.8981(.0077)	.0041
DE	.30	1.9992(.0244)	1.5001(.0260)	2.9999(.0235)	.0692(.0622)	.3000(.0227)	.0241
	.60	1.9993(.0246)	1.5002(.0257)	2.9997(.0235)	.0732(.0656)	.5992(.0173)	.0172
	.90	1.9995(.0246)	1.5003(.0258)	2.9994(.0241)	.0797(.0704)	.8981(.0075)	.0050

Table 2. Finite-sample performance of WLS estimator of  $\beta$  using estimated covariance structure

Noise	$\rho$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
N	.30	1.9994 <sub>(.0190)</sub>	1.4996 <sub>(.0217)</sub>	2.9993 <sub>(.0199)</sub>
	.60	1.9995 <sub>(.0143)</sub>	1.4996 <sub>(.0164)</sub>	2.9995 <sub>(.0151)</sub>
	.90	1.9998 <sub>(.0069)</sub>	1.4998 <sub>(.0079)</sub>	2.9998 <sub>(.0073)</sub>
DE	.30	1.9992 <sub>(.0192)</sub>	1.5005 <sub>(.0220)</sub>	3.0004 <sub>(.0189)</sub>
	.60	1.9994 <sub>(.0144)</sub>	1.5004 <sub>(.0167)</sub>	3.0004 <sub>(.0143)</sub>
	.90	1.9997 <sub>(.0069)</sub>	1.5002 <sub>(.0080)</sub>	3.0002 <sub>(.0069)</sub>

in Theorem 3, with estimated matrix  $\mathbf{I}(\theta_0)$  and  $\Sigma(\theta_0)$  based on 1,000 simulations.

The result indicates that our semiparametric covariance matrix estimation scheme works effectively. The RASE of  $\hat{\sigma}^2(\cdot)$  and the QMLE  $\hat{\rho}$  are close to 0, and the corresponding true  $\rho$ , for each case. This is consistent with the theoretical results. The standard deviation of QMLE  $\hat{\rho}$  is close to the corresponding ASE except for the high correlation case with  $\rho = .9$ . This can be explained by noting the proof of our asymptotic normality result. When proving Theorem 3, we need the upper bound of the eigenvalue of the inverse correlation matrix to control the effect of the estimation error in the previous estimation steps. For the AR(1) model, this upper bound is guaranteed by Proposition 1. But  $\rho = .9$  implies that in Proposition 1,  $\varphi = -1/\log(.9) = 9.49$ , which is relatively large. The problem disappears when the sample size is large. A similar phenomenon is observed for the ARMA(1, 1) correlation structure in Example 2.

For the parametric model  $m(\mathbf{x}) = \mathbf{x}^T \beta$ , we can incorporate the estimated covariance structure to estimate  $\beta$  using weighted least squares (WLS) regression, whose performance is reported in Table 2. Comparing the OLS and WLS estimators shows that the standard deviation of the estimator of  $\beta$  is reduced significantly by incorporating the estimated covariance structure, especially in the case of high correlation ( $\rho = .9$ ).

*Example 2* (Semiparametric varying-coefficient partially linear model). In this example, data are generated from model (2), with ARMA(1, 1) correlation structure on  $\varepsilon(t)$ . In this case  $\mathbf{x}$  is four-dimensional with  $\mathbf{x}_1 = (x_1, x_2)^T \in \mathbb{R}^2$  and  $\mathbf{x}_2 = (x_3, x_4)^T \in \mathbb{R}^2$ . We set the first component of  $\mathbf{x}_1$  to be constant 1 to include the intercept term, that is,  $x_1(t) \equiv 1$ . For any given time  $t$ ,  $x_2(t)$  and  $x_3(t)$  are generated jointly such that they have standard normal marginal distribution and correlation .5, and  $x_4(t)$  is Bernoulli-distributed with success probability .5, independent of  $x_2(t)$  and  $x_3(t)$ . The regression coefficients are

specified as

$$\alpha_1(t) = \sqrt{t/12}, \quad \alpha_2(t) = \sin(2\pi t/12),$$

and

$$\beta = (1, 2)^T.$$

After tuning, we select the best smoothing bandwidths,  $b = 1.25$  and  $h = 2.53$ , for estimating  $\alpha(\cdot)$  and  $\sigma^2(\cdot)$ .

*Performance for estimating  $\beta$ ,  $\alpha(\cdot)$ , and  $\sigma^2(\cdot)$ .* As  $\mathbf{x}_1 \in \mathbb{R}^2$ , DBE uses three neighboring (sorted) observations with weights given by (5) with positive sign. Mean and standard deviation (in parentheses) of our DBE of  $\beta_1$  and  $\beta_2$  over 1,000 repetitions are reported in the third column of Table 3 for three pairs of ARMA(1, 1) correlation structure parameters,  $\theta^T = (\gamma, \rho) = (.85, .30), (.85, .60), (.85, .90)$ . Table 3 also displays the mean and standard deviation (in parentheses) of the RASEs of  $\hat{\alpha}(\cdot)$  and  $\hat{\sigma}^2(\cdot)$  in the fourth and the fifth columns. From the table, we can see that the DBE of  $\beta$ , the local constant regression estimator of  $\alpha(\cdot)$ , and the local kernel smoothing estimator of  $\sigma^2(\cdot)$  are very precise in the finite-sample case.

*Performance of the QMLE of  $\theta$ .* Our QMLE  $\hat{\theta}$  is based on the estimation of other components in our model. Table 3 indicates that the estimators  $\hat{\beta}$ ,  $\hat{\alpha}(\cdot)$ , and  $\hat{\sigma}^2(\cdot)$  perform very well. Thus we expect similarly good performance for the QMLE  $\hat{\theta}$ . Table 4 gives its simulation results. The third column reports the mean of  $\hat{\gamma}$  and  $\hat{\rho}$ , with their standard deviations in parentheses and their correlations in square brackets, and the last column gives asymptotic standard errors and correlation of  $\hat{\gamma}$  and  $\hat{\rho}$  using the formula (13) in Theorem 3 with estimated matrix  $\mathbf{I}(\theta)$  and  $\Sigma(\theta)$  based on 1,000 simulations.

From Table 4, we can see that the estimated parameters in the correlation structure are very close to the corresponding true parameters for each case. This agrees with our consistency result. Next, we check the accuracy of the estimated standard errors. For the cases with true correlation structure parameters  $\theta^T = (\gamma, \rho) = (.85, .3)$  or  $(.85, .6)$ , the standard deviations and correlations of our QMLE  $\hat{\theta}$  are very close to the asymptotic standard errors and correlations (in the last column) using our asymptotic normality formulas. Similar to the previous example, we observe that the sample correlations deviate greatly from their corresponding simulated correlations using formula (13) for the two cases of higher correlation  $(\gamma, \rho) = (.85, .9)$  and the same explanation applies here.

*Prediction.* As mentioned in Section 1, estimating the covariance structure can improve the prediction accuracy for a particular trajectory. In this example we study the improvement

Table 3. Finite-sample performance of estimating  $\beta$ ,  $\alpha(\cdot)$ , and  $\sigma^2(\cdot)$

Noise	$(\gamma, \rho)$	$\hat{\beta}_1$	$\hat{\beta}_2$	RASE( $\hat{\alpha}_1$ )	RASE( $\hat{\alpha}_2$ )	RASE( $\hat{\sigma}^2$ )
N	(.85, .30)	1.0003 <sub>(.0322)</sub>	1.9993 <sub>(.0582)</sub>	.0695 <sub>(.0186)</sub>	.0814 <sub>(.0157)</sub>	.0618 <sub>(.0227)</sub>
	(.85, .60)	1.0006 <sub>(.0334)</sub>	2.0030 <sub>(.0565)</sub>	.0712 <sub>(.0218)</sub>	.0817 <sub>(.0155)</sub>	.0673 <sub>(.0245)</sub>
	(.85, .90)	1.0003 <sub>(.0347)</sub>	1.9985 <sub>(.0571)</sub>	.0718 <sub>(.0280)</sub>	.0827 <sub>(.0157)</sub>	.0786 <sub>(.0335)</sub>
DE	(.85, .30)	.9994 <sub>(.0339)</sub>	1.9993 <sub>(.0584)</sub>	.0667 <sub>(.0179)</sub>	.0821 <sub>(.0152)</sub>	.0884 <sub>(.0330)</sub>
	(.85, .60)	.9995 <sub>(.0337)</sub>	1.9987 <sub>(.0576)</sub>	.0694 <sub>(.0205)</sub>	.0822 <sub>(.0153)</sub>	.0919 <sub>(.0356)</sub>
	(.85, .90)	.9998 <sub>(.0331)</sub>	2.0001 <sub>(.0581)</sub>	.0718 <sub>(.0289)</sub>	.0826 <sub>(.0155)</sub>	.0950 <sub>(.0393)</sub>

