Semiparametric estimation of covariance matrices for longitudinal data

Jianqing Fan and Yichao Wu *

Princeton University and North Carolina State University

Abstract

Estimation of longitudinal data covariance structure poses significant challenges because the data are usually collected at irregular time points. A viable semiparametric model for covariance matrices was proposed in Fan, Huang and Li (2007) that allows one to estimate the variance function nonparametrically and to estimate the correlation function parametrically via aggregating information from irregular and sparse data points within each subject. However, the asymptotic properties of their quasi-maximum likelihood estimator (QMLE) of parameters in the covariance model are largely unknown. In the current work, we address this problem

*Jianqing Fan is Frederick L. Moore'18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (E-mail: jqfan@princeton.edu). Yichao Wu is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh NC 27695 (E-mail: wu@stat.ncsu.edu). This research is partially supported by National Science Foundation (NSF) grant DMS-03-54223 and National Institutes of Health (NIH) grant R01-GM07261. Address for correspondence: Yichao Wu, Department of Statistics, North Carolina State University, Raleigh NC 27695. Finally, we would like to thank the editor Professor Leonard Stefanski, the AE and two referees for their helpful comments that have led to the improvement of the manuscript.

in the context of more general models for the conditional mean function including parametric, nonparametric, or semi-parametric. 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.

Keywords: Correlation structure, difference-based estimation, quasi-maximum likelihood, varying-coefficient partially linear model.

AMS 2000 Subject Classification: 62F12, 62G08.

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, one 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. The longitudinal hormone study on progesterone (Sowers et al., 1998) collected urine samples from 34 healthy women in a menstrual cycle on alternative days. Zhang et al. (1998) analyzed the data using semi-parametric stochastic mixed models. A total of 492 observations were made on the 34 subjects in the study, with between 11 and 28 observations per subject. Menstrual cycle lengths of the subjects ranged from 23 to 56 days and averaged 29.6 days. Biologically it makes sense to assume that the change in progesterone level for a woman depends on the time during a menstrual cycle relative to her cycle length. So the menstrual cycle length of each woman was standardized (Sowers et al., 1998). A typical logarithmic transformation is applied on the progesterone level to make the data more homoscedastic. The progesterone data are unbalanced in that different subjects have different numbers of observations, and observation times are not regular and differ from one subject to another.

To address these challenges in covariance matrix estimation, Fan et al. (2007) modeled the variance function nonparametrically and correlation structure parametrically. They mainly focused 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 QMLE of parameters in correlation function. We therefore study the problem under a general mean-regression model

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

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

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

where $\rho(\cdot, \cdot, \boldsymbol{\theta})$ is a positive definite function for any $\boldsymbol{\theta} \in \Theta \subset I\!R^d$. 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 takes 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. However, generally correct specification of the correlation structure 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 higher correlation in the response. Hence we use an ARMA(1, 1) correlation structure while analyzing the progesterone data in Section 6.

Our approach of flexible covariance structure estimation sheds light on solving a long-standing problem on improving the efficiency of parameter estimation using ideally the unknown true covariance structure. In a seminal paper, Liang and Zeger (1986) introduced generalized estimating equations (GEE), extending generalized linear models to longitudinal data, and proposed using a working correlation matrix to improve efficiency. However, misspecification of the working correlation matrix is possible. To improve efficiency under misspecification, Qu, Lindsay and Li (2000) represented the inverse of the working correlation matrix by a linear combination of basis matrices 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 showed a rather unexpected result that higher efficiency is obtained by assuming independence, than by using the true correlation structure. In their later work, it was shown 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 semi-parametric, partially linear models. All these works require specifying a true correlation matrix, but do not provide a systematic estimation scheme. Our method provides a flexible approach to this important endeavor.

There are several approaches for estimating a covariance matrix. Most are nonparametric. Wu and Pourahmadi (2003) used non-parametric 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 Wu and Pourahmadi (2003)'s. Bickel and Levina (2006) investigated the regularization of covariance matrix via the idea of banding and obtained many insightful results. See also Rothman et al. (2007). All these aforementioned estimation methods have the same limitation that observations are assumed to be made over a grid, *i.e.*, balanced or nearly balanced longitudinal data. But this assumption is often not tenable, especially for longitudinal data that are commonly collected at irregular and possibly subject-specific time points. This feature of longitudinal data makes it challenging to study its covariance structure. Yao, Müller and Wang (2005a,b) took a different approach based on functional data analysis.

The remainder of the paper is organized as follows. In Section 2 we discuss estimation of the conditional mean function $m(\boldsymbol{x})$. We focus on semi-parametric, varying-coefficient, partially linear models for which we propose a more robust estimation scheme. Section 3 presents the estimation method for the semi-parametric covariance structure. The asymptotic properties of the estimated variance function and correlation structure parameter are given in Section 4. Extensive simulation studies and an application to the progesterone data are given in Sections 5 and 6, respectively. Technical proofs appear in the Appendix.

2 Semi-parametric 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 $(\boldsymbol{x}(t_j), y(t_j))$ at times $\{t_j\}$ from model (1) by $\mathbb{X} = \{J, (t_j, \boldsymbol{x}(t_j), y(t_j)), j = 1, 2, \cdots, J\}$, with data from subject i denoted by $\mathbb{X}_i = \{J_i, (t_{ij}, \boldsymbol{x}_i(t_{ij}), y_i(t_{ij})), j = 1, 2, \cdots, J_i\}$. Thus the complete data set

is represented as $\{\mathbb{X}_1, \mathbb{X}_2, \cdots, \mathbb{X}_n\}$.

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

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

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

To estimate the varying-coefficient partially linear model, Fan et al. (2007) assumed the existence of a second-order derivative of $\boldsymbol{\alpha}_0(\cdot)$. However 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 5.3 in Section 5, estimation can be badly 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 cause 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 remarkable 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 (Yatchew, 1997; Fan and Huang, 2001, 2005; Brown and Levine, 2007, to name a few), but all in the univariate nonparametric setup. In our setting, multiple nonparametric functions are needed 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 $\big\{(t_{(j)}, \boldsymbol{x}_{1(j)}, \boldsymbol{x}_{2(j)}, y_{(j)}), j = 1, 2, \cdots, N\big\}$, 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

$$\| \boldsymbol{\alpha}_0(t_{(i+j)}) - \boldsymbol{\alpha}_0(t_{(i)}) \| = O_p((t_{(i+j)} - t_{(i)})^{\kappa}) = O_p(1/n^{\kappa}), \text{ for } j = 1, 2, \cdots, d_1. \tag{3}$$

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

$$y^*_{(i)} = \sum_{j=i}^{i+d_1} w_{i,j} y_{(j)}, \qquad \boldsymbol{x}^*_{2(i)} = \sum_{j=i}^{i+d_1} w_{i,j} \boldsymbol{x}_{2(j)}, \qquad \varepsilon^*_{(i)} = \sum_{j=i}^{i+d_1} w_{i,j} \varepsilon_{(j)}.$$

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

$$\sum_{j=i}^{i+d_1} w_{i,j} \boldsymbol{x}_{1(j)} = \mathbf{0}, \qquad \forall 1 \le i \le N - d_1 \tag{4}$$

to remove the nonparametric component $\boldsymbol{x}_1(t)^T \boldsymbol{\alpha}(t)$. The weights are further normalized as in Hall, Kay and Titterington (1990) such that $\sum_{j=1}^{N} w_{i,j}^2 = 1$. Note that if $\boldsymbol{x}_{1(i)}, \boldsymbol{x}_{1(i+1)}, \cdots, \boldsymbol{x}_{1(i+d_1-1)}$ are linearly independent (which holds with probability one under some mild conditions for continuously distributed $\boldsymbol{x}_1$), the weights are uniquely

determined up to a sign change. More explicitly,

$$(w_{i,i}, w_{i,i+1}, \cdots, w_{i,i+d_1}) = \pm \frac{(((\boldsymbol{x}_{1(i)}, \boldsymbol{x}_{1(i+1)}, \cdots, \boldsymbol{x}_{1(i+d_1-1)})^{-1} \boldsymbol{x}_{1(i+d_1)})^T, -1)}{\parallel (((\boldsymbol{x}_{1(i)}, \boldsymbol{x}_{1(i+1)}, \cdots, \boldsymbol{x}_{1(i+d_1-1)})^{-1} \boldsymbol{x}_{1(i+d_1)})^T, -1) \parallel}. \quad (5)$$

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

A combination of (2)-(4) leads to

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

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

Model (6) is a standard multivariate linear regression problem. Applying the OLS technique on (6) with data $\left\{(\boldsymbol{x}_{2(i)}^*, y_{(i)}^*), i = 1, 2, \cdots, N - p\right\}$, we get the difference-based estimator (DBE) of $\boldsymbol{\beta}$. Due to the fact that the approximation error in (6) is of order $O_p(1/n^\kappa)$, standard result for the OLS implies that our DBE $\hat{\boldsymbol{\beta}}$ is consistent with order $n^{-\kappa \wedge 0.5}$, where $a \wedge b = \min(a, b)$.

