Semiparametric Mean-Covariance Regression Analysis for Longitudinal Data

July 14, 2009

Acknowledgments

We would like to thank the Editor, an associate editor and two referees for all the constructive comments and detailed suggestions, which have led to a substantially improved paper.

Abstract

Efficient estimation of the regression coefficients in longitudinal data analysis requires a correct specification of the covariance structure. Existing approaches usually focus on modeling the mean with specification of certain covariance structures, which may lead to inefficient or biased estimators of parameters in the mean if misspecification occurs. In this paper, we propose a data-driven approach based on semiparametric regression models for the mean and the covariance simultaneously, motivated by the modified Cholesky decomposition. A regression spline based approach using generalized estimating equations is developed to estimate the parameters in the mean and the covariance. The resulting estimators for the regression coefficients in both the mean and the covariance are shown to be consistent and asymptotically normally distributed. In addition, the nonparametric functions in these two structures are estimated at their optimal rate of convergence. Simulation studies and a real data analysis show that the proposed approach yields highly efficient estimators for the parameters in the mean, and provides parsimonious estimation for the covariance structure.

Some Keywords: Covariance misspecification; Efficiency; Generalized estimating equation; Longitudinal data; Modified Cholesky decomposition; Semiparametric models.

1 Background

Longitudinal data arise frequently in the biomedical, epidemiological, social and economical fields. A salient feature of longitudinal studies is that subjects are measured repeatedly over time. Thus, observations for the same subject are intrinsically correlated. Regression methods for such data sets accounting for within-subject correlation is abundant in the literature (Diggle et al., 2002; Wu and Zhang, 2006).

Within the framework of generalized linear models (GLM), the technique of generalized estimating equations (GEE, Liang and Zeger, 1986) is widely used for dealing with longitudinal data. GEE makes the use of a working correlation model to estimate the mean parameters in the marginal specification of the regression. Although consistency of the mean parameter estimators is not affected, misspecification of the correlation may result in a great loss of efficiency (Wang and Carey, 2003). On the other hand, the correlation structure itself may be of scientific interest (Prentice and Zhao, 1991; Diggle and Verbyla, 1998). Therefore, there is a great need to model the covariance structure. However, modeling the correlation matrix is more challenging than modeling the mean as there are usually much more parameters in the former and the positive definiteness of the covariance matrix has to be assured. Recently, Pourahmadi (1999, 2000) proposed a modified Cholesky decomposition to decompose the covariance matrix. This decomposition is attractive as it leads automatically to positive definite covariance matrices, and the parameters in it are related to well founded statistical concepts. Thereafter, the parameters in this decomposition can be modeled via regression techniques, enabling model based inference for the parameters in the mean and the covariances. See Pan and MacKenzie (2003) for a related discussion. More recently, Ye and Pan (2006) further proposed to use GEE to model the parameters in this decomposition by several sets of parametric estimating equations.

There is a clear need to relax the parametric assumption posed in Ye and Pan (2006), as model misspecification may result in biased estimation, a problem even more severe than misspecification of the covariance. An attractive approach is the semiparametric regression model, or the partly linear model (PLM), which retains the flexibility of the nonparametric model but avoids the need to model a fully nonparametric model (Härdle, Liang and Gao, 2000). Existing applications of PLM

to longitudinal data accounting for within-subject dependence mainly focus on regression analysis of the mean. The covariance is usually assumed known up to a few parameters. See for example, the kernel and spline approaches in Lin and Carroll (2001), Welch, Lin and Carroll (2002), He, Zhu and Fung (2002), Wang (2003), Wang, Carroll and Lin (2005), He, Fung and Zhu (2005), and Lin and Carroll (2006).

Compared to the models for the mean in longitudinal data analysis, model based analysis for the covariance is much less studied. To address this issue, we propose semiparametric models for the mean and the covariance structure for longitudinal data. Our formulation builds on the modified Cholesky decomposition advocated by Pourahmadi (1999) such that the entries in this decomposition can be modeled by unconstrained semiparametric regression models. Our approach is more flexible than the model in Ye and Pan (2006) in both the mean and the covariance. There are also related works in this regard. Fan, Fan and Lv (2008) studied a factor model method to parametrize a high dimensional covariance matrix. Wu and Pourahmadi (2003) proposed nonparametric estimates of the covariance matrix, but their method does not deal with irregular observed measurements and only models a few of the subdiagonals of the lower triangular matrix in the modified Cholesky decomposition. Fan, Huang and Li (2007) and Fan and Wu (2008) studied a different semiparametric model for the covariance structure by imposing the familiar variance-correlation decomposition. They estimated the marginal variance via kernel smoothing and proposed a parametric model for the correlation matrix. Similar to the method in Fan, Huang and Li (2007), our approach can handle irregularly and possibly subject-specific times points. We show that the resulting estimators for the regression coefficients in the mean and the variance are consistent and asymptotically normally distributed. Furthermore, the nonparametric parts are estimated at the optimal convergence rate by

using regression splines.

The rest of the paper is organized as follows. Section 2 introduces the models and estimation methods. Theoretical properties of the proposed estimators are given in Section 3. Extensive simulations and data analysis are presented in Section 4. Section 5 gives some concluding remarks. All the proofs are relegated to the Appendix.