Table 4. Finite-sample performance of QMLE of  $\theta$

Noise	$(\gamma, \rho)$	$(\hat{\gamma}, \hat{\rho})$			Asymptotic covariance
N	(.85, .30)	.8458 <sub>(.0677)</sub>	.2995 <sub>(.0439)</sub>	[-.7749]	(.0636), (.0407) [-.7547]
	(.85, .60)	.8428 <sub>(.0327)</sub>	.6000 <sub>(.0301)</sub>	[-.6432]	(.0303), (.0272) [-.7281]
	(.85, .90)	.8427 <sub>(.0153)</sub>	.8990 <sub>(.0106)</sub>	[.0351]	(.0112), (.0083) [-.5898]
DE	(.85, .30)	.8479 <sub>(.0772)</sub>	.2996 <sub>(.0462)</sub>	[-.7734]	(.0741), (.0442) [-.7492]
	(.85, .60)	.8441 <sub>(.0353)</sub>	.5983 <sub>(.0315)</sub>	[-.6638]	(.0344), (.0304) [-.6545]
	(.85, .90)	.8421 <sub>(.0155)</sub>	.8984 <sub>(.0113)</sub>	[-.0011]	(.0129), (.0094) [-.3307]

of prediction after estimating the covariance structure. For each case, we generate an independent prediction data set of size 400 exactly as the training data set for estimation is generated. For each observation of these 400 subjects in the prediction data set, an independent Bernoulli random variable with success probability .5 is generated. If it is 0, then the response of this observation is treated as “missing” and must be predicted; if it is 1, then this observation is fully observed and used to predict the “missing” observations. Prediction is made using the prediction formula given by Fan et al. (2007, sec. 5.3).

Table 5 reports the means and standard deviations (in parentheses) over 1,000 repetitions of the sum of squared prediction errors (SSPEs) for five different types of prediction in the case of normal standardized random errors. A similar result can be found for the case with DE standardized random errors. Prediction 1 corresponds to the oracle [i.e., using true  $\beta$ ,  $\alpha(\cdot)$ ,  $\sigma^2(\cdot)$ , and  $\theta$ ]; prediction 2 uses estimated  $\hat{\beta}$ ,  $\hat{\alpha}(\cdot)$ ,  $\hat{\sigma}^2(\cdot)$  and true  $\theta$ ; prediction 3 uses all estimated parameters [i.e.,  $\hat{\beta}$ ,  $\hat{\alpha}(\cdot)$ ,  $\hat{\sigma}^2(\cdot)$ , and  $\hat{\theta}$ ]; and predictions 4 and 5 correspond to predictions that ignore the covariance structure, based on true  $\beta$  and  $\alpha(\cdot)$  and estimated  $\hat{\beta}$  and  $\hat{\alpha}(\cdot)$ . The numbers in parentheses are their corresponding standard deviations. The last column of Table 5 gives the number of points where the prediction is made for each case, namely the number of “missing” observations in the independent prediction data set.

Note first that in Table 5, the difference between SSPE(prediction 5) and SSPE(prediction 1) is much larger than that between SSPE(prediction 3) and SSPE(prediction 1). This implies that, compared with prediction ignoring the covariance structure, incorporating the estimated covariance structure reduces the prediction error significantly. Furthermore, we see that the difference between SSPE(prediction 3) and SSPE(prediction 1) is much larger than that between SSPE(prediction 3) and SSPE(prediction 2). This indicates that using the true correlation structure parameter  $\theta$  and the estimated  $\hat{\theta}$  does not make much difference after the estimated covariance structure is included and demonstrates the accuracy of  $\hat{\theta}$  for prediction.

*Example 3* (Comparison with the method of Fan et al. (2007)). For estimating the varying-coefficient partially linear model, the major difference between our method and the method of Fan et al. (2007) is that we estimate  $\beta$  using the difference-based technique, plug it into model (2), and estimate  $\alpha(\cdot)$  through the local constant regression, whereas they estimate  $\beta$  and  $\alpha(\cdot)$  simultaneously using the local linear regression and profile-likelihood techniques. Next, we use another simulation to study the impact of rough  $\alpha(\cdot)$  on these two different estimation schemes. This occurs when there is some structure break or technological innovation that cause the change point in time.

In this comparison study, we set  $\alpha_1(t) = 2$  and  $\alpha_2(t) = 4/(1 + \exp(-40(t - 7))) - 2$ . Note that  $\alpha_2(\cdot)$  is a scaled sigmoid function and is close to a jump function, as shown by the dotted line in Figure 1(d). For each  $t$ , the regressor  $x_1(t)$  takes constant 8, and  $x_2(t)$  and  $x_3(t)$  are jointly simulated from a bivariate normal distribution, namely

$$\begin{bmatrix} x_2(t) \\ x_3(t) \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \right),$$

where  $a_1 = 64$ ,  $a_2 = a_3 = .95 \times \text{sign}(7 - t) \times 8$ , and  $a_4 = 1$ . All other components in our model are simulated the same way as in the previous example. The sample size is  $n = 100$ ; the true ARMA(1, 1) correlation structure parameters are  $\gamma = .85$  and  $\rho = .6$ . After tuning, the profiling method of Fan et al. (2007) chooses the best smoothing bandwidth pair  $b = .8$ ,  $h = 3.80$ , and our new proposed estimation scheme selects  $b = .4$ ,  $h = .75$  for estimation  $\alpha(\cdot)$  and  $\sigma^2(\cdot)$ . A boxplot of the estimated correlation structure parameters over these 100 samples for each method is shown in Figure 1(a).

From the boxplot in Figure 1, we see that the estimator of Fan et al. (2007) is far off the true correlation structure parameters, whereas our method still performs very well. The underlying reason for this is that in the method of Fan et al. (2007), smoothing the rough  $\alpha(\cdot)$  causes a large bias in the profile-likelihood estimator of  $\beta$  because of the strong correlation between  $x_2(\cdot)$  and  $x_3(t)$ . To the profile-likelihood estimators of  $\beta$  for different smoothing bandwidths are depicted in Figure 2, showing that

Table 5. Finite-sample performance of prediction

Noise	$(\gamma, \rho)$	Prediction 1	Prediction 2	Prediction 3	Prediction 4	Prediction 5	Number of predictions
N	(.85, .30)	1,784.06	1,799.02 <sub>(10.28)</sub>	1,799.77 <sub>(10.48)</sub>	1,884.51	1,899.89 <sub>(10.22)</sub>	2,124
	(.85, .60)	1,327.10	1,352.88 <sub>(9.65)</sub>	1,353.64 <sub>(9.82)</sub>	1,766.82	1,794.50 <sub>(10.32)</sub>	2,095
	(.85, .90)	754.34	788.16 <sub>(11.33)</sub>	788.27 <sub>(11.32)</sub>	1,891.00	1,917.81 <sub>(19.54)</sub>	2,068