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

2.2 Kernel smoothing estimator of $\boldsymbol{\alpha}(\cdot)$

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

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

with approximation error of order $O_p(n^{-\kappa\wedge 0.5})$. Model (7) is exactly a varying-coefficient model, and was studied by Hastie and Tibshirani (1993) for the case of i.i.d. observations and also by Fan and Zhang (2000) in the context of longitudinal data. Local smoothing technique can be used to estimate $\boldsymbol{\alpha}(\cdot)$. Due to 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 $\{(t_{ij}, \boldsymbol{x}_{1,i}(t_{ij}), \tilde{y}_i(t_{ij})), j = 1, 2, \cdots, J_i; i = 1, 2, \cdots, n\}$.

Note that, for any t in a neighborhood of t_0, due to Condition [L] we have

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

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

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

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

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

where $\boldsymbol{X}_1 = (\boldsymbol{x}_{1,1}(t_{11}), \boldsymbol{x}_{1,1}(t_{12}), \cdots, \boldsymbol{x}_{1,n}(t_{nJ_n}))^T$. 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 semi-parametric 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 (iii) in the

Appendix. We next present our estimation scheme of the semi-parametric 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 follows

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

that 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 the kernel smoothing to estimate the variance function $\sigma^2(t) = E\varepsilon^2(t)$, via

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

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 $\mathrm{corr}(\boldsymbol{\varepsilon}_i)$ by $\boldsymbol{C}(\boldsymbol{\theta}; i)$ whose (j,k)-element is $\rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$, $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \cdots, r_{iJ_i})^T$, and $\hat{\boldsymbol{V}}_i = \mathrm{diag}\{\hat{\sigma}(t_{i1}), \hat{\sigma}(t_{i2}), \cdots, \hat{\sigma}(t_{iJ_i})\}$.

We estimate the correlation structure parameter $\boldsymbol{\theta}$ by quasi-maximum likelihood method,

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

Denote by $\zeta(t) \equiv \varepsilon(t)/\sigma(t)$. For a generic subject $\mathbb{X}$ with J observations, the "stan-

dardized" random error vector $\boldsymbol{\zeta} = (\zeta(t_1), \zeta(t_2), \cdots, \zeta(t_J))^T$ is assumed to follow an elliptically contoured distribution, having a multivariate density function proportional to $|\boldsymbol{C}(\boldsymbol{\theta}_0)|^{-1/2} h_0(\boldsymbol{\zeta}^T \boldsymbol{C}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\zeta})$, where $\boldsymbol{C}(\boldsymbol{\theta}_0)$ 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 (below a threshold) for some individuals, the matrix $\boldsymbol{C}(\boldsymbol{\theta}; i)$ can be ill-conditioned. In this case, we can either delete some of their observations or remove those cases, thus reducing the influence of those individuals. Under Condition (ii), such cases are rare as 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 semi-parametric covariance structure in our model (1).

To derive asymptotic properties, we assume that the data is a random sample collected from the population process $\{y(t), \boldsymbol{x}(t)\}$ 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, $i = 1, 2, \cdots, n$, are independent and identically distributed and Condition (ii) is satisfied. For each subject i, given the number of observations J_i, observation times t_{ij}, $j = 1, 2, \cdots, J_i$, are independently and identically distributed with density function $f(t)$. Large-sample properties of our estimators are stated in Theorems 1-3. Technical conditions and proofs are relegated to the Appendix.

4.1 Estimator of the covariance structure

Denote by $\dot{\sigma}_0^2(\cdot)$ and $\ddot{\sigma}_0^2(\cdot)$ the first and second derivatives of the variance function $\sigma_0^2(\cdot)$, respectively.

Theorem 1 *Under Conditions (i-v), if $h \propto n^{-\frac{1}{5}}$, then, as $n \to \infty$,*

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

where the bias and variance are given by $b(t) = h^2\mu_2\left[\ddot{\sigma}_0^2(t) + 2\dot{\sigma}_0^2(t)f'(t)/f(t)\right]/2$ and $v(t) = \mathrm{Var}(\varepsilon^2(t))\nu_0/(f(t)E(J_1))$, respectively, with $\mu_2 = \int u^2K(u)du$ and $\nu_0 = \int K^2(u)du$.

When the correlation structure $\rho(\cdot,\cdot,\boldsymbol{\theta})$ is correctly specified, our 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 (i)-(x) listed in Appendix, and with h specified in Theorem 1, the QMLE $\hat{\boldsymbol{\theta}}$ is consistent, i.e.,*

$$\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0, \quad \textit{in probability, as } n \to \infty. \tag{12}$$

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

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

where $\boldsymbol{C}(\boldsymbol{\theta}_0)$ is the correlation matrix of a generic subject $\mathbb{X}$, $\boldsymbol{C}_i(\boldsymbol{\theta}) = \dfrac{\partial \boldsymbol{C}(\boldsymbol{\theta})}{\partial \theta_i}$, $\boldsymbol{C}^i(\boldsymbol{\theta}) = \dfrac{\partial \boldsymbol{C}^{-1}(\boldsymbol{\theta})}{\partial \theta_i} = -\boldsymbol{C}^{-1}(\boldsymbol{\theta})\boldsymbol{C}_i(\boldsymbol{\theta})\boldsymbol{C}^{-1}(\boldsymbol{\theta})$ and the expectation is due to the randomness of the observation times and taken with respect to the true underlining 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 $\boldsymbol{C}(\boldsymbol{\theta}, m)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denotes a $d \times d$ matrix whose (i,j)-element

is given by

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

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

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

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

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

and

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

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 matrices

Note that the structural condition (viii) in the Appendix is very common when studying covariance matrices. 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 (viii) 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 $\rho(s,t,(\gamma,\varphi)^T) = \gamma\exp(-|s-t|/\varphi)$, respectively, 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, *i.e.*, $\rho(s,t,\boldsymbol{\theta}) = \sum_{i=1}^{p}\gamma_i\exp(-|s-t|/\varphi_i)$ with $\varphi_i \geq 0$, $0 \leq \gamma_i \leq 1$, and $\sum_{i=1}^{p}\gamma_i = 1$.

Proposition 1 *Let $\boldsymbol{A}$ be a $K \times K$ matrix with (i,j)-element $(\boldsymbol{A})_{ij} = \exp(-a|s_i - s_j|)$ where $s_1 < s_2 < \cdots < s_K$ and $a > 0$.*

[a] Then the (i,j)-element of $\boldsymbol{A}^{-1}$ is given by: $(\boldsymbol{A}^{-1})_{11} = (1+\coth(a(s_2-s_1)))/2$; $(\boldsymbol{A}^{-1})_{ii} = (\coth(a(s_i-s_{i-1})) + \coth(a(s_{i+1}-s_i)))/2$ for $i = 2,3,\cdots,K-1$; $(\boldsymbol{A}^{-1})_{KK} = (1+\coth(a(s_K-s_{K-1})))/2$; $(\boldsymbol{A}^{-1})_{ij} = -(\mathrm{csch}(a|s_i-s_j|))/2$ for $|i-j| = 1$; $(\boldsymbol{A}^{-1})_{ij} = 0$ when $|i-j| > 1$, where $\coth(\cdot)$ and $\mathrm{csch}(\cdot)$ are the hyperbolic cotangent and cosecant functions.