2 The Models and The Estimation Methods

2.1 The models

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})'$ be the n_i repeated measurements at time points $\mathbf{t}_i = (t_{i1}, \cdots, t_{in_i})'$ of the ith subject $(i = 1, \cdots, m)$. Here t_{ij} may be the time or any time-dependent covariate which is modeled nonparametrically. Without loss of generality, we assume that all $\{t_{ij}\}$ are scaled into the interval $[0, 1]$. Furthermore, we assume that the first two moments of the response satisfy $E(y_{ij}|\mathbf{x}_{ij}, t_{ij}) = \mu_{ij}^0$ and $V(\mathbf{y}_i|\mathbf{x}_i, \mathbf{t}_i) = \mathbf{\Sigma}_{0i}$, where $\mathbf{x}_{ij}$ is a p-vector covariate and $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{in_i})'$ is the covariate matrix for the ith subject. To guarantee the positive definiteness of the matrices $\mathbf{\Sigma}_{0i}$, an explicit way of modeling $\mathbf{\Sigma}_{0i}$ is via its modified Cholesky decomposition as $\mathbf{\Phi}_i \mathbf{\Sigma}_{0i} \mathbf{\Phi}_i' = \mathbf{D}_{0i}$, where $\mathbf{\Phi}_i$ is a lower triangular matrix with 1's on its diagonal, and $\mathbf{D}_{0i}$ is a diagonal matrix. As indicated by Pourahmadi (1999), this decomposition has a clear statistical interpretation. The below-diagonal entries of $\mathbf{\Phi}_i$ are the negatives of the autoregressive coefficients ϕ_{ijk} defined in

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}). \tag{1}$$

That is, the autoregressive coefficients are the population regression coefficients of the linear regression of y_{ij} on its predecessors $y_{i(j-1)}, \cdots, y_{i1}$. The diagonal entries σ_{0ij}^2 of $\mathbf{D}_{0i}$ can be seen as the innovation variance $\sigma_{0ij}^2 = var(\epsilon_{ij})$, for $\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$. Clearly, the modified Cholesky decomposition is advantageous in that ϕ and $\log(\sigma^2)$

are unconstrained, rather than a constrained parameter $\mathbf{\Sigma}_{0i}$ that must be positive definite. To use the semiparametric regression tools, we postulate three sets of models

$$g(\mu_{ij}^0) = \mathbf{x}_{ij}'\boldsymbol{\beta}_0 + f_0(t_{ij}), \qquad \phi_{ijk} = \mathbf{w}_{ijk}'\boldsymbol{\gamma}_0, \qquad \log(\sigma_{0ij}^2) = \mathbf{z}_{ij}'\boldsymbol{\lambda}_0 + f_1(t_{ij}), \qquad (2)$$

where $\mathbf{x}_{ij}, \mathbf{w}_{ijk}$ and $\mathbf{z}_{ij}$ are the $p \times 1$, $q \times 1$ and $d \times 1$ vectors of covariates respectively; $\boldsymbol{\beta}_0$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\lambda}_0$ are the regression coefficients; $f_0(\cdot)$ and $f_1(\cdot)$ are unknown smooth functions. The known link function $g(\cdot)$ is assumed to be monotone and differentiable. The covariates $\mathbf{x}_{ij}, \mathbf{w}_{ijk}$ and $\mathbf{z}_{ij}$ may contain the baseline covariates, the time and the associated interactions. The idea of using (2) reflects the belief that regression models for the autoregressive coefficients and innovation variances are as important as those for the mean. Furthermore, model based analysis of these parameters permit more accessible statistical inference. This technique was also used by Ye and Pan (2006), but they assumed parametric models for $g(\mu_{ij}^0)$ and $\log(\sigma_{0ij}^2)$. As discussed earlier, the semiparametric estimating equations are more flexible and can be less biased if the parametric assumption is violated.

It is interesting to make some comparisons with the approach in Fan et al. (2007). They adopted the variance-correlation decomposition by using a well-behaved matrix such as AR(1) or ARMA(1,1) as the correlation matrix to ensure positive definiteness of the covariance. We use the modified Cholesky decomposition, which allows unconstrained parametrization. Thus regression based analysis for entries in this decomposition are possible. These two approaches depend on different ways to decompose covariance and each has its own merits.

2.2 The estimating equations

The two nonparametric functions f_0 and f_1 are parametrized by regression splines, because splines can provide optimal rates of convergence for both the parametric

and the nonparametric components in PLM with a small number of knots (Heckman 1986; He and Shi 1996, He, et al., 2002). Additionally, any computational algorithm developed for GLM can be used for fitting a semiparametric extension of GLM, since they treat a nonparametric function as a linear function with the basis functions as covariates. For simplicity, we assume that f_0 and f_1 have the same smoothness property. Let $0 = s_0 < s_1 < \cdots < s_{k_n} < s_{k_n+1} = 1$ be a partition of the interval $[0, 1]$. Using $\{s_i\}$ as the internal knots, we have $K = k_n + l$ normalized B-spline basis functions of order l that form a basis for the linear spline space. We use the B-spline basis functions because they have bounded support and are numerically stable (Schumaker, 1981). Thus $f_0(t)$ and $f_1(t)$ are approximated by $\boldsymbol{\pi}'(t)\boldsymbol{\alpha}$ and $\boldsymbol{\pi}'(t)\tilde{\boldsymbol{\alpha}}$ respectively, where $\boldsymbol{\pi}(t) = (B_1(t), \cdots, B_K(t))'$ is the vector of basis functions and $\boldsymbol{\alpha}, \tilde{\boldsymbol{\alpha}} \in \mathbb{R}^K$. Let $\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}(t_{ij})$. With this notation, the nonlinear regression models in (2) can be linearized as following:

$$g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{\pi}'(t_{ij})\boldsymbol{\alpha} = \mathbf{b}'_{ij}\boldsymbol{\theta}, \qquad \log(\sigma^2_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\lambda} + \boldsymbol{\pi}'(t_{ij})\tilde{\boldsymbol{\alpha}} = \mathbf{h}'_{ij}\boldsymbol{\rho}, \tag{3}$$

where $\mathbf{b}'_{ij} = (\mathbf{x}'_{ij}, \boldsymbol{\pi}'_{ij})$, $\mathbf{h}'_{ij} = (\mathbf{z}'_{ij}, \boldsymbol{\pi}'_{ij})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and $\boldsymbol{\rho} = (\boldsymbol{\lambda}', \tilde{\boldsymbol{\alpha}}')'$. We then let $\boldsymbol{\mu}_i = (\mu_{i1}, \cdots, \mu_{in_i})'$, $\mathbf{B}_i = (b_{i1}, \cdots, b_{in_i})'$ and define $\mathbf{x}_i$, $\boldsymbol{\pi}_i$, $\mathbf{z}_i$ and $\mathbf{H}_i$ in a similar fashion. Throughout this paper, a scalar function acting on a vector is set to be the vector of the function on each component, for example, $g(\boldsymbol{\mu}_i) = (g(\mu_{i1}), \cdots, g(\mu_{in_i}))'$. Using the GEE method from Liang and Zeger (1986), we construct the estimating equations for $\boldsymbol{\theta}, \boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ as follows