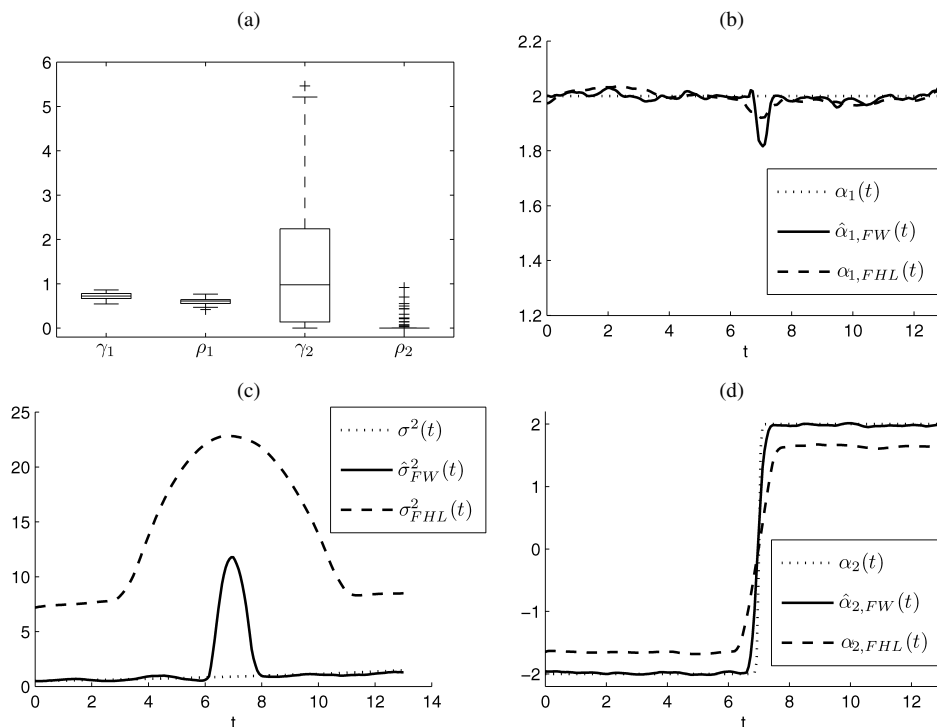


Figure 1. Example 3. (a) The boxplot of our estimator ( $\gamma_1$  and  $\rho_1$ ) and Fan et al. (2007)'s estimator ( $\gamma_2$  and  $\rho_2$ ) of the correlation structure parameter. (b), (c), and (d) Plots of the true ( $\cdots$ ), our new estimate ( $\text{—}$ ), and Fan et al. (2007)'s estimate ( $\text{- - -}$ ) of  $\alpha_1(\cdot)$ ,  $\sigma^2(\cdot)$ , and  $\alpha_2(\cdot)$ , for a typical sample.

the profile-likelihood method does not provide a consistent estimator of  $\beta_1$ . But our DBEs of  $\beta$  are  $\hat{\beta}_1 = .9163$  (.1651) and  $\hat{\beta}_2 = 2.0037$  (.0828), very close to their corresponding true values. This means that our difference-based method still performs very well. Thus our DBE of  $\beta$  is more robust to the smoothness assumption of  $\alpha(\cdot)$ . For a random sample, we plot the estimated  $\alpha_1(\cdot)$ ,  $\alpha_2(\cdot)$ , and  $\sigma^2(\cdot)$  using two different methods in Figure 1(b)–(d). The RASE for estimating  $\alpha_1(\cdot)$  and  $\alpha_2(\cdot)$  is reported in Table 6. We can see that our new estimation scheme significantly improves the estimation of the rough component  $\alpha_2(\cdot)$ ; however, there is no improvement on the estimation of

$\alpha_1(\cdot)$ . Note that here we use the same bandwidth to estimate  $\alpha_1(\cdot)$  and  $\alpha_2(\cdot)$ . The performance of our method can be improved by allowing different bandwidths for different components of  $\alpha(\cdot)$ , as studied by Fan and Zhang (1999).

This comparison study reinforces our use of the profiling technique in the case of a rough varying-regression coefficient  $\alpha(\cdot)$  in the varying-coefficient, partially linear model. But while handling real data, we never know the a priori smoothness of  $\alpha(\cdot)$ . In our new estimation scheme, we can do a visual check or even apply some advanced technique to diagnose the smoothness of  $\alpha(\cdot)$  after getting the DBE  $\hat{\beta}$ .

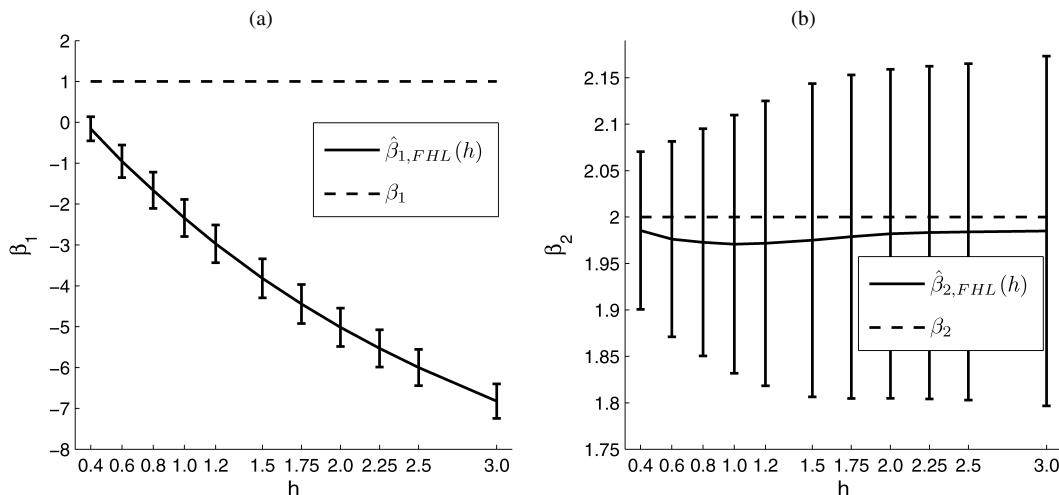


Figure 2. The solid curves (average estimate) with error bars (one sample standard deviation) plot the profile-likelihood estimators of the regression coefficient  $\beta_1$  (a) and  $\beta_2$  (b) for different smoothing bandwidth  $h$ . The dashed curves denote their corresponding true coefficients.

Table 6. RASE for estimating  $\alpha_1(\cdot)$  and  $\alpha_2(\cdot)$  using our method (FW) and the method of Fan et al. (2007) (FHL)

Noise	FW	FHL
$\alpha_1(\cdot)$	.0381(.0164)	.0281(.0090)
$\alpha_2(\cdot)$	.1946(.0277)	.4525(.0403)

6. APPLICATION TO THE PROGESTERONE DATA

In this section we apply the proposed methods to the longitudinal progesterone data. For the  $i$ th subject, let  $x_i$  and  $z_i$  denote age and body mass index, both of which are standardized to have mean 0 and standard deviation 1. We consider the response to be the difference between the  $j$ th log-transformed progesterone level measured at standardized day  $t_{ij}$  and the individual's average log-transformed progesterone level. We consider the semiparametric model

$$y_{ij} = \beta_1 x_i + \beta_2 z_i + f(t_{ij}) + \varepsilon. \tag{14}$$

Note that  $f(t)$  is the sole varying-coefficient term. After sorting, our DBE is based on two neighboring observations with weights  $(1/\sqrt{2}, -1/\sqrt{2})$ , because there is only one varying-coefficient term,  $f(t_{ij})$ . The DBE of (14) gives estimates  $\hat{\beta}_1 = .0306$  and  $\hat{\beta}_2 = .0195$ . The leave-one-subject-out cross-validation procedure suggested by Rice and Silverman (1991) is used to select the bandwidth for estimating  $f(\cdot)$  using local constant regression. Figure 3(a) depicts the cross-validation score function, defined as the sum of residual squares, and suggests the optimal bandwidth 1.9349. Figure 3(b) plots the corresponding estimate  $\hat{f}(\cdot)$  with a pointwise 95% confidence interval. The plug-in bandwidth selector (Ruppert, Sheather, and Wand 1995) is implemented to choose the smoothing bandwidth for the one-dimensional kernel regression of the variance function  $\sigma^2(\cdot)$ ; it selects the bandwidth 2.5864. Figure 3(c) shows the resulting estimate of the variance function.

We consider an ARMA(1, 1) correlation structure  $\text{corr}(\varepsilon(s), \varepsilon(t)) = \gamma\rho^{|s-t|}$  if  $s \neq t$  and 1 otherwise. The QMLEs of  $\gamma$  and  $\rho$  are .6900 (.3662) and .6452 (.1319). The numbers in parentheses are the corresponding standard errors, obtained using (13) based on estimated  $\gamma$ ,  $\rho$ , and  $\varepsilon(t_{ij})$ . This indicates a strong correlation structure.

We next consider testing the hypothesis  $H_0: \gamma = 1$  versus  $H_1: \gamma < 1$ , that is, whether the correlation structure is AR(1). The  $p$  value for this hypothesis test exceeds .10, and as a result,  $H_0$  cannot be rejected.

Because the null hypothesis is not rejected, an AR(1) correlation structure,  $\text{corr}(\varepsilon(s), \varepsilon(t)) = \rho^{|s-t|}$ , is applied on these data. The QMLE  $\hat{\rho}$  is .5049 with standard error .0831.

By incorporating the estimate covariance structure in DBE, WLS regression gives new estimates of  $\beta$ ,  $\hat{\beta}_1 = -.0059$  and  $\hat{\beta}_2 = -.0074$ . The corresponding new plots of Figure 3 are very similar and thus are not reproduced here.