[b] If $\min_{i=2,3,\cdots,K}(s_i - s_{i-1}) = s_0 > 0$, the eigenvalues of $\boldsymbol{A}^{-1}$ 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, both δ_0 and δ_1 do not depend 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 (viii) is satisfied for either 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: $\rho(s,t,(\gamma_1,\gamma_2,\cdots,\gamma_p,\varphi_1,\varphi_2,\cdots,\varphi_p)^T) = \sum_{i=1}^{p}\gamma_i\exp(-|s-t|/\varphi_i)$ *with* $q<p$, $(\gamma_i,\varphi_i)\in[0,1]\times[0,\varphi_0]$ *for* $i=1,2,\cdots,p$, *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 AIC/BIC-related procedure because the estimation of the parametric correlation structure is based on maximum likelihood once we have estimated the regression function $m(\boldsymbol{x})$ and the variance function $\sigma^2(t)$.

5 Monte Carlo Study

In this section, we study the finite-sample performance of the semi-parametric covariance matrix estimator presented in Section 3. While focusing on the varying-coefficient partially linear model in Example 5.2, we use Example 5.1 to demonstrate the efficiency improvement by incorporating the estimated covariance structure for the case of parametric models. A comparison study is provided in Example 5.3 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 $\{0,1,\cdots,12\}$ and each scheduled time has a probability of 20% being skipped except the time 0. For each non-skipped 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 $\sigma^2(t)=0.5\exp(t/12)$. Either a true AR(1) (Example 5.1) or a true ARMA(1,1) (Examples 5.2 and 5.3) correlation structure is assumed. That is, $\mathrm{corr}(\epsilon(s),\epsilon(t))=\gamma\rho^{|s-t|}$ 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 $t_1,t_2,\cdots,t_J$, the "standardized" random error vector $(\zeta(t_1),\zeta(t_2),\cdots,\zeta(t_J))^T$ have

marginal Normal (N) or Double Exponential (DE) distribution with mean zero, variance one, and correlation matrix $\boldsymbol{C}(\gamma, \rho; t_1, t_2, \cdots, t_J)$ whose (i,j)-element is $\gamma\rho^{|t_i - t_j|}$ when $i \neq j$ and 1 otherwise.

If not specified, our simulation result is based on 1000 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 Errors (RASE), which is defined as $\text{RASE}(\hat{\sigma}^2) = \sqrt{\sum_{k=1}^{K}(\hat{\sigma}^2(t_k) - \sigma^2(t_k))^2/K}$, where $\{t_k : k = 1, 2, \cdots, K\}$ form a uniform grid. The number of grid points K is set to be 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, respectively.

Example 5.1[Parametric model]

In this example, the covariate is three-dimensional, *i.e.*, $\boldsymbol{x} = (x_1, x_2, x_3)^T \in I\!R^3$, with standard normal marginal distribution and correlation $\text{corr}(x_i, x_j) = 0.5^{|i-j|}$. The true regression parameter vector is set to be $\boldsymbol{\beta}_0 = (2, 1.5, 3)^T$. An AR(1) correlation structure is used with three different parameter values, $\rho = 0.3, 0.6, 0.9$. 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 parenthesis) of the OLS estimator of $\boldsymbol{\beta}$ in the third column block. The fourth and fifth correspond to the kernel estimator of $\sigma^2(\cdot)$ and the QMLE of ρ, respectively. The last column gives the asymptotic standard errors (ASE) of QMLE $\hat{\rho}$ using the formula (13) in Theorem 3 with estimated matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ based on 1000 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 zero and the cor-

Table 1: Finite-sample performance for estimating $\boldsymbol{\beta}$, $\sigma^2(\cdot)$ and ρ.

Noise	ρ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	RASE($\hat{\sigma}^2$)	$\hat{\rho}$	ASE
N	0.30	1.9993 (0.0236)	1.5002 (0.0266)	2.9991 (0.0237)	0.0493 (0.0456)	0.2997 (0.0219)	0.0263
	0.60	1.9995 (0.0237)	1.5002 (0.0264)	2.9993 (0.0235)	0.0565 (0.0509)	0.5988 (0.0172)	0.0165
	0.90	1.9996 (0.0238)	1.4998 (0.0259)	3.0000 (0.0231)	0.0702 (0.0605)	0.8981 (0.0077)	0.0041
DE	0.30	1.9992 (0.0244)	1.5001 (0.0260)	2.9999 (0.0235)	0.0692 (0.0622)	0.3000 (0.0227)	0.0241
	0.60	1.9993 (0.0246)	1.5002 (0.0257)	2.9997 (0.0235)	0.0732 (0.0656)	0.5992 (0.0173)	0.0172
	0.90	1.9995 (0.0246)	1.5003 (0.0258)	2.9994 (0.0241)	0.0797 (0.0704)	0.8981 (0.0075)	0.0050

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

Noise	ρ	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
N	0.30	1.9994 (0.0190)	1.4996 (0.0217)	2.9993 (0.0199)
	0.60	1.9995 (0.0143)	1.4996 (0.0164)	2.9995 (0.0151)
	0.90	1.9998 (0.0069)	1.4998 (0.0079)	2.9998 (0.0073)
DE	0.30	1.9992 (0.0192)	1.5005 (0.0220)	3.0004 (0.0189)
	0.60	1.9994 (0.0144)	1.5004 (0.0167)	3.0004 (0.0143)
	0.90	1.9997 (0.0069)	1.5002 (0.0080)	3.0002 (0.0069)

responding true ρ, respectively, 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 $\rho = 0.9$. This can be explained by noting the proof of our asymptotic normality result. While 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. However $\rho = 0.9$ implies that, in Proposition 1, $\varphi = -1/\log(0.9) = 9.49$, which is relatively large. The problem disappears when the sample size is large. Similar phenomenon is observed as well for ARMA(1,1) correlation structure in Example 5.2.

For the parametric model $m(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, we can incorporate the estimated covariance structure to estimate $\boldsymbol{\beta}$ using weighted least squares regression, whose performance is reported in Table 2. Comparing OLS and WLS estimators, we see that the standard deviation of the estimator of $\boldsymbol{\beta}$ is significantly reduced by incorporating the estimated covariance structure, especially in the case of high correlation $\rho = 0.9$.

Example 5.2[Semi-parametric varying-coefficient partially linear model]

In this example, data are generated from model (2), with ARMA(1,1) correlation

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

Noise	(γ, ρ)	$\hat{\beta}_1$	$\hat{\beta}_2$	RASE($\hat{\alpha}_1$)	RASE($\hat{\alpha}_2$)	RASE($\hat{\sigma}^2$)
N	(0.85, 0.30)	1.0003 (0.0322)	1.9993 (0.0582)	0.0695 (0.0186)	0.0814 (0.0157)	0.0618 (0.0227)
	(0.85, 0.60)	1.0006 (0.0334)	2.0030 (0.0565)	0.0712 (0.0218)	0.0817 (0.0155)	0.0673 (0.0245)
	(0.85, 0.90)	1.0003 (0.0347)	1.9985 (0.0571)	0.0718 (0.0280)	0.0827 (0.0157)	0.0786 (0.0335)
DE	(0.85, 0.30)	0.9994 (0.0339)	1.9993 (0.0584)	0.0667 (0.0179)	0.0821 (0.0152)	0.0884 (0.0330)
	(0.85, 0.60)	0.9995 (0.0337)	1.9987 (0.0576)	0.0694 (0.0205)	0.0822 (0.0153)	0.0919 (0.0356)
	(0.85, 0.90)	0.9998 (0.0331)	2.0001 (0.0581)	0.0718 (0.0289)	0.0826 (0.0155)	0.0950 (0.0393)

structure on $\varepsilon(t)$. In this case, $\boldsymbol{x}$ is four-dimensional with $\boldsymbol{x}_1 = (x_1, x_2)^T \in I\!R^2$ and $\boldsymbol{x}_2 = (x_3, x_4)^T \in I\!R^2$. We set the first component of $\boldsymbol{x}_1$ to be constant one to include the intercept term, *i.e.*, $x_1(t) \equiv 1$. For any given time t, $x_2(t)$ and $x_3(t)$ are jointly generated such that they have standard normal marginal distribution and correlation 0.5, and $x_4(t)$ is Bernoulli-distributed with success probability 0.5, independent of $x_2(t)$ and $x_3(t)$. The regression coefficients are specified as follows,

$$\alpha_1(t) = \sqrt{t/12}, \quad \alpha_2(t) = \sin(2\pi t/12), \quad \text{and } \boldsymbol{\beta} = (1, 2)^T.$$