$$\begin{aligned} &\mathbf{S}_1(\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbf{B}'_i\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mu_i(\mathbf{B}_i\boldsymbol{\theta})) = 0,\ \mathbf{S}_2(\boldsymbol{\gamma}) = \sum_{i=1}^{m} \mathbf{V}'_i\mathbf{D}_i^{-1}(\mathbf{r}_i - \hat{\mathbf{r}}_i) = 0, \\ &\mathbf{S}_3(\boldsymbol{\rho}) = \sum_{i=1}^{m} \mathbf{H}'_i\mathbf{D}_i\mathbf{W}_i^{-1}(\boldsymbol{\epsilon}_i^2 - \sigma_i^2(\mathbf{H}_i\boldsymbol{\rho})) = 0, \end{aligned} \tag{4}$$

where $\boldsymbol{\Delta}_i = \boldsymbol{\Delta}_i(\mathbf{B}_i\boldsymbol{\theta}) = diag\{\dot{g}^{-1}(\mathbf{b}'_{ij}\boldsymbol{\theta}), \cdots, \dot{g}^{-1}(\mathbf{b}'_{in_i}\boldsymbol{\theta})\}$ and $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse function $g^{-1}(\cdot)$; $\mathbf{r}_i$ and $\hat{\mathbf{r}}_i$ are the $n_i \times 1$ vectors with jth components $r_{ij} =$

$y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = E(r_{ij}|r_{i1}, \cdots, r_{i(j-1)}) = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$ $(j = 1, \cdots, n_i)$. Note that when $j = 1$ the notation $\sum_{k=1}^{0}$ means zero throughout this paper. It can be shown that $\mathbf{D}_i = \text{diag}\{\sigma_{i1}^2, \cdots, \sigma_{in_i}^2\}$ in $\mathbf{S}_2(\boldsymbol{\gamma})$ is actually the covariance matrix of $\mathbf{r}_i - \hat{\mathbf{r}}_i$ and that $\mathbf{V}_i' = \partial \hat{\mathbf{r}}_i' / \partial \boldsymbol{\gamma}$ is the $q \times n_i$ matrix with jth column $\partial \hat{r}_{ij} / \partial \boldsymbol{\gamma} = \sum_{k=1}^{j-1} r_{ik} w_{ijk}$. On the other hand, $\boldsymbol{\epsilon}_i^2$ and $\boldsymbol{\sigma}_i^2$ in $\mathbf{S}_3(\boldsymbol{\lambda})$ are the $n_i \times 1$ vectors with jth components ϵ_{ij}^2 and σ_{ij}^2 $(j = 1, \cdots, n_i)$, respectively, where $\epsilon_{ij} = y_{ij} - \hat{y}_{ij}$ and $\hat{y}_{ij}$ are given in (1). Obviously, we have $E(\boldsymbol{\epsilon}_i^2) = \boldsymbol{\sigma}_i^2$. In addition, $\mathbf{W}_i$ is the covariance matrix of $\boldsymbol{\epsilon}_i^2$, that is, $\mathbf{W}_i = V(\boldsymbol{\epsilon}_i^2)$. The solutions of these generalized estimating equations, $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\rho}}$ say, are termed the GEE estimators of $\boldsymbol{\theta}, \boldsymbol{\gamma}$ and $\boldsymbol{\rho}$. As suggested by Ye and Pan (2006), a sandwich 'working' covariance structure $\mathbf{W}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\delta}) \mathbf{A}_i^{1/2}$ can be used to approximate the true $\mathbf{W}_i$'s, where $\mathbf{A}_i = 2\text{diag}\{\sigma_{i1}^4, \cdots, \sigma_{in_i}^4\}$ and $\mathbf{R}_i(\boldsymbol{\delta})$ mimics the correlation between ϵ_{ij}^2 and ϵ_{ik}^2 $(j \neq k)$ by introducing a new parameter $\boldsymbol{\delta}$. Typical structures for $\mathbf{R}_i(\boldsymbol{\delta})$ include compound symmetry (exchangeable) and $AR(1)$. As with the conventional generalized estimating equations for the mean, the parameter $\boldsymbol{\delta}$ may have very little effect on the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$. Our real data analysis and simulation studies reported in later sections confirm this point very well.

The three GEE equations in (4) can be seen as a generalization of the conventional GEE for the mean parameters. If we use a working covariance structure for $\boldsymbol{\Sigma}_i$ in $\mathbf{S}_1(\boldsymbol{\theta})$ and ignore $\mathbf{S}_2(\boldsymbol{\gamma})$ and $\mathbf{S}_3(\boldsymbol{\rho})$, we have the PLM for the mean. The modified Cholesky decomposition allows us to proceed a step further to impose partly linear structures for the variance components.