It is straightforward to understand how the estimated regression function and the estimated variance function affect the pointwise prediction as shown by Fan et al. (2007). However, it is not easy to quantify the sensitivity of pointwise prediction with respect to the estimated correlation structure parameter. We studied this sensitivity using two randomly selected subjects. To provide a visual quantification of this sensitivity, the pointwise prediction and 95% predictive interval using the estimated AR(1) correlation structure parameter  $\hat{\rho}$  with a perturbation of one standard error are shown in the left column panels of Figure 4 for one subject and in the right column panels for the other subject, with the same prediction formula of Fan et al. (2007) and the same estimated  $\hat{\beta}$ ,  $\hat{\alpha}(\cdot)$ , and  $\hat{\sigma}^2(\cdot)$  used for both. These figures show not much change in the prediction for different correlation structure parameters, suggesting that the pointwise prediction is not sensitive to the correlation structure parameter.

APPENDIX: PROOFS

First, the following technical conditions are imposed:

- (A) On the domain  $[0, T]$ , the density function  $f(\cdot)$  is Lipschitz-continuous and bounded away from 0. The kernel function  $K(\cdot)$  is a symmetric density function with a compact support.
- (B) The moment-generating function of  $J_i$  is finite in some neighborhood of the origin.
- (C) The estimator of the mean function,  $\hat{m}(\cdot)$ , is consistent with the polynomial convergence rate, that is,  $\exists \tau > 1/4$  such that  $m_0(\mathbf{x}) - \hat{m}(\mathbf{x}) = O_p(n^{-\tau})$  uniformly in  $\mathbf{x}$ .
- (D) It holds that  $\sup |\hat{m}(\mathbf{x}) - \hat{m}_{-i}(\mathbf{x})| = O_p(n^{-\varsigma})$  uniformly in  $\mathbf{x}$  for some  $\varsigma > 2/5$ , where  $m_{-i}(\cdot)$  is the leave-one-subject-out estimation of the conditional mean function by excluding the  $i$ th subject.

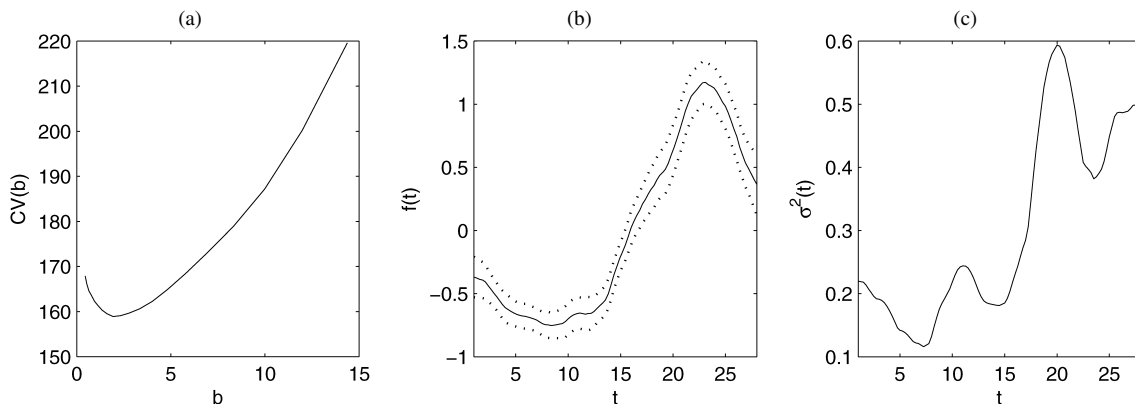


Figure 3. (a) The leave-one-subject-out cross-validation score against the bandwidth. (b) The estimated varying-coefficient function  $f(\cdot)$  with a pointwise 95% confidence interval ( $\cdots$ ). (c) The estimated variance function  $\sigma^2(\cdot)$ .

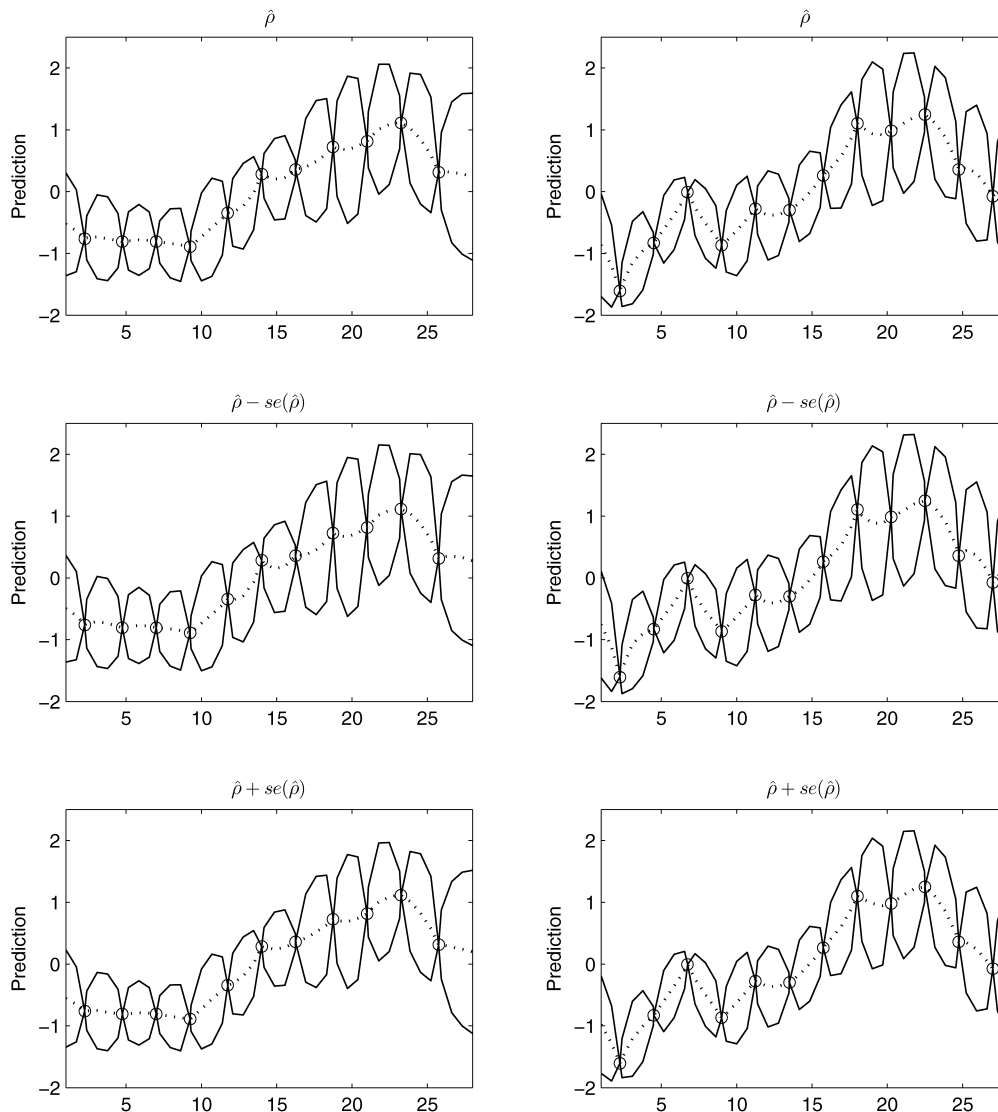


Figure 4. The observed values ( $\circ$ ), the corresponding pointwise predictions ( $\cdots$ ), and 95% predictive intervals ( $—$ ) for two randomly selected subjects, one in the left three panels and the other in the right three panels. For each column, the top, middle, and bottom panels correspond to the prediction using  $\hat{\rho}$ ,  $\hat{\rho} - se(\hat{\rho})$ , and  $\hat{\rho} + se(\hat{\rho})$ .