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

Performance of estimating $\boldsymbol{\beta}$, $\boldsymbol{\alpha}(\cdot)$ and $\sigma^2(\cdot)$

As $\boldsymbol{x}_1 \in I\!R^2$, DBE uses three neighboring (sorted) observations with weights given by (5) with positive sign. Mean and standard deviation (in parenthesis) of our DBE of β_1 and β_2 over 1000 repetitions are reported in the third column of Table 3 for three pairs of ARMA(1,1) correlation structure parameter $\boldsymbol{\theta}^T = (\gamma, \rho) = (0.85, 0.30), (0.85, 0.60), (0.85, 0.90)$. Table 3 also displays the mean and standard deviation (in parenthesis) of the RASEs of $\hat{\boldsymbol{\alpha}}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ in the fourth and the fifth column blocks, respectively. From Table 3, we can see that the DBE of $\boldsymbol{\beta}$, the local constant regression estimator of $\boldsymbol{\alpha}(\cdot)$, and the local kernel smoothing estimator of $\sigma^2(\cdot)$ are very precise in the finite-sample case.

Performance of QMLE of $\boldsymbol{\theta}$

Our QMLE $\hat{\boldsymbol{\theta}}$ is based on the estimation of other components in our model. The

Table 4: Finite-sample performance of QMLE of $\boldsymbol{\theta}$

Noise	(γ, ρ)	$(\hat{\gamma}, \hat{\rho})$	asymptotic covariance
N	(0.85, 0.30)	0.8458 (0.0677) 0.2995 (0.0439) [-0.7749]	(0.0636), (0.0407) [-0.7547]
	(0.85, 0.60)	0.8428 (0.0327) 0.6000 (0.0301) [-0.6432]	(0.0303), (0.0272) [-0.7281]
	(0.85, 0.90)	0.8427 (0.0153) 0.8990 (0.0106) [0.0351]	(0.0112), (0.0083) [-0.5898]
DE	(0.85, 0.30)	0.8479 (0.0772) 0.2996 (0.0462) [-0.7734]	(0.0741), (0.0442) [-0.7492]
	(0.85, 0.60)	0.8441 (0.0353) 0.5983 (0.0315) [-0.6638]	(0.0344), (0.0304) [-0.6545]
	(0.85, 0.90)	0.8421 (0.0155) 0.8984 (0.0113) [-0.0011]	(0.0129), (0.0094) [-0.3307]

result in Table 3 indicates that the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}(\cdot)$ and $\hat{\sigma}^2(\cdot)$ perform very well. Hence we expect similarly good performance for the QMLE $\hat{\boldsymbol{\theta}}$. Its simulation results are shown in Table 4. In Table 4, column three reports the mean of $\hat{\gamma}$ and $\hat{\rho}$, with their standard deviations in parentheses and their correlation in square bracket, 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 $\boldsymbol{I}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ based on 1000 simulations.

From Table 4, we can see that the estimated parameters in the correlation structure are very close to the corresponding true ones 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 $\boldsymbol{\theta}^T = (\gamma, \rho) = (0.85, 0.3)$ or $(0.85, 0.6)$, the standard deviation and correlation of our QMLE $\hat{\boldsymbol{\theta}}$ is very close to the asymptotic standard errors and correlation (in the last column) using our asymptotic normality formulas. Similarly as in the previous example, we observe that the sample correlations deviate a lot from their corresponding simulated correlations using formula (13) for the two cases of higher correlation $(\gamma, \rho) = (0.85, 0.9)$ and the same explanation applies here.

Prediction

As mentioned in the introduction, estimating the covariance structure can improve the prediction accuracy for a particular trajectory. In this example, we study the improvement of prediction after estimating the covariance structure. For each case, we generate an independent prediction data set of size 400 exactly in the same way 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

Table 5: Finite-sample performance of prediction

Noise	(γ, ρ)	prediction 1	prediction 2	prediction 3	prediction 4	prediction 5	# of predictions
N	(0.85, 0.30)	1784.06	1799.02 (10.28)	1799.77 (10.48)	1884.51	1899.89 (10.22)	2124
	(0.85, 0.60)	1327.10	1352.88 (9.65)	1353.64 (9.82)	1766.82	1794.50 (10.32)	2095
	(0.85, 0.90)	754.34	788.16 (11.33)	788.27 (11.32)	1891.00	1917.81 (19.54)	2068

0.5 is generated; if it is zero, the response of this observation is treated as "missing" and needs to be predicted; if it is one, this observation is fully observed and used to predict the "missing" ones. Prediction is made using the prediction formula given in Section 5.3 of Fan et al. (2007).

Table 5 reports the mean and standard deviation (in parenthesis) over 1000 repetitions of the sum of squared prediction errors (SSPE) for five different types of prediction in the case of normal "standardized" random errors. Similar result can be carried out for the case with double exponential "standardized" random errors. Prediction 1 corresponds to the oracle, *i.e.*, using true $\boldsymbol{\beta}$, $\boldsymbol{\alpha}(\cdot)$, $\sigma^2(\cdot)$ and $\boldsymbol{\theta}$; prediction 2 uses estimated $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}(\cdot)$, $\hat{\sigma}^2(\cdot)$ and true $\boldsymbol{\theta}$; prediction 3 use all estimated parameters, *i.e.*, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}(\cdot)$, $\hat{\sigma}^2(\cdot)$ and $\hat{\boldsymbol{\theta}}$; predictions 4 and 5 correspond to predictions that ignore the covariance structure, based on true $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$ and estimated $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}(\cdot)$, respectively. 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.

Notice first, in Table 5, that 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 to prediction ignoring the covariance structure, incorporating the estimated covariance structure reduces 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 $\boldsymbol{\theta}$ and the estimated $\hat{\boldsymbol{\theta}}$ does not make much difference after including the estimated covariance structure and demonstrates the accuracy

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

Noise	FW	FHL
$\alpha_1(\cdot)$	0.0381 (0.0164)	0.0281 (0.0090)
$\alpha_2(\cdot)$	0.1946 (0.0277)	0.4525 (0.0403)

of $\hat{\boldsymbol{\theta}}$ for prediction.

Example 5.3[Comparison with the method in Fan et al. (2007)]

For estimating the varying-coefficient partially linear model, the major difference between our method and Fan et al. (2007)'s is that we estimate $\boldsymbol{\beta}$ using the difference-based technique, plug it into model (2), and estimate $\boldsymbol{\alpha}(\cdot)$ via the local constant regression whereas they estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$ simultaneously using the local linear regression and profile-likelihood techniques. Next we use another simulation to study the impact of rough $\boldsymbol{\alpha}(\cdot)$ 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 $\alpha_2(t) = 4/(1 + \exp(-40(t - 7))) - 2$. Notice that $\alpha_2(\cdot)$ is a scaled sigmoid function and is close to a jump function as shown by the dotted line in panel (D) of Figure 1. 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 $a_1 = 64$, $a_2 = a_3 = 0.95 \times \text{sign}(7 - t) \times 8$, and $a_4 = 1$. All other components in our model are simulated the same way as in the previous example. Sample size is $n = 100$; the true ARMA(1,1) correlation structure parameters are $\gamma = 0.85$ and $\rho = 0.6$. After tuning, Fan et al. (2007)'s profiling method chooses the best smoothing bandwidth pair $b = 0.8, h = 3.80$ and our new proposed estimation scheme selects $b = 0.4, h = 0.75$ for estimation $\boldsymbol{\alpha}(\cdot)$ and $\sigma^2(\cdot)$. A box-plot of the estimated correlation structure parameters over these 100 samples for each method is shown in panel (A) of Figure 1.