2.3 The main algorithm

The solutions of $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$ satisfy the equations in (4). These parameters can be solved iteratively by fixing the other parameters. For example, for fixed values of

$\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$ can be computed via the third equation in (4). An application of the quasi-Fisher scoring algorithm on equation (4) directly yields the numerical solutions for these parameters. More specifically, given $\boldsymbol{\Sigma}_i$, $\boldsymbol{\theta}$ can be updated by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left\{ \left[\sum_{i=1}^{m} \mathbf{B}_i' \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_i \mathbf{B}_i \right]^{-1} \sum_{i=1}^{m} \mathbf{B}_i' \boldsymbol{\Delta}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\mathbf{B}_i \boldsymbol{\theta})) \right\} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}. \quad (5)$$

Given $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$, $\boldsymbol{\gamma}$ can be updated approximately through

$$\boldsymbol{\gamma}^{(k+1)} = \left\{ \left[E \sum_{i=1}^{m} \mathbf{V}_i' \mathbf{D}_i^{-1} \mathbf{V}_i \right]^{-1} \sum_{i=1}^{m} \mathbf{V}_i' \mathbf{D}_i^{-1} \mathbf{r}_i \right\} \Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(k)}}. \quad (6)$$

In this update, the true errors are replaced by the residuals. We remark that this replacement has minimal effects on the estimation if the mean model is correctly specified. However, if the mean model is misspecified, this replacement would normally give inconsistent estimators for the autoregressive and log innovation parameters. This argument, on the other hand, demonstrates that the partly linear mean model of our approach is at least more robust than a parametric mean model with respect to model misspecification. Finally, given $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, the innovation variance parameters $\boldsymbol{\rho}$ can be updated using

$$\boldsymbol{\rho}^{(k+1)} = \boldsymbol{\rho}^{(k)} + \left\{ \left[\sum_{i=1}^{m} \mathbf{H}_i' \mathbf{D}_i \mathbf{W}_i^{-1} \mathbf{D}_i \mathbf{H}_i \right]^{-1} \sum_{i=1}^{m} \mathbf{H}_i' \mathbf{D}_i \mathbf{W}_i^{-1} (\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_i^2) \right\} \Big|_{\boldsymbol{\rho}=\boldsymbol{\rho}^{(k)}}, \quad (7)$$

Equation (5)-(7) indicate that, iteratively, the parameters can be estimated using weighted generalized least squares. We summarize the algorithm as follows

1. Initialization step: given a starting value $(\boldsymbol{\theta}^{(0)'}, \boldsymbol{\gamma}^{(0)'}, \boldsymbol{\rho}^{(0)'})'$, use the model (2) to form the lower triangular matrices $\mathbf{T}_i^{(0)}$ and diagonal matrices $\mathbf{D}_i^{(0)}$. Set $k = 0$ to obtain $\boldsymbol{\Sigma}_i^{(0)}$, the starting values of $\boldsymbol{\Sigma}_i$;

2. Iteration step: use $(5) - (7)$ to calculate $\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}$ and $\boldsymbol{\rho}^{(k+1)}$;

3. Updating step: replace $\boldsymbol{\theta}^{(k)}, \boldsymbol{\gamma}^{(k)}$ and $\boldsymbol{\rho}^{(k)}$ with the estimators $\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\gamma}^{(k+1)}$ and $\boldsymbol{\rho}^{(k+1)}$. Repeat Steps 2-3 until a desired convergence criterion is met.

A good starting value of $\boldsymbol{\Sigma}_i^{(0)}$ can be simply chosen as $\mathbf{I}_i$, the identity matrix for the ith subject. This initial value of $\boldsymbol{\Sigma}_i$ guarantees the consistency of the initial estimators in the mean, which in return guarantees consistency of the autoregressive parameters and innovative parameters after the first iteration. In the analysis presented in this paper, the convergence criterion is met as long as the successive difference in the Euclidean norm is less than 10^{-6}. Our numerical experience shows that this iterative algorithm converges very quickly, usually in a few iterations.

2.4 Knot selection

Knot selection is an important issue in spline smoothing. The number of knots plays the same role as the smoothing parameter in the smoothing spline model and the bandwidth parameter in kernel smoothing. In this article, we follow the spline literature (He, et al. 2005) and use the sample quartiles of $\{t_{ij}, i = 1, \cdots, m, j = 1, \cdots, n_i\}$ as knots. For example, if we use three internal knots, they are taken to be the three quartiles of the observed $\{t_{ij}\}$. We use cubic splines (splines of order 4) in the numerical simulation section, and the number of internal knots is taken to be the integer part of $n^{1/5}$, where n is the sample size. This particular choice is consistent with the asymptotic theory of Section 3. According to our empirical experience, it works well in a wide variety of problems. A data adaptive procedure is to use the leave-one-subject-out cross validation method, which is computationally more demanding. Theoretical justification of the leave-one-subject-out cross validation is possible but is beyond the scope of the paper.

3 Asymptotic Properties

Here and throughout, $\|\cdot\|$ for a vector denotes its Euclidean norm, and for any square matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes its modulus of the largest singular value of $\mathbf{A}$. To study the rates of convergence for $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}}$ and $\hat{f}_0, \hat{f}_1$, we first give a set of regularity conditions. If the estimating equation (4) has multiple solutions, then only a sequence of consistent estimator $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}})$ is considered in this section. A sequence $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\rho}})$ is said to be a consistent sequence if $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\lambda}}')' \to (\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0', \boldsymbol{\lambda}_0')'$ and $\sup_t |\boldsymbol{\pi}'(t)\hat{\boldsymbol{\alpha}} - f_0(t)| \to 0$, $\sup_t |\tilde{\boldsymbol{\pi}}'(t)\hat{\tilde{\boldsymbol{\alpha}}} - f_1(t)| \to 0$ in probability as $m \to \infty$. The fact that our iterative algorithm starts from consistent estimators of the parameters ensures that the final estimators are also consistent. Our basic conditions are as follows:

(A1) The dimensions p, q and d of covariates $\mathbf{x}_{ij}, \mathbf{w}_{ijk}$ and $\mathbf{z}_{ij}$ are fixed; $m \to \infty$ and $\max_i\{n_i\}$ is bounded, and the distinct values of t_{ij} form a quasi-uniform sequence that grows dense on $[0, 1]$. The first four moments of y_{ij} exist.

(A2) The sth derivatives of f_0 and f_1 are bounded for some $s \geq 2$.

(A3) The covariates $\mathbf{w}_{ijk}$ and the matrices $\mathbf{W}_i^{-1}$ are all bounded, which means that all the elements of the vectors and matrices are bounded. The function $g^{-1}(\cdot)$ has bounded second derivatives.