- (E)  $\sigma_0^2(\cdot)$  has a continuous second derivative and is bounded away from 0 in its domain,  $E\varepsilon(t)^4 < \infty$ .
- (F) The true parameter of the correlation structure,  $\theta_0$ , lies in the interior of a compact set  $\Theta$ .
- (G) For any  $\theta$ ,  $\frac{\partial}{\partial \theta_i} \rho(t, s, \theta)$  and  $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho(t, s, \theta)$  are bounded bivariate functions of  $t$  and  $s$ .
- (H) For any  $\theta \in \Theta$ , it holds with probability 1 that the eigenvalues of the correlation matrix,  $\mathbf{C}(\theta)$ , of a generic subject  $\mathbb{X}$  are between  $\varrho_0$  and  $\varrho_1$ , where  $0 < \varrho_0 < \varrho_1 < \infty$ .
- (I) Assume that  $E \log(g_0(\boldsymbol{\varepsilon}; \mathbf{C}(\theta_0))/f(\boldsymbol{\varepsilon}; \mathbf{C}(\theta)))$  exists, where the expectation is taken with respect to the true elliptical density  $g_0(\boldsymbol{\varepsilon}; \mathbf{C}(\theta_0))$  of the  $\boldsymbol{\varepsilon}$  [i.e.,  $\boldsymbol{\varepsilon} \sim g_0(\boldsymbol{\varepsilon}; \mathbf{C}(\theta_0)) \propto |\mathbf{C}(\theta_0)|^{-1/2} h_0(\boldsymbol{\varepsilon}^T \mathbf{C}(\theta_0)^{-1} \boldsymbol{\varepsilon})$ ], and  $f(\boldsymbol{\varepsilon}; \mathbf{C}(\theta))$  corresponds to the density used in the QMLE by treating  $\boldsymbol{\varepsilon}$  as normally distributed.
- (J) The correlation structure is identifiable, that is,  $\rho(s, t, \theta) \neq \rho(s, t, \theta)$  for any  $\theta \neq \theta_0$  when  $s \neq t$ .

Remark A.1. Technical Condition (C) seems strong. But once the form of the conditional mean function is available, it can be relaxed.

For parametric model  $m(\mathbf{x}(t)) = \mathbf{x}(t)^T \boldsymbol{\beta}$ , Condition (C) can be replaced by  $\sup_{t \in [0, T]} E \|\mathbf{x}(t)\| < \infty$ , and the estimator of  $\boldsymbol{\beta}$  is consistent with a polynomial rate  $O_p(n^{-\tau})$  for any  $\tau > 1/4$ . For the varying-coefficient model  $m(\mathbf{x}(t)) = \mathbf{x}(t)^T \boldsymbol{\alpha}(t)$ , Condition (C) can be replaced by  $\sup_{t \in [0, T]} E \|\mathbf{x}(t)\| < \infty$  and the estimator of  $\boldsymbol{\alpha}(t)$  is consistent with a polynomial rate  $O_p(n^{-\tau})$  for any  $\tau > 1/4$  uniformly in  $t$ ; for nonparametric model  $m(\mathbf{x}) = m(\mathbf{x})$  without any constraint, it is sufficient to assume that  $m(\cdot)$  can be consistently estimated at a polynomial rate  $O_p(n^{-\tau})$  for  $\tau > 1/4$  uniformly in the domain of  $m(\cdot)$ . These are reasonable assumptions for the corresponding models of the conditional mean function. A similar argument applies to Condition (D).

Proof of Theorem 1

Note that Condition (B) implies  $\max_{i=1}^n J_i = O_p(\log n)$ , and it is unlikely for each individual to have two observations in the same neighborhood  $[t - h, t + h]$ . Thus in what follows, the  $\varepsilon_i(t_{ij})$ s can be treated as independent.

Define  $e_{ij} = \hat{m}(\mathbf{x}_i(t_{ij})) - m(\mathbf{x}_i(t_{ij}))$ . Noting that  $r_{ij}^2 = \varepsilon_i^2(t_{ij}) - 2\varepsilon_i(t_{ij})e_{ij} + e_{ij}^2$ , we can accordingly decompose  $\hat{\sigma}^2(t)$  as

$$\begin{aligned} \hat{\sigma}^2(t) &= \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} \varepsilon_i^2(t_{ij}) K_h(t_{ij} - t)}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t)} \\ &\quad - 2 \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} \varepsilon_i(t_{ij}) e_{ij} K_h(t_{ij} - t)}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t)} \\ &\quad + \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} e_{ij}^2 K_h(t_{ij} - t)}{\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t)} \\ &= A_1 + A_2 + A_3. \end{aligned}$$

Using Condition (C), we get  $\sqrt{nh}A_3 = O_p(\sqrt{nh}n^{-2\tau}) = o_p(1)$  when  $\tau > 1/5$  and  $h \propto n^{-1/5}$ . Note that  $\text{var}(A_2) \leq n^{-2\tau} \times \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} \sigma^2(t_{ij}) K_h^2(t_{ij} - t)}{(\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t))^2} \leq n^{-2\tau} (\sup \sigma^2(t)) / (nh)$  and

$$\begin{aligned} |E(A_2)| &= \left| E \left( \left( \sum_{i=1}^n \sum_{j=1}^{J_i} \varepsilon_i(t_{ij}) (\hat{m}(\mathbf{x}_i(t_{ij})) - \hat{m}_{-i}(\mathbf{x}_i(t_{ij}))) \right) \right. \right. \\ &\quad \left. \left. \times K_h(t_{ij} - t) \right) / \left( \sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t) \right) \right| \\ &\leq E \frac{\sum_{i=1}^n \sum_{j=1}^{J_i} |\varepsilon_i(t_{ij})| |K_h(t_{ij} - t)|}{|\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij} - t)|} \\ &\quad \times \sup |\hat{m}(\mathbf{x}) - \hat{m}_{-i}(\mathbf{x})|, \end{aligned}$$

where  $\hat{m}_{-i}(\cdot)$  is the leave-one-subject-out estimator of  $m(\cdot)$  by excluding the  $i$ th subject. By Condition (D) and the mean-variance decomposition, we get

$$\sqrt{nh}A_2 = O_p(\sqrt{nh}[n^{-\tau}/\sqrt{nh} + O_p(n^{-\tau})]) = o_p(1).$$

It remains to show that the main term is asymptotically normal. Applying the standard techniques to derive asymptotic bias and variance for a kernel regression estimator of type  $A_1$ , it follows from  $E\varepsilon_i(t)^2 = \sigma_0^2(t)$  that

$$\sqrt{nh}[A_1 - \sigma_0^2(t) - b(t)] \xrightarrow{L} N(0, v(t)).$$

Using Slutsky's theorem, we have that

$$\sqrt{nh}[\hat{\sigma}^2(t) - \sigma_0^2(t) - b(t)] \xrightarrow{L} N(0, v(t)).$$

This completes the proof of Theorem 1.

**Proof of Theorem 2**

We need the following observations and results to prove Theorem 2. Note that  $\zeta(t) = \varepsilon(t)/\sigma(t)$ . Then  $E\zeta(t) = 0$ ,  $\text{var}(\zeta(t)) = 1$ , and  $\text{cov}(\zeta_i(t_{ij}), \zeta_i(t_{ik})) = \rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$ . After plugging the estimators  $\hat{m}(\cdot)$  and  $\hat{\sigma}^2(\cdot)$ , we obtain the corresponding estimators of the standardized random errors  $\zeta_i(t_{ij})$ 's and denote them by  $\hat{\zeta}_i(t_{ij})$ 's. They then can be decomposed as follows:

$$\begin{aligned} \hat{\zeta}_i(t_{ij}) &= \frac{y_{ij} - \hat{m}(\mathbf{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} \\ &= \frac{m(\mathbf{x}_i(t_{ij})) - \hat{m}(\mathbf{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} + \frac{\sigma_0(t_{ij})\zeta_i(t_{ij})}{\hat{\sigma}(t_{ij})} \\ &= \zeta_i(t_{ij}) + \frac{m(\mathbf{x}_i(t_{ij})) - \hat{m}(\mathbf{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} \end{aligned}$$

$$+ \frac{[\sigma_0(t_{ij}) - \hat{\sigma}(t_{ij})]\zeta_i(t_{ij})}{\hat{\sigma}(t_{ij})}. \tag{A.1}$$

For each subject  $i$ , by vectorizing the residuals, we denote  $\boldsymbol{\zeta}_i = (\zeta_i(t_{i1}), \zeta_i(t_{i2}), \dots, \zeta_i(t_{iJ_i}))^T$  and its corresponding estimator  $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_i(t_{i1}), \hat{\zeta}_i(t_{i2}), \dots, \hat{\zeta}_i(t_{iJ_i}))^T$ . Note that while proving Theorem 1, we also can use a technique related to that of Fan and Huang (2005) to show that  $\hat{\sigma}^2(t) - \sigma_0^2(t)$  converges to 0 uniformly in  $t$ . Based on Conditions (C) and the result of Theorem 1, we have that  $\boldsymbol{\zeta}_i - \hat{\boldsymbol{\zeta}}_i$  converges in probability to 0 at a polynomial rate both elementwise and in the 2-norm due to Condition (E) and the fact that  $\max_i J_i = O_p(\log n)$ .