From the box-plot in Figure 1, we see that Fan et al. (2007)'s estimator is far off the true correlation structure parameters while our method still performs very well. The underlining reason is that, in Fan et al. (2007)'s method, smoothing the rough $\boldsymbol{\alpha}(\cdot)$ causes a big bias in the profile-likelihood estimator of $\boldsymbol{\beta}$ due to the strong correlation between $x_2(\cdot)$ and $x_3(t)$. To see this, the profile-likelihood estimators of $\boldsymbol{\beta}$ for different smoothing bandwidths are depicted in Figure 2, showing that the profile-likelihood method does not provide a consistent estimator of β_1 at all. However our DBEs of $\boldsymbol{\beta}$ are $\hat{\beta}_1 = 0.9163(0.1651)$ and $\hat{\beta}_2 = 2.0037(0.0828)$, which are very close to their corresponding true values. This means that our difference-based method still performs very well. Hence our DBE of $\boldsymbol{\beta}$ is more robust to the smoothness assumption of $\boldsymbol{\alpha}(\cdot)$. For a random sample, we plot the estimated $\alpha_1(\cdot)$, $\alpha_2(\cdot)$, and $\sigma^2(\cdot)$ using two different methods in panels (B-D) of Figure 1. The RASE for estimating $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ is reported in Table 6. We can see that our new estimation scheme improves significantly 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 $\boldsymbol{\alpha}(\cdot)$, as studied in Fan and Zhang (1999).

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

6 Application to the Progesterone data

In this section, we apply the proposed methods to the longitudinal progesterone data.

For the i-th subject, denote x_i and z_i to be age and body mass index (both are stan-

dardized to have mean zero and standard deviation one); we consider as the response 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 following semi-parametric 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 $(1/\sqrt{2}, -1/\sqrt{2})$ as there is only one varying-coefficient term $f(t_{ij})$. DBE of (14) gives estimates $\hat{\beta}_1 = 0.0306$ and $\hat{\beta}_2 = 0.0195$. 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. The left panel of Figure 3 depicts the cross-validation score function defined as the sum of residual squares and suggests the optimal bandwidth 1.9349. The corresponding estimate $\hat{f}(\cdot)$ is plotted in the center panel of Figure 3 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 $\sigma^2(\cdot)$ and selects the bandwidth 2.5864. The resulting estimate of the variance function is shown in the right panel of Figure 3.

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 γ and ρ are 0.6900(0.3662) and 0.6452(0.1319), respectively. The numbers in parentheses are the corresponding standard errors, obtained using (13) based on estimated γ, ρ and $\varepsilon(t_{ij})$. This indicates a strong correlation structure.

We next consider testing the hypothesis $H_0 : \gamma = 1$ vs $H_1 : \gamma < 1$, *i.e.*, 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.

As the null hypothesis is not rejected, an AR(1) correlation structure $\text{corr}(\varepsilon(s), \varepsilon(t)) =$

$\rho^{|s-t|}$ is applied on this data. The QMLE $\hat{\rho}$ is 0.5049 with standard error given by 0.0831.

Via incorporating the estimate covariance structure in DBE, weighted least squares regression gives new estimates of $\boldsymbol{\beta}$: $\hat{\beta}_1 = -0.0059$ and $\hat{\beta}_2 = -0.0074$. The corresponding new plots of Figure 3 are very similar and thus not reproduced.

It is straightforward to understand how the estimated regression function and the estimated variance function affect the pointwise prediction as shown in 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 study this sensitivity using two randomly selected subjects. In order 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 the right column panels for the other one, where the same prediction formula of Fan et al. (2007) is used. In this prediction, the same estimated $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}(\cdot)$, and $\hat{\sigma}^2(\cdot)$ are used. From these figures, we can not see much change in the prediction for different correlation structure parameters and this suggests that the pointwise prediction is not sensitive to the correlation structure parameter.

Appendix

The following technical conditions are imposed.

(i) On the domain $[0, T]$, the density function $f(\cdot)$ is Lipschitz continuous and bounded away from zero. The kernel function $K(\cdot)$ is a symmetric density function with a compact support.

(ii) The moment generating function of J_i is finite near origin.

(iii) The estimator $\hat{m}(\cdot)$ of the mean function is consistent with polynomial convergence rate, *i.e.*, $\exists\ \tau > 1/4$ such that $m_0(\boldsymbol{x}) - \hat{m}(\boldsymbol{x}) = O_p(n^{-\tau})$ uniformly in $\boldsymbol{x}$.

(iv) It holds that $\sup|\hat{m}(\boldsymbol{x}) - \hat{m}_{-i}(\boldsymbol{x})| = O_p(n^{-\varsigma})$ uniformly in $\boldsymbol{x}$ 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.

(v) $\sigma_0^2(\cdot)$ has a continuous second derivative and is bounded away from zero in its domain. $E\varepsilon(t)^4 < \infty$.

(vi) The true parameter $\boldsymbol{\theta}_0$ of the correlation structure lies in the interior of a compact set Θ.

(vii) For any $\boldsymbol{\theta}$, $\frac{\partial}{\partial\theta_i}\rho(t,s,\boldsymbol{\theta})$ and $\frac{\partial^2}{\partial\theta_i\partial\theta_j}\rho(t,s,\boldsymbol{\theta})$ are bounded bivariate functions of t and s.

(viii) For any $\boldsymbol{\theta} \in \Theta$, it holds with probability one that the eigenvalues of the correlation matrix $\boldsymbol{C}(\boldsymbol{\theta})$ of a generic subject $\mathbb{X}$ are between ϱ_0 and ϱ_1, where $0 < \varrho_0 < \varrho_1 < \infty$.

(ix) Assume $E\log(g_0(\boldsymbol{\varepsilon};\boldsymbol{C}(\boldsymbol{\theta}_0))/f(\boldsymbol{\varepsilon};\boldsymbol{C}(\boldsymbol{\theta})))$ exists, where the expectation is taken with respect to the true elliptical density $g_0(\boldsymbol{\varepsilon};\boldsymbol{C}(\boldsymbol{\theta}_0))$ of the $\boldsymbol{\varepsilon}$, *i.e.*, $\boldsymbol{\varepsilon} \sim g_0(\boldsymbol{\varepsilon};\boldsymbol{C}(\boldsymbol{\theta}_0)) \propto |\boldsymbol{C}(\boldsymbol{\theta}_0)|^{-1/2}h_0(\boldsymbol{\varepsilon}^T\boldsymbol{C}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\varepsilon})$, and $f(\boldsymbol{\varepsilon};\boldsymbol{C}(\boldsymbol{\theta}))$ is corresponds to the density used in QMLE by treating $\boldsymbol{\varepsilon}$ normally distributed.

(x) Correlation structure is identifiable, *i.e.*, $\rho(s,t,\boldsymbol{\theta}_0) \neq \rho(s,t,\boldsymbol{\theta})$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ when $s \neq t$.

Remark 3 : Technical condition (iii) seems strong. However, once the form of the conditional mean function is available, it can be relaxed. For parametrical model $m(\boldsymbol{x}(t)) = \boldsymbol{x}(t)^T\boldsymbol{\beta}$, Condition (iii) can be replaced by $\sup_{t\in[0,T]} E \parallel \boldsymbol{x}(t) \parallel < \infty$ 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 $m(\boldsymbol{x}(t)) = \boldsymbol{x}(t)^T\boldsymbol{\alpha}(t)$, it can be replaced by $\sup_{t\in[0,T]} E \parallel \boldsymbol{x}(t) \parallel < \infty$ 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 $m(\boldsymbol{x}) = m(\boldsymbol{x})$ without any constraint, it is enough 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. Similar argument applies to Condition (iv).

Proof of Theorem 1: Note that Condition (ii) 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 $[t-h, t+h]$. So in the following, the $\varepsilon_i(t_{ij})$s can be treated as independent.