(A4) The parametric space Θ is a compact subset of R^{p+q+d}, and the parameter value $(\boldsymbol{\beta}_0', \boldsymbol{\gamma}_0', \boldsymbol{\lambda}_0')'$ is in the interior of the parameter space Θ.

The assumptions that the number of independent subjects goes to infinity and that the maximum number of repeated measurements is bounded are standard. They were used by Liang and Zeger (1986) to study the GEE estimates. The existence of the first four moments of the response is needed for consistently estimating the parameters in the variance. The smoothness conditions on f_0 and f_1 given by Condition (A2)

determine the rate of convergence of the spline estimates. Condition (A3) is satisfied as t is bounded. Assumption (A4) is routinely made in linear models.

To study the asymptotic properties of estimators, we assume the dependence between $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ and t_{ij} as following

$$x_{ijk} = g_k(t_{ij}) + \delta_{ijk}, \quad k = 1, \cdots, p. \tag{8}$$

$$z_{ijl} = \tilde{g}_l(t_{ij}) + \tilde{\delta}_{ijl}, \quad l = 1, \cdots, d; i = 1, \cdots, m; j = 1, \cdots, n_i; \tag{9}$$

where δ_{ijk}'s and $\tilde{\delta}_{ijl}$'s are mean zero random variables independent of the corresponding random errors and of one another. Relationships (8) and (9) are commonly made in the spline literature. See He, et al. (2002) and references therein. Indeed, Härdle et al. (2000, Chapter 1.3) considered similar assumptions for partly linear kernel regression. Our assumptions correspond to their random design case. Equations (8) and (9) can be applied when categorical variables are present (Speckman, 1988). For example, when x_k is a binary zero one variable denoting treatment assignment, $g_k(t) = P(x_k = 1|t)$. Independence between x_k and covariate t implies that g_k is constant. However, a systematic trend resulting in unbalanced allocation to treatment could result in non-constant g_k. Let $\mathbf{\Lambda}_n$ and $\tilde{\mathbf{\Lambda}}_n$ be the $n \times p$ and $n \times d$ matrices whose kth column are $\boldsymbol{\delta}_k = (\delta_{11k}, \cdots, \delta_{1n_1k}, \cdots, \delta_{mn_mk})'$ and $\tilde{\boldsymbol{\delta}}_k = (\tilde{\delta}_{11k}, \cdots, \tilde{\delta}_{1n_1k}, \cdots, \tilde{\delta}_{mn_mk})'$ respectively. We make the following assumption:

(A5) (1) $E\mathbf{\Lambda}_n$=0, $\sup_n \frac{1}{n} E\|\mathbf{\Lambda}_n\|^2 < \infty$; $E\tilde{\mathbf{\Lambda}}_n$=0, $\sup_n \frac{1}{n} E\|\tilde{\mathbf{\Lambda}}_n\|^2 < \infty$;

(2) $k_n\mathbf{M}'\mathbf{\Sigma}^0\mathbf{M}$ and $k_n\mathbf{M}'\mathbf{W}^0\mathbf{M}$ are nonsingular for sufficiently large n, and the eigenvalues of $k_n\mathbf{M}'\mathbf{\Sigma}^0\mathbf{M}/n$ and $k_n\mathbf{M}'\mathbf{W}^0\mathbf{M}/n$ are bounded away from 0 and infinity, where $\mathbf{M} = (\boldsymbol{\pi}_1', \cdots, \boldsymbol{\pi}_m')'$, $\mathbf{\Sigma}^0 = diag\{\mathbf{\Sigma}_1^0, \cdots, \mathbf{\Sigma}_m^0\}$ with $\mathbf{\Sigma}_i^0 = \mathbf{\Delta}_{0i}\mathbf{\Sigma}_{0i}^{-1}\mathbf{\Delta}_{0i} = \mathbf{\Delta}_i(\boldsymbol{\eta}_i^0)\mathbf{\Sigma}_{0i}^{-1}\mathbf{\Delta}_i(\boldsymbol{\eta}_i^0)$, and $\mathbf{W}^0$ is defined in a similar fashion.

We take the number of knots k_n as the integer part of $n^{1/(2s+1)}$, where s is defined

in (A2) and taken as 2 in this work. For this knot number, Condition (2) of (A5) is expected to hold as this is a property of the B-spline basis functions (He and Shi, 1996).

The asymptotic properties of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\lambda}})$ involve computation of the covariance matrix $\boldsymbol{\Delta}_m = (\boldsymbol{\delta}_m^{kl})_{k,l=1,2,3}$ of $(\tilde{\mathbf{S}}_1', \tilde{\mathbf{S}}_2', \tilde{\mathbf{S}}_3')'/\sqrt{m}$, where $\tilde{\mathbf{S}}_1, \tilde{\mathbf{S}}_2$ and $\tilde{\mathbf{S}}_3$ are defined by

$$\tilde{\mathbf{S}}_1 = \sum_{i=1}^{m} \mathbf{X}_i^{*\prime} \boldsymbol{\Delta}_{0i} \boldsymbol{\Sigma}_{0i}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_{0i}),\ \tilde{\mathbf{S}}_2 = \sum_{i=1}^{m} \mathbf{V}_i^{0\prime} \mathbf{D}_{0i}^{-1} (\mathbf{r}_{0i} - \hat{\mathbf{r}}_{0i}),\ \tilde{\mathbf{S}}_3 = \sum_{i=1}^{m} \mathbf{Z}_i^{*\prime} \mathbf{D}_{0i} \mathbf{W}_{0i}^{-1} (\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2). \tag{10}$$

Here $\mathbf{X}^* = (\mathbf{I} - \mathbf{P})\mathbf{X}$ with $\mathbf{P} = \mathbf{M}(\mathbf{M}'\boldsymbol{\Sigma}^0\mathbf{M})^{-1}\mathbf{M}'\boldsymbol{\Sigma}^0$; $\mathbf{r}_{0i} = \mathbf{y}_i - \boldsymbol{\mu}_{0i}$, $\hat{\mathbf{r}}_{0i} = (r_{0i1}, \cdots, r_{0ij}, \cdots, r_{0in_i})'$ with $\hat{\mathbf{r}}_{0ij} = \sum_{k=1}^{j-1} r_{0ik} \mathbf{w}_{ijk}' \boldsymbol{\gamma}_0$; $\mathbf{V}_i^0 = (0, r_{0i1}\mathbf{w}_{i21}', \cdots, \sum_{k=1}^{j-1} r_{0ik}\mathbf{w}_{ijk}')'$; and $\mathbf{Z}^* = (\mathbf{I} - \mathbf{P})\mathbf{Z}$. We make the following assumption similar to Assumption 11 in Ye and Pan (2006).