The QMLE  $\hat{\boldsymbol{\theta}}$  is defined through

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg \max_{\boldsymbol{\theta}} \sum_{m=1}^n \left\{ -\frac{1}{2} \log |\mathbf{C}(\boldsymbol{\theta}; m)| - \frac{1}{2} (\hat{\boldsymbol{\zeta}}_m)^T \mathbf{C}(\boldsymbol{\theta}; m)^{-1} \hat{\boldsymbol{\zeta}}_m \right\} \\ &= \arg \max_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}). \end{aligned} \tag{A.2}$$

To save space, when there is no confusion, we use the generic notation  $\mathbf{C}$  or  $\mathbf{C}(\boldsymbol{\theta})$  to denote the correlation matrix function  $\mathbf{C}(\boldsymbol{\theta}; m)$ . For  $1 \leq i, j \leq d$ , write  $\mathbf{C}_i = \frac{\partial \mathbf{C}}{\partial \theta_i}$ ,  $\mathbf{C}^i = \frac{\partial \mathbf{C}^{-1}}{\partial \theta_i} = -\mathbf{C}^{-1} \mathbf{C}_i \mathbf{C}^{-1}$ ,  $\mathbf{C}_{ij} = \frac{\partial^2 \mathbf{C}}{\partial \theta_i \partial \theta_j}$ , and  $\mathbf{C}^{ij} = \frac{\partial^2 \mathbf{C}^{-1}}{\partial \theta_i \partial \theta_j} = \mathbf{C}^{-1} (\mathbf{C}_i \mathbf{C}^{-1} \mathbf{C}_j + \mathbf{C}_j \mathbf{C}^{-1} \mathbf{C}_i - \mathbf{C}_{ij}) \mathbf{C}^{-1}$ .

*Lemma A.1.* For any given  $J_i$  and  $\mathbf{T}_i$ , if  $\mathbf{C}(\boldsymbol{\theta}; i) \neq \mathbf{C}(\boldsymbol{\theta}_0; i)$  for any  $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ , then there is a unique minimizer of  $\log |\mathbf{C}(\boldsymbol{\theta}; i)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; i)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; i)]$ , and the unique minimizer is  $\boldsymbol{\theta}_0$ .

*Proof.* Given  $J_i$  and  $\mathbf{T}_i$ , to prove Lemma A.1, tentatively we assume that  $\boldsymbol{\zeta}_i$  is normally distributed with mean  $\mathbf{0}$  and covariance  $\mathbf{C}(\boldsymbol{\theta}_0; i)$ . Noting that  $\log x \leq x - 1$ , we have

$$E_{\boldsymbol{\theta}_0} \log \frac{f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}; i))}{f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}_0; i))} \leq E_{\boldsymbol{\theta}_0} \left\{ \frac{f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}; i))}{f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}_0; i))} - 1 \right\} = 0,$$

where equality holds only when  $f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}; i))/f(\boldsymbol{\zeta}_i; \mathbf{C}(\boldsymbol{\theta}_0; i)) = 1$  almost surely, that is,  $\mathbf{C}(\boldsymbol{\theta}; i) = \mathbf{C}(\boldsymbol{\theta}_0; i)$  due to the normality assumption.

Noting that the left side of the foregoing equation is equal to

$$\begin{aligned} &-\frac{1}{2} (\log |\mathbf{C}(\boldsymbol{\theta}; i)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; i)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; i)]) \\ &\quad + \frac{1}{2} (\log |\mathbf{C}(\boldsymbol{\theta}_0; i)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}_0; i)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; i)]), \end{aligned}$$

we have

$$\begin{aligned} \log |\mathbf{C}(\boldsymbol{\theta}_0; i)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}_0; i)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; i)] \\ \leq \log |\mathbf{C}(\boldsymbol{\theta}; i)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; i)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; i)], \end{aligned}$$

with equality holding only when  $\mathbf{C}(\boldsymbol{\theta}; i) = \mathbf{C}(\boldsymbol{\theta}_0; i)$ . Thus Lemma A.1 is proved.

To prove Theorem 2, the log-likelihood function can be decomposed as follows:

$$\begin{aligned} &\frac{1}{n} l_n(\boldsymbol{\theta}) \\ &= -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + (\hat{\boldsymbol{\zeta}}_m)^T \mathbf{C}(\boldsymbol{\theta}; m)^{-1} \hat{\boldsymbol{\zeta}}_m \} \\ &= -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \hat{\boldsymbol{\zeta}}_m \hat{\boldsymbol{\zeta}}_m^T] \} \\ &= -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} (\boldsymbol{\zeta}_m + (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)) \\ &\quad \times (\boldsymbol{\zeta}_m + (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m))^T] \} \end{aligned}$$

$$= -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \\ + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)(\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m)^T] \}.$$

Condition (H) implies that for any  $\boldsymbol{\theta} \in \Theta$ ,

$$|(\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m)^T \mathbf{C}(\boldsymbol{\theta}; m)^{-1} (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)| \\ \leq (1/\varrho_0) \|(\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m)\| \cdot \|\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m\|.$$

Condition (E),  $\max_{m=1}^n J_m = O(\log n)$ , and the fact that  $\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m \xrightarrow{P} \mathbf{0}$  at a polynomial rate imply that  $E\|\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m\| \cdot \|\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m\| \rightarrow 0$ . As a result, the law of large numbers implies that

$$\left| \frac{1}{n} l_n(\boldsymbol{\theta}) - \left( -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m (\boldsymbol{\zeta}_m)^T] \} \right) \right| \\ \leq \frac{1}{2n} \sum_{m=1}^n |(\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m)^T \mathbf{C}(\boldsymbol{\theta}; m)^{-1} (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)| \\ \leq \frac{1}{2n\varrho_0} \sum_{m=1}^n \|\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m\| \cdot \|\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m\| \xrightarrow{P} 0$$

uniformly for  $\boldsymbol{\theta} \in \Theta$ . Thus we have

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(1)} \xrightarrow{P} \mathbf{0},$$

where

$$\hat{\boldsymbol{\theta}}^{(1)} = \arg \max_{\boldsymbol{\theta} \in \Theta} \left( -\frac{1}{2} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \} \right).$$

Condition (H) implies that for each  $m$ , the absolute value of

$$\log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T]$$

is bounded by

$$J_i \max(|\log \varrho_0|, |\log \varrho_1|) + (1/\varrho_0) \text{tr}(\boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T),$$

the expectation of which exists, is finite, and does not depend on  $\boldsymbol{\theta}$ . Thus Condition (I) and the law of large numbers imply that for each  $\boldsymbol{\theta}$ ,

$$-\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \} \\ \xrightarrow{P} E_{g_0} \log f(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta})) \\ \equiv -E_{g_0} \log \frac{g_0(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}_0))}{f(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}))} + E_{g_0} \log g_0(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}_0)) \quad (\text{A.3})$$

(cf. White 1982), where  $g_0(\cdot; \cdot)$  specifies the true spherical density and  $f(\cdot; \cdot)$  denotes the multivariate normal density of  $\boldsymbol{\zeta}$  with mean  $\mathbf{0}$  and covariance matrix  $\mathbf{C}(\boldsymbol{\theta})$ . Note that the continuity enforced by Condition (G) on our correlation structure  $\rho(\cdot, \cdot, \boldsymbol{\theta})$  implies that  $E_{g_0} \log f(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}))$  is continuous with respect to  $\boldsymbol{\theta}$ . Then for any  $\epsilon > 0$ , there is a neighborhood  $\mathcal{N}(\boldsymbol{\theta}_1)$  of  $\boldsymbol{\theta}_1$  such that

$$\sup_{\boldsymbol{\theta} \in \mathcal{N}(\boldsymbol{\theta}_1)} \left| -\frac{1}{2n} \sum_{m=1}^n \{ \log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \} - E_{g_0} \log f(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta})) \right| < \epsilon$$

in probability for large  $n$ . It follows that  $\Theta$  may be covered by such neighborhoods and, because  $\Theta$  is compact, as enforced by Condition (F), by a finite collection of such neighborhoods. As a result, the convergence of (A.3) is uniform with respect to  $\boldsymbol{\theta}$ , because the small

number  $\epsilon$  is arbitrary. Thus  $\hat{\boldsymbol{\theta}}^{(1)}$  converges in probability to the minimizer of the Kullback–Leibler information criterion,

$$I(g_0 : f, \boldsymbol{\theta}) \equiv E_{g_0} \log(g_0(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}_0))/f(\boldsymbol{\zeta}; \mathbf{C}(\boldsymbol{\theta}))).$$