Define $e_{ij} = \hat{m}(\boldsymbol{x}_i(t_{ij})) - m(\boldsymbol{x}_i(t_{ij}))$. 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

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

Using Condition (iii), 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 $\mathrm{Var}(A_2) \le n^{-2\tau} \dfrac{\sum_{i=1}^n \sum_{j=1}^{J_i} \sigma^2(t_{ij}) K_h^2(t_{ij}-t)}{(\sum_{i=1}^n \sum_{j=1}^{J_i} K_h(t_{ij}-t))^2} \le n^{-2\tau}(\sup \sigma^2(t))/(nh)$ and

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

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

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

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

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

Using *Slutsky*'s Theorem, we have

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

This completes the proof of Theorem 1. ∎

We need the following observations and results to prove Theorem 2. Notice that $\zeta(t) = \varepsilon(t)/\sigma(t)$. Then $E\zeta(t) = 0$ and $\operatorname{Var}\zeta(t) = 1$, $\operatorname{cov}(\zeta_i(t_{ij}), \zeta_i(t_{ik})) = \rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$. 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. Then they can be decomposed as follows.

$$\begin{aligned}
\hat{\zeta}_i(t_{ij}) &= \frac{y_{ij} - \hat{m}(\boldsymbol{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} = \frac{m(\boldsymbol{x}_i(t_{ij})) - \hat{m}(\boldsymbol{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} + \frac{\sigma_0(t_{ij})\zeta_i(t_{ij})}{\hat{\sigma}(t_{ij})} \\
&= \zeta_i(t_{ij}) + \frac{m(\boldsymbol{x}_i(t_{ij})) - \hat{m}(\boldsymbol{x}_i(t_{ij}))}{\hat{\sigma}(t_{ij})} + \frac{[\sigma_0(t_{ij}) - \hat{\sigma}(t_{ij})]\zeta_i(t_{ij})}{\hat{\sigma}(t_{ij})}. \qquad (15)
\end{aligned}$$

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

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

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{m=1}^{n} \left\{ -\frac{1}{2} \log |\boldsymbol{C}(\boldsymbol{\theta}; m)| - \frac{1}{2} (\hat{\boldsymbol{\zeta}}_m)^T \boldsymbol{C}(\boldsymbol{\theta}; m)^{-1} \hat{\boldsymbol{\zeta}}_m \right\} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l_n(\boldsymbol{\theta}). \quad (16)$$

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

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

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

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

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

Noticing that the left hand side of the above equation is equal to

$$-\frac{1}{2}(\log |C(\boldsymbol{\theta}; i)| + tr[\boldsymbol{C}(\boldsymbol{\theta}; i)^{-1}\boldsymbol{C}(\boldsymbol{\theta}_0; i)]) + \frac{1}{2}(\log |\boldsymbol{C}(\boldsymbol{\theta}_0; i)| + tr[\boldsymbol{C}(\boldsymbol{\theta}_0; i)^{-1}\boldsymbol{C}(\boldsymbol{\theta}_0; i)]),$$

we have

$$\log |\boldsymbol{C}(\boldsymbol{\theta}_0; i)| + tr[\boldsymbol{C}(\boldsymbol{\theta}_0; i)^{-1}\boldsymbol{C}(\boldsymbol{\theta}_0; i)] \le \log |\boldsymbol{C}(\boldsymbol{\theta}; i)| + tr[\boldsymbol{C}(\boldsymbol{\theta}; i)^{-1}\boldsymbol{C}(\boldsymbol{\theta}_0; i)]$$

with equality holds only when $\boldsymbol{C}(\boldsymbol{\theta}; i) = \boldsymbol{C}(\boldsymbol{\theta}_0; i)$. So Lemma 1 is proved. ∎

Proof of Theorem 2: The log-likelihood function can be decomposed as follows:

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

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

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

Condition (v), $\max_{m=1}^{n} J_m = O(\log n)$, and the fact that $\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m \xrightarrow{P} \mathbf{0}$ at a polynomial rate imply that $E \parallel (\boldsymbol{\zeta}_m + \hat{\boldsymbol{\zeta}}_m) \parallel \cdot \parallel \hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m) \parallel \to 0$. As a result, the Law of Large Numbers implies that

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

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

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

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

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

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

is bounded by

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

whose expectation exists, is finite, and does not depend on $\boldsymbol{\theta}$. Hence, Condition (ix) and the Law of Large Numbers imply that, for each $\boldsymbol{\theta}$,

$$\begin{aligned} & -\frac{1}{2n} \sum_{m=1}^{n} \left\{ \log |\boldsymbol{C}(\boldsymbol{\theta}; m)| + tr[\boldsymbol{C}(\boldsymbol{\theta}; m)^{-1} \boldsymbol{\zeta}_m \boldsymbol{\zeta}_m^T] \right\} \\ \xrightarrow{P} \; & E_{g_0} \log f(\boldsymbol{\zeta}; \boldsymbol{C}(\boldsymbol{\theta})) \\ \equiv \; & -E_{g_0} \log \frac{g_0(\boldsymbol{\zeta}; \boldsymbol{C}(\boldsymbol{\theta}_0))}{f(\boldsymbol{\zeta}; \boldsymbol{C}(\boldsymbol{\theta}))} + E_{g_0} \log g_0(\boldsymbol{\zeta}; \boldsymbol{C}(\boldsymbol{\theta}_0)) \end{aligned} \tag{17}$$

(c.f. 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 $\boldsymbol{C}(\boldsymbol{\theta})$. Note that the continuity enforced by Condition (vii) on our correlation structure $\rho(\cdot,\cdot,\boldsymbol{\theta})$ implies that $E_{g_0} \log f(\boldsymbol{\zeta}; \boldsymbol{C}(\boldsymbol{\theta}))$ 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

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

in probability for large n. It follows that Θ may be covered by such neighborhoods and, since Θ is compact as enforced by Condition (vi), by a finite collection of such neighborhoods. As a result, the convergence of (17) is uniform with respect to $\boldsymbol{\theta}$ since the small number ϵ is arbitrary. Thus $\hat{\boldsymbol{\theta}}^{(1)}$ converges in probability to the minimizer of

the Kullback-Leibler Information Criterion

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

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

$$\begin{aligned} & E_{g_0}\left(\log(g_0(\boldsymbol{\zeta}_m; \boldsymbol{C}(\boldsymbol{\theta}_0; m))/f(\boldsymbol{\zeta}_m; \boldsymbol{C}(\boldsymbol{\theta}; m)))|J_m, \boldsymbol{T}_m\right) \\ = \; & \frac{1}{2}(\log|\boldsymbol{C}(\boldsymbol{\theta}; m)| + tr[\boldsymbol{C}(\boldsymbol{\theta}; m)^{-1}\boldsymbol{C}(\boldsymbol{\theta}_0; m)]) + constant \end{aligned} \tag{18}$$

has global minimizer $\boldsymbol{\theta}_0$ due to Lemma 1 and Condition (x).

Notice that

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

Hence $\boldsymbol{\theta}_0$ minimizes $I(g_0 : f, \boldsymbol{\theta})$ globally, which implies that $\hat{\boldsymbol{\theta}}^{(1)} \xrightarrow{P} \boldsymbol{\theta}_0$ due to (17). Theorem 1 is proved by noting that we have shown $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(1)} \xrightarrow{P} \mathbf{0}$. ∎

Proof of Theorem 3: By routine calculation as in the proof of Theorem 2, we have

$$\begin{aligned} \frac{1}{n}\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \; = \; & -\frac{1}{2n}\sum_{m=1}^{n}\left\{2\boldsymbol{\zeta}_m^T\boldsymbol{C}^i(\boldsymbol{\theta}; m)(\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m) + (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)^T\boldsymbol{C}^i(\boldsymbol{\theta}; m)(\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)\right\} \\ & -\frac{1}{2n}\sum_{m=1}^{n}\left\{tr[\boldsymbol{C}^{-1}(\boldsymbol{\theta}; m)\boldsymbol{C}_i(\boldsymbol{\theta}; m)] + tr[\boldsymbol{C}^i(\boldsymbol{\theta}; m)\boldsymbol{\zeta}_m\boldsymbol{\zeta}_m^T]\right\}. \end{aligned} \tag{19}$$

Note that $\boldsymbol{C}^i(\boldsymbol{\theta}, m) = \boldsymbol{C}^{-1}(\boldsymbol{\theta}, m)\boldsymbol{C}_i(\boldsymbol{\theta}, m)\boldsymbol{C}^{-1}(\boldsymbol{\theta}, m)$ and $\boldsymbol{\zeta}_m$ has mean zero and finite variance. The uniform convergence of $m(\cdot)$ and $\sigma^2(\cdot)$ in the decomposition (15) implies that the first term on the right hand side of (19) converges to zero in probability. As a

result, we have

$$\frac{1}{n}\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - \left[-\frac{1}{2n}\sum_{m=1}^{n}\left\{tr[\boldsymbol{C}^{-1}(\boldsymbol{\theta};m)\boldsymbol{C}_i(\boldsymbol{\theta};m)] + tr[\boldsymbol{C}^{i}(\boldsymbol{\theta};m)\boldsymbol{\zeta}_m\boldsymbol{\zeta}_m^T]\right\}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \xrightarrow{P} 0. \tag{20}$$