(A6) The covariance matrix $\boldsymbol{\Delta}_m$ is positive definite, and

$$\boldsymbol{\Delta}_m = \begin{pmatrix} \boldsymbol{\delta}_m^{11} & \boldsymbol{\delta}_m^{12} & \boldsymbol{\delta}_m^{13} \\ \boldsymbol{\delta}_m^{21} & \boldsymbol{\delta}_m^{22} & \boldsymbol{\delta}_m^{23} \\ \boldsymbol{\delta}_m^{31} & \boldsymbol{\delta}_m^{32} & \boldsymbol{\delta}_m^{33} \end{pmatrix} \to \boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{\delta}^{11} & \boldsymbol{\delta}^{12} & \boldsymbol{\delta}^{13} \\ \boldsymbol{\delta}^{21} & \boldsymbol{\delta}^{22} & \boldsymbol{\delta}^{23} \\ \boldsymbol{\delta}^{31} & \boldsymbol{\delta}^{32} & \boldsymbol{\delta}^{33} \end{pmatrix}, \quad \text{as } m \to \infty, \tag{11}$$

were $\boldsymbol{\Delta}$ is a positive definite matrix.

Theorem 1. *If (A1) to (A6) hold and $k_n = O(n^{1/(2s+1)})$, we have that*

$$\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \hat{f}_0(t_{ij}) - f_0(t_{ij}) \right\}^2 = O_p(n^{-2s/(2s+1)}), \tag{12}$$

$$\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \hat{f}_1(t_{ij}) - f_1(t_{ij}) \right\}^2 = O_p(n^{-2s/(2s+1)}), \tag{13}$$

where $\hat{f}_0(t) = \boldsymbol{\pi}'(t)\hat{\boldsymbol{\alpha}}$ and $\hat{f}_1(t) = \boldsymbol{\pi}'(t)\widehat{\tilde{\boldsymbol{\alpha}}}$.

As pointed out by He *et al.* (2005), (12) and (13) imply that $\int (\hat{f}_i(t) - f_i(t))^2 dt = O_p(n^{-2s/(2s+1)})$, $i = 0, 1$ under some general conditions (Stone, 1985, Lemmas 8 and 9). This is the optimal rate of convergence for estimating f_0 and f_1 under the smoothness condition (A2). For the parametric coefficients, we have the following.

Theorem 2. *Under conditions (A1)-(A6), the generalized estimating equation estimator* $(\hat{\boldsymbol{\beta}}_m', \hat{\boldsymbol{\gamma}}_m', \hat{\boldsymbol{\lambda}}_m')'$ *is* $\sqrt{m}$*-consistent and asymptotically normal, that is*

$$\sqrt{m}\begin{pmatrix} \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_0 \\ \hat{\boldsymbol{\lambda}}_m - \boldsymbol{\lambda}_0 \end{pmatrix} \to N\left\{0, \begin{pmatrix} \boldsymbol{\delta}^{11} & 0 & 0 \\ 0 & \boldsymbol{\delta}^{22} & 0 \\ 0 & 0 & \boldsymbol{\delta}^{33} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\delta}^{11} & \boldsymbol{\delta}^{12} & \boldsymbol{\delta}^{13} \\ \boldsymbol{\delta}^{21} & \boldsymbol{\delta}^{22} & \boldsymbol{\delta}^{23} \\ \boldsymbol{\delta}^{31} & \boldsymbol{\delta}^{32} & \boldsymbol{\delta}^{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}^{11} & 0 & 0 \\ 0 & \boldsymbol{\delta}^{22} & 0 \\ 0 & 0 & \boldsymbol{\delta}^{33} \end{pmatrix}^{-1}\right\}$$

in distribution as $m \to \infty$.

The asymptotic variance reduces to a diagonal matrix for normally distributed response variable, which is a semiparametric analog to Theorem 2 of Ye and Pan (2006).

For inference, we use a robust estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_m$ as follows

$$V(\hat{\boldsymbol{\beta}}_m) = \mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1}, \tag{14}$$

where $\mathbf{M}_0 = \sum_{i=1}^m \mathbf{X}_i^{*\prime}\hat{\boldsymbol{\Delta}}_i\hat{\boldsymbol{\Sigma}}_i\hat{\boldsymbol{\Delta}}_i\mathbf{X}_i^*$, $\mathbf{M}_1 = \sum_{i=1}^m \mathbf{X}_i^{*\prime}\hat{\boldsymbol{\Delta}}_i\hat{\boldsymbol{\Sigma}}_i(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)'\hat{\boldsymbol{\Delta}}_i\mathbf{X}_i^*$. The estimated covariance matrices of $\hat{\boldsymbol{\gamma}}_m$ and $\hat{\boldsymbol{\lambda}}_m$ can be obtained in a similar way, and the covariances $\boldsymbol{\delta}^{kl}(k \neq l)$ can also be estimated by their sample versions.

4 Numerical Study

Whenever appropriate, we compare our approach with GEE and that in Fan et al. (2007). We use local linear regression to model $f_0(t)$ in the mean and local constant regression for the marginal variance for Fan et al's method. For the conventional GEE, we use the same spline representation for f_0. All the code is written in R and function geese in R package geepack is used for GEE. For brevity, our approach is referred to as the semiparametric mean-variance method (SMV). The approach by Fan et al. (2007) is referred to as the quasi-likelihood method (QL) which is used for estimating the correlation.