For each subject  $m$ , conditional on its number  $J_m$  of observations and observation times  $\mathbf{T}_m$ ,

$$E_{g_0}(\log(g_0(\boldsymbol{\zeta}_m; \mathbf{C}(\boldsymbol{\theta}_0; m))/f(\boldsymbol{\zeta}_m; \mathbf{C}(\boldsymbol{\theta}; m))) | J_m, \mathbf{T}_m) \\ = \frac{1}{2} (\log |\mathbf{C}(\boldsymbol{\theta}; m)| + \text{tr}[\mathbf{C}(\boldsymbol{\theta}; m)^{-1} \mathbf{C}(\boldsymbol{\theta}_0; m)]) + \text{constant} \quad (\text{A.4})$$

has global minimizer  $\boldsymbol{\theta}_0$  due to Lemma A.1 and condition (J).

Note that

$$I(g_0 : f, \boldsymbol{\theta}) = E[E_{g_0}(\log(g_0(\boldsymbol{\zeta}_m; \mathbf{C}(\boldsymbol{\theta}_0; m))/f(\boldsymbol{\zeta}_m; \mathbf{C}(\boldsymbol{\theta}; m))) | J_m, \mathbf{T}_m)].$$

Thus  $\boldsymbol{\theta}_0$  minimizes  $I(g_0 : f, \boldsymbol{\theta})$  globally, which implies that  $\hat{\boldsymbol{\theta}}^{(1)} \xrightarrow{P} \boldsymbol{\theta}_0$  due to (A.3). Theorem 1 is proved by noting that we have shown that  $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(1)} \xrightarrow{P} \mathbf{0}$ .

### Proof of Theorem 3

By routine calculation as in the proof of Theorem 2, we have

$$\frac{1}{n} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \\ = -\frac{1}{2n} \sum_{m=1}^n \{ 2\boldsymbol{\zeta}_m^T \mathbf{C}^i(\boldsymbol{\theta}; m) (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m) \\ + (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)^T \mathbf{C}^i(\boldsymbol{\theta}; m) (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m) \} \\ - \frac{1}{2n} \sum_{m=1}^n \{ \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}; m) \mathbf{C}_i(\boldsymbol{\theta}; m)] + \text{tr}[\mathbf{C}^i(\boldsymbol{\theta}; m) \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \}. \quad (\text{A.5})$$

Note that  $\mathbf{C}^i(\boldsymbol{\theta}, m) = \mathbf{C}^{-1}(\boldsymbol{\theta}, m) \mathbf{C}_i(\boldsymbol{\theta}, m) \mathbf{C}^{-1}(\boldsymbol{\theta}, m)$ , and that  $\boldsymbol{\zeta}_m$  has mean 0 and finite variance. The uniform convergence of  $m(\cdot)$  and  $\sigma^2(\cdot)$  in the decomposition (A.1) implies that the first term on the right side of (A.5) converges to 0 in probability. As a result, we have

$$\frac{1}{n} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \left[ -\frac{1}{2n} \sum_{m=1}^n \{ \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}_0; m) \mathbf{C}_i(\boldsymbol{\theta}_0; m)] \\ + \text{tr}[\mathbf{C}^i(\boldsymbol{\theta}_0; m) \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right] \xrightarrow{P} 0. \quad (\text{A.6})$$

Thus

$$\frac{1}{n} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{P} -\frac{1}{2} E \{ \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}_0; m) \mathbf{C}_i(\boldsymbol{\theta}_0; m)] \\ + \text{tr}[\mathbf{C}^i(\boldsymbol{\theta}_0; m) E_{\boldsymbol{\theta}_0}(\boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T)] \} \\ \equiv 0.$$

Similarly, we can show that

$$\frac{1}{n} \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ \xrightarrow{P} -\frac{1}{2} E \{ \text{tr}[\mathbf{C}^{-1}(\boldsymbol{\theta}_0; m) \mathbf{C}_i(\boldsymbol{\theta}_0; m) \mathbf{C}^{-1}(\boldsymbol{\theta}_0; m) \mathbf{C}_j(\boldsymbol{\theta}_0; m)] \}, \\ \text{and that } \frac{1}{n} \frac{\partial^3 l_n(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \text{ converges in probability to} \\ -\frac{1}{2} E \{ \text{tr}[\mathbf{C}_{ij} \mathbf{C}^{-1} \mathbf{C}_k \mathbf{C}^{-1} + \mathbf{C}_{jk} \mathbf{C}^{-1} \mathbf{C}_i \mathbf{C}^{-1} + \mathbf{C}_j \mathbf{C}^{-1} \mathbf{C}_{ik} \mathbf{C}^{-1} \\ - 2\mathbf{C}_i \mathbf{C}^{-1} \mathbf{C}_j \mathbf{C}^{-1} \mathbf{C}_k \mathbf{C}^{-1} - 2\mathbf{C}_j \mathbf{C}^{-1} \mathbf{C}_i \mathbf{C}^{-1} \mathbf{C}_k \mathbf{C}^{-1}] \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \}.$$

The existence and boundedness of the foregoing expectations can be easily proved using Conditions (B), (G), and (H).

Note that  $\mathbf{I}(\theta_0) = \lim_{n \rightarrow \infty} (-\frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial \theta \partial \theta^T} |_{\theta=\theta_0})$  is a  $d \times d$  matrix with  $(i, j)$  element that is the limit of  $-\frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial \theta_i \partial \theta_j} |_{\theta=\theta_0}$  as  $n \rightarrow \infty$ . Taylor-expand  $\frac{\partial}{\partial \theta} l_n(\hat{\theta})$  at  $\theta_0$ , that is,

$$\begin{aligned} \mathbf{0} &= \frac{1}{n} \frac{\partial}{\partial \theta} l_n(\hat{\theta}) \\ &= \frac{1}{n} \frac{\partial}{\partial \theta} l_n(\theta_0) + \frac{1}{n} \frac{\partial^2 l_n(\theta_0)}{\partial \theta \partial \theta^T} (\hat{\theta} - \theta_0) + \mathbf{R}_n(\theta^*), \end{aligned} \tag{A.7}$$

where  $\theta^*$  is between  $\theta_0$  and  $\theta$  and  $\mathbf{R}_n(\theta) = \mathbf{M}(\theta)(\hat{\theta} - \theta_0)$ , where  $\mathbf{M}(\theta)$  is a  $d \times d$  matrix with  $m$ th row given by

$$\frac{1}{2} (\hat{\theta} - \theta_0)^T \left[ \frac{1}{n} \frac{\partial^2}{\partial \theta \partial \theta^T} \left( \frac{\partial}{\partial \theta_m} l_n(\theta) \right) \right].$$

Thus we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = - \left( \frac{1}{n} \frac{\partial^2 l_n(\theta_0)}{\partial \theta \partial \theta^T} + \mathbf{M}(\theta^*) \right)^{-1} \left( \frac{\sqrt{n}}{n} \frac{\partial}{\partial \theta} l_n(\theta_0) \right).$$

Because  $\hat{\theta}$  is consistent, every element of the matrix  $\mathbf{M}(\theta^*)$  converges to 0 in probability.