Hence

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

Similarly, we can show that

$$\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_i\partial\theta_j}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \xrightarrow{P} -\frac{1}{2}E\left\{tr[\boldsymbol{C}^{-1}(\boldsymbol{\theta}_0;m)\boldsymbol{C}_i(\boldsymbol{\theta}_0;m)\boldsymbol{C}^{-1}(\boldsymbol{\theta}_0;m)\boldsymbol{C}_j(\boldsymbol{\theta}_0;m)]\right\},$$

and $\frac{1}{n}\frac{\partial^3 l_n(\boldsymbol{\theta})}{\partial \theta_i\partial\theta_j\partial\theta_k}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ converges in probability to

$$\begin{aligned}-\frac{1}{2}E\{&tr[\boldsymbol{C}_{ij}\boldsymbol{C}^{-1}\boldsymbol{C}_k\boldsymbol{C}^{-1} + \boldsymbol{C}_{jk}\boldsymbol{C}^{-1}\boldsymbol{C}_i\boldsymbol{C}^{-1} + \boldsymbol{C}_j\boldsymbol{C}^{-1}\boldsymbol{C}_{ik}\boldsymbol{C}^{-1}\\ &-2\boldsymbol{C}_i\boldsymbol{C}^{-1}\boldsymbol{C}_j\boldsymbol{C}^{-1}\boldsymbol{C}_k\boldsymbol{C}^{-1} - 2\boldsymbol{C}_j\boldsymbol{C}^{-1}\boldsymbol{C}_i\boldsymbol{C}^{-1}\boldsymbol{C}_k\boldsymbol{C}^{-1}]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\}.\end{aligned}$$

The existence and boundedness of the above expectations can easily be proved by using Conditions (ii), (vii), and (viii).

Notice that $\boldsymbol{I}(\boldsymbol{\theta}_0) = \lim_{n\to 0}(-\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0})$ is a d by d matrix whose (i,j)-element is the limit of $-\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ as $n\to\infty$. Taylor expand $\frac{\partial}{\partial\boldsymbol{\theta}}l_n(\hat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}_0$, *i.e.*,

$$\mathbf{0} = \frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}}l_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}}l_n(\boldsymbol{\theta}_0) + \frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \boldsymbol{R}_n(\boldsymbol{\theta}^*), \tag{21}$$

where $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$ and $\boldsymbol{R}_n(\boldsymbol{\theta}) = \boldsymbol{M}(\boldsymbol{\theta})((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, where $\boldsymbol{M}(\boldsymbol{\theta})$ is a d-by-d matrix whose m-th row is given by

$$\frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T[\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\frac{\partial}{\partial\theta_m}l_n(\boldsymbol{\theta}))].$$

So we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -(\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} + \boldsymbol{M}(\boldsymbol{\theta}^*))^{-1}(\frac{\sqrt{n}}{n}\frac{\partial}{\partial\boldsymbol{\theta}}l_n(\boldsymbol{\theta}_0)).$$

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

Notice that, from above, we have $\frac{1}{n}\frac{\partial l_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\ |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \rightarrow \mathbf{0}$ and $-\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \rightarrow \boldsymbol{I}(\boldsymbol{\theta}_0)$ as $n \rightarrow \infty$. To get the desired asymptotic normality result we need to show that $-\frac{1}{2n}\sum_{m=1}^{n}\left\{2\boldsymbol{\zeta}_m^T\boldsymbol{C}^i(\boldsymbol{\theta};m)(\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m) + (\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)^T\boldsymbol{C}^i(\boldsymbol{\theta};m)(\hat{\boldsymbol{\zeta}}_m - \boldsymbol{\zeta}_m)\right\} = o_p(n^{-1/2})$. The second term can be easily shown to be of order $o_p(n^{-1/2})$ due to Conditions (iii), (vii), (viii) and the result of Theorem 1. Denote $\tilde{\boldsymbol{C}}^i(\boldsymbol{\theta};m)$ to be a $(\max_{j=1}^{n} J_j)$-by-$(\max_{j=1}^{n} J_j)$ matrix whose top-left J_m-by-J_m submatrix is $\boldsymbol{C}^i(\boldsymbol{\theta};m)$ and other elements are all filled by zeros. Similarly, $\widetilde{\hat{\boldsymbol{\zeta}}_m}$ and $\tilde{\boldsymbol{\zeta}}_m$ are vectors of length $(\max_{i=1}^{n} J_i)$ with first J_m elements given by $\hat{\boldsymbol{\zeta}}_m$ and $\boldsymbol{\zeta}_m$, respectively, and other elements are filled by zeros. Then the first term is equal to $-\frac{1}{n}\sum_{m=1}^{n}\tilde{\boldsymbol{\zeta}}_m^T\tilde{\boldsymbol{C}}^i(\boldsymbol{\theta};m)(\widetilde{\hat{\boldsymbol{\zeta}}_m} - \tilde{\boldsymbol{\zeta}}_m)$, which can be shown to be of order $o_p(n^{-1/2})$ by using techniques similar to the proof of Lemma 2 in Lam and Fan (2008). Applying the Central Limit Theorem, we have that

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

Proof of Proposition 1: Part [a] of Proposition 1 can be verified using basic algebra and is skipped here.

Next we prove part [b]. With the assumption that $\min_{i=2,3,\cdots,K}(s_i - s_{i-1}) = s_0 > 0$ and based on part [a], we can easily show that $\boldsymbol{A}^{-1}$ is diagonal dominated. The basic

properties of hyperbolic function implies that

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

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

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

Noticing that $\boldsymbol{A}^{-1}$ is symmetric, we have

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

This immediately implies that the smallest eigenvalue of $\boldsymbol{A}^{-1}$ is not smaller than δ_0 which does not depend on the number J of observations. Similarly we can show that the largest eigenvalue of $\boldsymbol{A}^{-1}$ is at most $2\max_{i=1}^{J}(\boldsymbol{A}^{-1})_{ii} \leq \delta_1(t_0, a)$. ∎

Proof of Proposition 2: The i-th subject has observations at times $t = t_{i1}, t_{i2}, \cdots, t_{iJ_i}$, 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

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

So part A is proved.

Part [B] can be easily proved by noticing that the correlation matrix for ARMA(1, 1) model is exactly $(1-\gamma)\boldsymbol{I}+\gamma\boldsymbol{D}$, where $\boldsymbol{D}$ is the corresponding correlation matrix of AR(1) with the same parameter φ and $\boldsymbol{I}$ is the identity matrix.

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

References

Bickel, P. and Levina, E. (2006). Regularized estimation of large covariance matrices. Technical report #716, Dept. of Statistics, UC Berkeley.

Brown, L. D. and Levine, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *The Annals of Statistics*. To appear.

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). *Analysis of longitudinal data*. 2nd ed. Oxford University Press, USA.

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *Journal of American Statistical Association*, **96** 640–652.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semi-parametric varying-coefficient partially linear models. *Bernoulli*, **11** 1031–1059.

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

Fan, J. and 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 Series B*, **62** 303–322.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying-coefficient models. *The Annals of Statistics*, **27** 1491–1518.

Hall, P., Kay, J. W. and Titterington, 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, B*, **55** 757–796.

Huang, J. Z., Liu, L. and Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis fuction approximations. *Journal of Computational and Graphical statistcs*. To appear.

Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *The Annals of Statistics*. To appear.

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 Royal Statistical Society Series B*, **53** 233–243.

Rothman, A. J., Bickel, P., Levina, L. and Zhu, J. (2007). Sparse permutation invariant covariance estimation. Technical report.

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 Americal 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.

Yao, F., Müller, H.-G. and Wang, J.-L. (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 stochasitc mixed models for longitudinal data. *Journal of the American Statistical Association*, **93** 710–719.