4.1 Real data analysis

We apply the proposed estimation method to the CD4 cell study, which was analyzed by many authors (Zeger and Diggle, 1994; Ye and Pan, 2006). This data set comprises

CD4 cell counts of 369 HIV-infected men with six covariates including time since seroconversion (t_{ij}), age (relative to arbitrary origin, x_{ij1}), packs of cigarettes smoked per day (x_{ij2}), recreation drug use (x_{ij3}), number of sexual partners (x_{ij4}), mental illness score (x_{ij5}). Altogether there are 2,376 values of CD4 cell counts, with multiple repeated measurements taken for each individual at different times, covering a period of approximately eight and a half years. The number of measurements for each individual varies from 1 to 12 and the time points are not equally spaced. Thus, the CD4 cell data are highly unbalanced. We use square root transformation on the response by the suggestion in Zeger and Diggle (1994), where further details about the design and the medical implications of the study can be found.

To model jointly the mean and covariance structures for the CD4 cell data, we use the following mean model

$$y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + x_{ij3}\beta_3 + x_{ij4}\beta_4 + x_{ij5}\beta_5 + f_0(t_{ij}) + e_{ij}.$$

We take covariates for the autoregressive components as $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2, (t_{ij} - t_{ik})^3)'$ following the arguments in Ye and Pan (2006), and for the innovation variances as $\mathbf{z}_{ij} = \mathbf{x}_{ij}$. The latter specification allows us to examine whether the innovations are dependent on the covariates. Finally the number of knots is taken to be $[(2376)^{1/5}] = 7$, which is also the optimal number of knots according to the leave-one-subject-out cross validation. Table 1 lists the results for $\boldsymbol{\beta}$ by our modified Cholesky decomposition method, where a working AR(1) model with $\delta = 0.2$ is used for the innovation. For comparison, we list the conventional GEE method for the partly linear mean model using different working correlations, including independent, AR(1) and exchangeable structures. We also include Fan et al's approach by using AR(1) and ARMA(1,1) for the correlation. The results show that both SMV and QL give estimators with generally smaller standard errors. For our approach, smoking

and drug use are highly significant variables, while mental illness score is marginally significant. QL identifies smoking and mental illness as significant covariates, and estimates drug use as borderline insignificant. The significance of smoking is missed by GEE using AR(1) covariance structure, while that of drug use is missed by GEE using either AR(1) or exchangeable variance structure. Finally, GEE using the independent working correlation indicates that mental score is not significant at all, which contradicts with the GEE results using other working correlations. For the autoregressive and log innovation parameters, our model yields

$$\underset{(0.048)}{\gamma_1 = 0.675},\ \underset{(0.077)}{\gamma_2 = -0.563},\ \underset{(0.033)}{\gamma_3 = 0.170},\ \underset{(0.004)}{\gamma_4 = -0.017};$$

$$\underset{(0.011)}{\lambda_1 = -0.003},\ \underset{(0.045)}{\lambda_2 = 0.085},\ \underset{(0.122)}{\lambda_3 = -0.048},\ \underset{(0.015)}{\lambda_4 = 0.006},\ \underset{(0.005)}{\lambda_5 = -0.004},$$

where the standard errors are in parentheses. All the autoregressive parameters are highly significant.

Figure 1 displays the three fitted curves for f_0, ϕ as a function of the time lag and f_1, when $\mathbf{R}_i(\boldsymbol{\delta})$ is specified by AR(1) with $\delta = 0.2$. The trajectory of the mean curve is consistent with that in Zeger and Diggle (1994). Figure 1 (B) plots the estimated generalized autoregressive parameters ϕ against the time lag between measurements in the same subject, which, according to our model, is simply a cubic polynomial. This plot shows that the generalized autoregressive parameters decrease sharply if the time lag is less than two years and then drop slowly when the lag becomes larger. The innovation curve seems to fluctuate around a constant. These observations basically agree with those in Ye and Pan (2006).

It is helpful to compare our method with Fan et al's approach in terms of prediction. To this end, we randomly split the data into three parts, each with 369/3=123 subjects. We use the first two parts of this partitioning as the training data set to