Note that from the foregoing, we have that  $\frac{1}{n} \frac{\partial l_n(\theta)}{\partial \theta} |_{\theta=\theta_0} \rightarrow \mathbf{0}$  and  $-\frac{1}{n} \frac{\partial^2 l_n(\theta)}{\partial \theta \partial \theta^T} |_{\theta=\theta_0} \rightarrow \mathbf{I}(\theta_0)$  as  $n \rightarrow \infty$ . To get the desired asymptotic normality result, we need to show that  $-\frac{1}{2n} \sum_{m=1}^n \{2\xi_m^T \mathbf{C}^i(\theta; m) \times (\hat{\xi}_m - \xi_m) + (\hat{\xi}_m - \xi_m)^T \mathbf{C}^i(\theta; m)(\hat{\xi}_m - \xi_m)\} = o_p(n^{-1/2})$ . The second term can be easily shown to be of order  $o_p(n^{-1/2})$  due to Conditions (C), (G), and (H) and the result of Theorem 1. Denote  $\tilde{\mathbf{C}}^i(\theta; m)$  to be a  $(\max_{j=1}^n J_j) \times (\max_{j=1}^n J_j)$  matrix with top-left  $J_m \times J_m$  submatrix  $\mathbf{C}^i(\theta; m)$  and other elements all filled by 0's. Similarly,  $\tilde{\xi}_m$  and  $\tilde{\xi}_m$  are vectors of length  $(\max_{j=1}^n J_j)$  with first  $J_m$  elements given by  $\hat{\xi}_m$  and  $\xi_m$  and other elements filled by 0's. Then the first term is equal to  $-\frac{1}{n} \sum_{m=1}^n \tilde{\xi}_m^T \tilde{\mathbf{C}}^i(\theta; m)(\tilde{\xi}_m - \xi_m)$ , which can be shown to be of order  $o_p(n^{-1/2})$  using techniques similar to the proof of lemma 2 of Lam and Fan (2008). Applying the central limit theorem, we have that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, \mathbf{I}(\theta_0)^{-1} \Sigma(\theta_0) \mathbf{I}(\theta_0)^{-1}).$$

**Proof of Proposition 1**

Part a of Proposition 1 can be verified using basic algebra and is skipped here. We prove part b. With the assumption that  $\min_{i=2,3,\dots,K} (s_i - s_{i-1}) = s_0 > 0$ , and based on part a, we can easily show that  $\mathbf{A}^{-1}$  is diagonal-dominated. The basic properties of hyperbolic function imply that

$$\min_{i=1,2,\dots,K} \left( (\mathbf{A}^{-1})_{ii} - \sum_{j \neq i} (\mathbf{A}^{-1})_{ij} \right) \geq \delta_0(t_0, a).$$

Let  $\mathbf{b} = (b_1, b_2, \dots, b_K)^T$  be an arbitrary vector in the  $K$ -dimensional space. We have

$$\begin{aligned} \mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} &= \sum_{i=1}^K \sum_{j=1}^K (\mathbf{A}^{-1})_{ij} b_i b_j \\ &\geq \sum_{i=1}^K (\mathbf{A}^{-1})_{ii} b_i^2 + \sum_{1 \leq i \neq j \leq K} (\mathbf{A}^{-1})_{ij} (b_i^2 + b_j^2) / 2 \\ &= \sum_{i=1}^K (\mathbf{A}^{-1})_{ii} b_i^2 + \sum_{i=1}^K b_i^2 \left( \sum_{j \neq i} (\mathbf{A}^{-1})_{ij} + \sum_{j \neq i} (\mathbf{A}^{-1})_{ji} \right) / 2. \end{aligned}$$

Noting that  $\mathbf{A}^{-1}$  is symmetric, we have

$$\mathbf{b}^T \mathbf{A}^{-1} \mathbf{b} \geq \sum_{i=1}^K \left[ (\mathbf{A}^{-1})_{ii} - \sum_{j \neq i} (\mathbf{A}^{-1})_{ij} \right] b_i^2 \geq \sum_{i=1}^K \delta_0 b_i^2 = \delta_0 \|\mathbf{b}\|^2.$$

This immediately implies that the smallest eigenvalue of  $\mathbf{A}^{-1}$  is not smaller than  $\delta_0$ , which does not depend on the number  $J$  of observations. Similarly, we can show that the largest eigenvalue of  $\mathbf{A}^{-1}$  is at most  $2 \max_{i=1}^J (\mathbf{A}^{-1})_{ii} \leq \delta_1(t_0, a)$ .

**Proof of Proposition 2**

The  $i$ th subject has observations at times  $t = t_{i1}, t_{i2}, \dots, t_{ij}$ , which are assumed to be in increasing order. According to Proposition 1, the eigenvalues of the correlation matrix of the  $i$ th subject are between

$$\frac{1}{\delta_1(t_0, 1/\varphi_0)} = \inf_{a \geq 1/\varphi_0} \frac{1}{\delta_1(t_0, a)}$$

and

$$\frac{1}{\delta_0(t_0, 1/\varphi_0)} = \sup_{a \geq 1/\varphi_0} \frac{1}{\delta_0(t_0, a)}.$$

Thus part a is proved.

Part b can be easily proved by noting that the correlation matrix for ARMA(1, 1) model is exactly  $(1 - \gamma)\mathbf{I} + \gamma\mathbf{D}$ , where  $\mathbf{D}$  is the corresponding correlation matrix of AR(1) with the same parameter  $\varphi$  and  $\mathbf{I}$  is the identity matrix.

To prove part c, note that the correlation matrix for CARMA( $p, q$ ) model can be expressed as  $\sum_{i=1}^p \gamma_i \mathbf{D}_i$ , where  $\mathbf{D}_i$  is the corresponding correlation matrix of AR(1) with parameter  $\varphi_i$ . Part c follows straightforwardly by applying Weyl's inequality.

[Received March 2008. Revised June 2008.]

**REFERENCES**

Bickel, P., and Levina, E. (2006), "Regularized Estimation of Large Covariance Matrices," Technical Report 716, University of California Berkeley, Dept. of Statistics.

Brown, L. D., and Levine, M. (2007), "Variance Estimation in Nonparametric Regression via the Difference Sequence Method," *The Annals of Statistics*, 35, 2219–2232.

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2002), *Analysis of Longitudinal Data* (2nd ed.), New York: Oxford University Press.

Fan, J., and Huang, L. (2001), "Goodness-of-Fit Test for Parametric Regression Models," *Journal of American Statistical Association*, 96, 640–652.

— (2005), "Profile Likelihood Inferences on Semi-Parametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1059.

Fan, J., and Yao, Q. (1998), "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrika*, 85, 645–660.

Fan, J., and Zhang, J.-T. (2000), "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of Royal Statistical Society, Ser. B*, 62, 303–322.

Fan, J., and Zhang, W. (1999), "Statistical Estimation in Varying-Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Fan, J., Huang, T., and Li, R. (2007), "Analysis of Longitudinal Data With Semiparametric Estimation of Covariance Function," *Journal of the American Statistical Association*, 102, 632–641.

Hall, P., Kay, J. W., and Titterton, D. M. (1990), "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression," *Biometrika*, 77, 521–528.

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models" (with discussion), *Journal of Royal Statistical Society, Ser. B*, 55, 757–796.

Huang, J. Z., Liu, L., and Liu, N. (2007), "Estimation of Large Covariance Matrices of Longitudinal Data With Basis Function Approximations," *Journal of Computational and Graphical Statistics*, 16, 189–209.

Lam, C., and Fan, J. (2008), "Profile-Kernel Likelihood Inference With Diverging Number of Parameters," *The Annals of Statistics*, 36, 2232–2260.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Model," *Biometrika*, 73, 13–22.

Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.

- Phadke, M. S., and Wu, S. M. (1974), "Modeling of Continuous Stochastic Processes From Discrete Observations With Application to Sunspots Data," *Journal of the American Statistical Association*, 69, 325–329.
- Qu, A., Lindsay, B. G., and Li, B. (2000), "Improving Generalized Estimating Equations Using Quadratic Inference Functions," *Biometrika*, 87, 823–836.
- Rice, J., and Silverman, B. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When Data Are Curves," *Journal of the Royal Statistical Society, Ser. B*, 53, 233–243.
- Rothman, A. J., Bickel, P., Levina, L., and Zhu, J. (2007), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics*, 2, 494–515.
- Ruppert, D., Sheather, S., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.
- Sowers, M., Randolph, J. F., Crutchfield, M., Jannausch, M. L., Shapiro, B., Zhang, B., and La Pietra, M. (1998), "Urinary Ovarian and Gonadotropin Hormone Levels in Premenopausal Women With Low Bone Mass," *Journal of Bone and Mineral Research*, 13, 1191–1202.
- Wang, N. (2003), "Marginal Nonparametric Kernel Regression Accounting Within-Subject Correlation," *Biometrika*, 90, 29–42.
- Wang, N., Carroll, R. J., and Lin, X. (2005), "Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data," *Journal of the American Statistical Association*, 100, 147–157.
- White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–26.
- Wu, B., and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika*, 90, 831–844.
- Yao, F., Müller, H.-G., and Wang, J.-L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590.
- (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.
- Yatchew, A. (1997), "An Elementary Estimator for the Partially Linear Model," *Economics Letters*, 57, 135–143.
- Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.