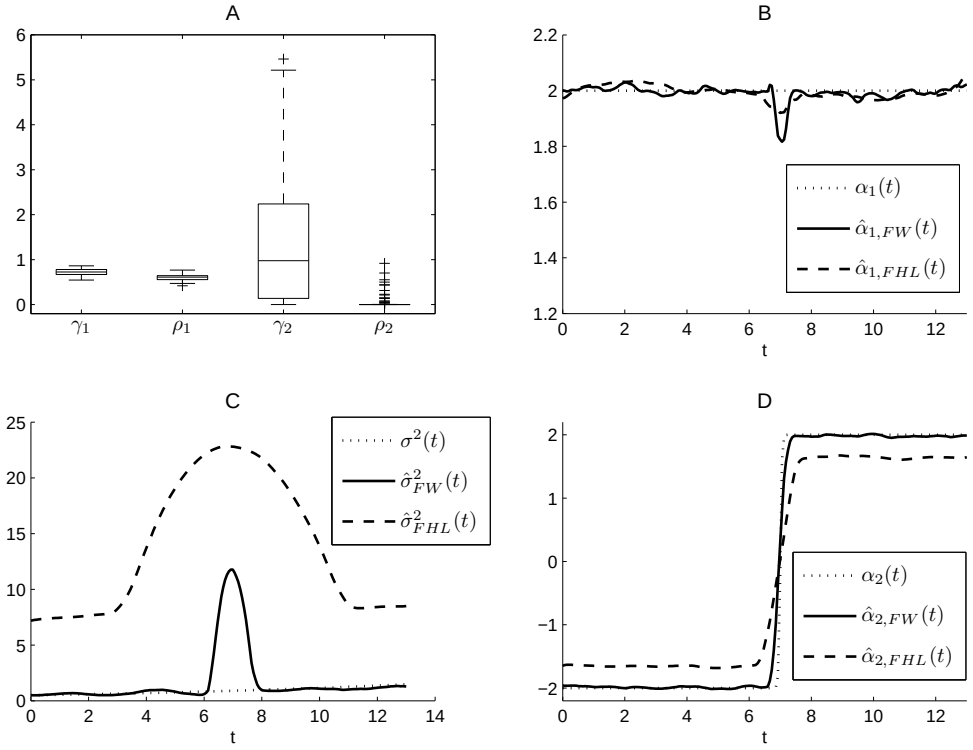

Figure 1: For Example 5.3, panel (A) gives the box-plot of our estimator (γ_1 and ρ_1) and Fan et al. (2007)'s estimator (γ_2 and ρ_2) of the correlation structure parameter; panels (B-D) plot the true (dotted line), our new estimate (solid line), and Fan et al. (2007)'s estimate (dashed line) of $\alpha_1(\cdot)$, $\sigma^2(\cdot)$, and $\alpha_2(\cdot)$, respectively, for one typical sample.

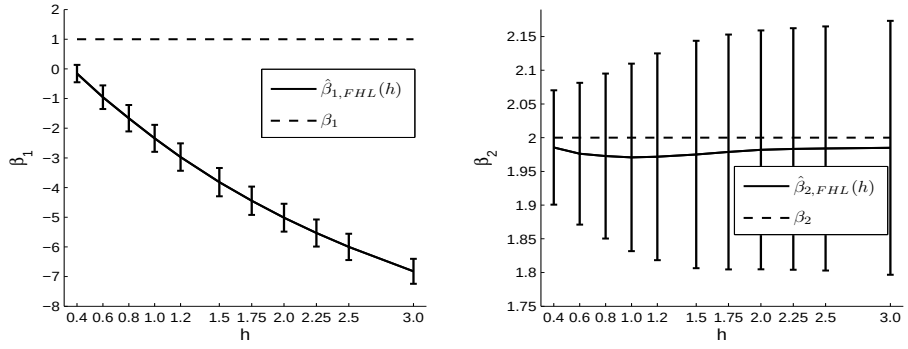

Figure 2: The solid curves (average estimate) with error bars (one sample standard deviation) in the left and right panels plot the profile-likelihood estimators of the regression coefficient β_1 and β_2 respectively for different smoothing bandwidth h. The dashed curves denote their corresponding true coefficients.

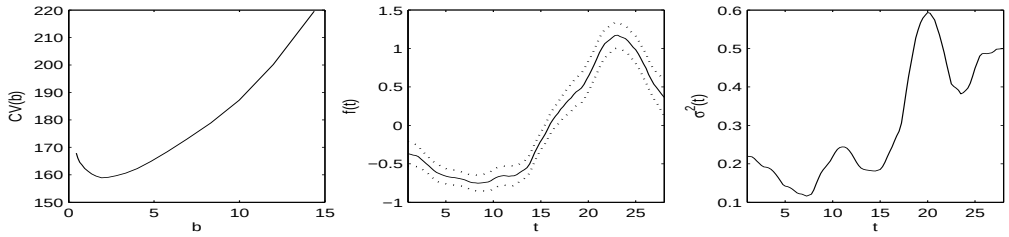

Figure 3: The left panel plots the leave-one-subject-out cross-validation score against the bandwidth. The center panel depicts the estimated varying-coefficient function $f(\cdot)$ with a pointwise 95% confidence interval (dotted line). The right panel gives the estimated variance function $\sigma^2(\cdot)$.

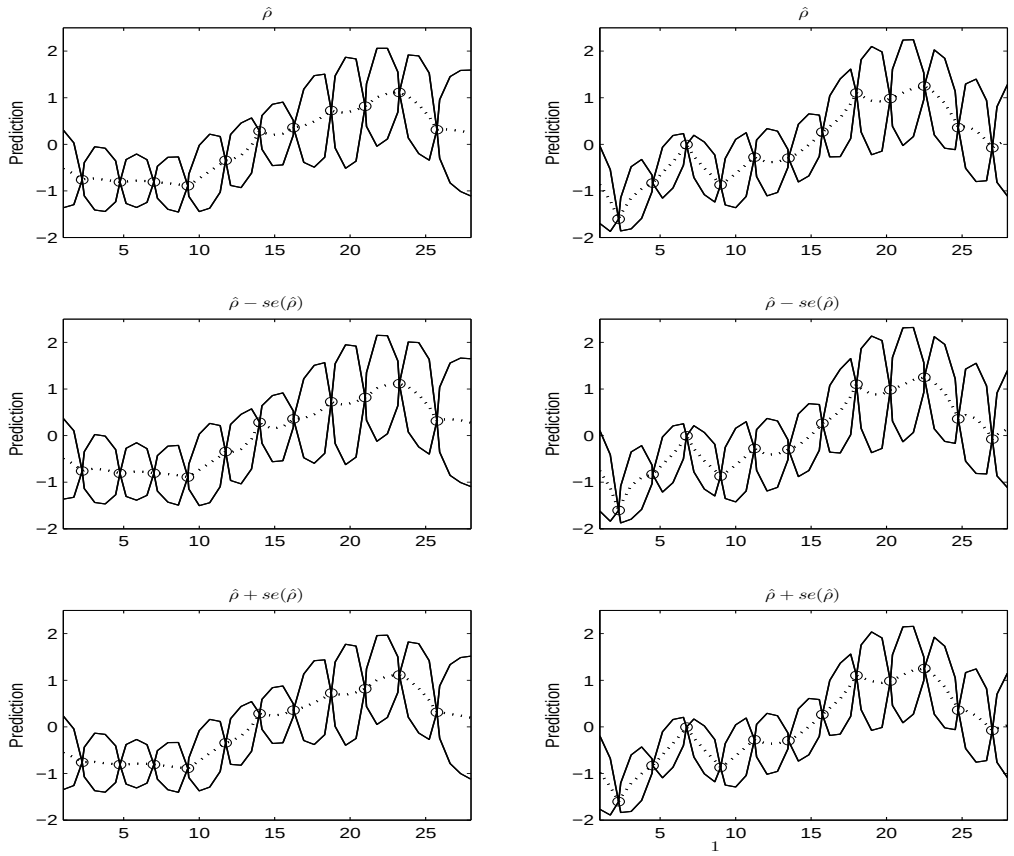

Figure 4: The observed values (circle), the corresponding pointwise predictions (dotted line), and 95% predictive intervals (solid line) are plotted for two randomly selected subjects, one in the left three panels and the other in the right three panels. For each column, the top, middle, and bottom panels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