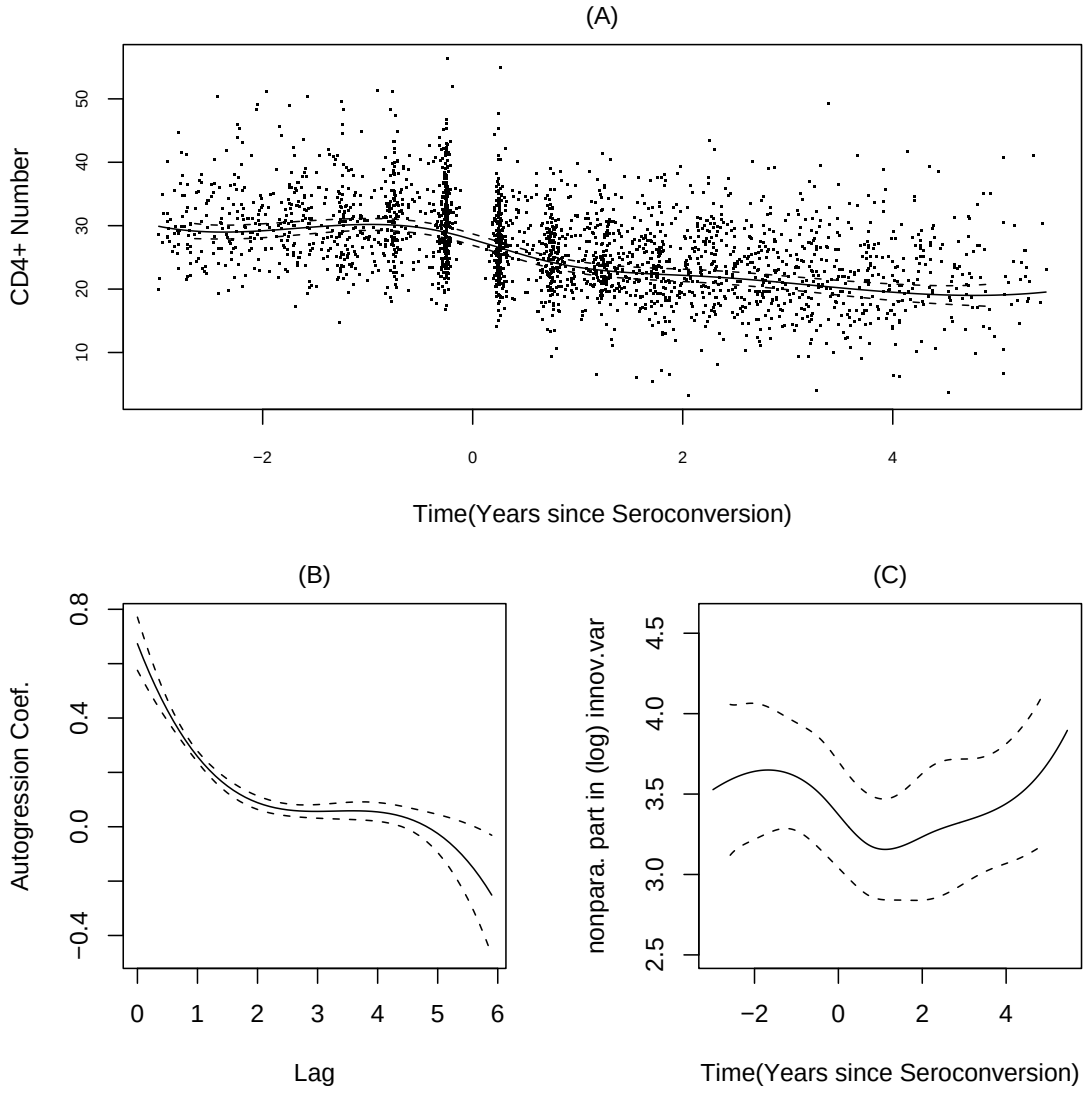

Figure 1: The CD4 cell data. The fitted curves of (A) nonparametric part in mean against time, (B) the generalized autoregressive parameters against lag and (C) the nonparametric part in (log) innovation variances against time based on AR(1) structure with $\delta = 0.2$ and square root CD4 cell numbers. Dashed curves represent asymptotic 95% confidence intervals.

Table 1: CD4 cell data. The estimates of parameters based on square root CD4 cell numbers, with standard errors in parentheses.

	SMV	Generalized Estimating Equations			QL	
		Independence	AR(1)	Exchangeable	AR(1)	ARMA(1,1)
β_1	$0.005_{(0.030)}$	$0.015_{(0.035)}$	$0.016_{(0.034)}$	$0.002_{(0.032)}$	$0.016_{(0.032)}$	$0.016_{(0.031)}$
β_2	$0.768_{(0.130)}$	$0.981_{(0.184)}$	$0.262_{(0.190)}$	$0.596_{(0.136)}$	$0.665_{(0.152)}$	$0.587_{(0.138)}$
β_3	$0.821_{(0.345)}$	$1.075_{(0.528)}$	$0.471_{(0.350)}$	$0.494_{(0.358)}$	$0.700_{(0.358)}$	$0.616_{(0.329)}$
β_4	$0.044_{(0.038)}$	$-0.064_{(0.059)}$	$0.050_{(0.041)}$	$0.060_{(0.043)}$	$0.011_{(0.040)}$	$0.067_{(0.038)}$
β_5	$-0.030_{(0.014)}$	$-0.031_{(0.021)}$	$-0.046_{(0.014)}$	$-0.048_{(0.015)}$	$-0.034_{(0.014)}$	$-0.043_{(0.014)}$

fit the model, and then assess the out-of-sample performance on the testing data set (denoted as TD) which is left out. We use the predictive mean square errors defined as $\sum_{i\in \mathrm{TD}} \sum_{j=1}^{n_i} (\hat{y}_{ij} - y_{ij})^2 / n_{\mathrm{TD}}$ as the performance measure, where $\hat{y}_{ij}$ is the fitted response using the training data set only and $n_{\mathrm{TD}} = \sum_{i\in \mathrm{TD}} n_i$ is the test data sample size. This process is repeated 100 times. The average mean square errors for QL AR(1), QL ARMA(1,1) and our approach are 37.19, 36.98, 37.53 with standard errors 2.98, 2.93 and 3.40 respectively. Pairwise t-tests of the mean square errors show that these approaches perform similarly. Note that our model is more flexible and potential more powerful because covariates other than t are allowed in modeling the log innovation variances.

As suggested by an anonymous reviewer, we can use graphical tools to explore the covariance structure produced by various methods. For a fair comparison, we fit the mean by using the regression spline approximation for $f_0(t)$, assuming errors are iid. The fitted residuals are then used for covariance modeling. For SMV, since none of $\boldsymbol{\lambda}$ is significant, we apply SMV with only a nonparametric function for the log innovation. Note that iteration for $\boldsymbol{\theta}$ in (5) is no longer needed for this comparison. Define the variogram $v(u,t) = E\big[\{e(t) - e(t-u)\}^2\big]/2,\ u \geq 0$ as a function of time-lag u and observing time t. We then compare the sample variogram, the variogram estimated by QL with ARMA(1,1) structure and variogram estimated by SMV at

the sample points by plotting $v(u,t)$ versus u and t in Figure 2. The vertical axis is truncated at 40 to accentuate the shape of the smooth estimate. It is seen that the sample variogram increases with time-lag, corresponding to decreasing correlation as observations are separated in time. Both QL and SMV capture this feature nicely. On the other hand, the plot of sample variogram versus observation time indicates that $v(u,t)$ is likely independent of t, which is missed by QL and to a lesser degree by SMV. Note that the enormous random fluctuation of the sample variogram is a typical phenomenon (Diggle, et al. 2002).

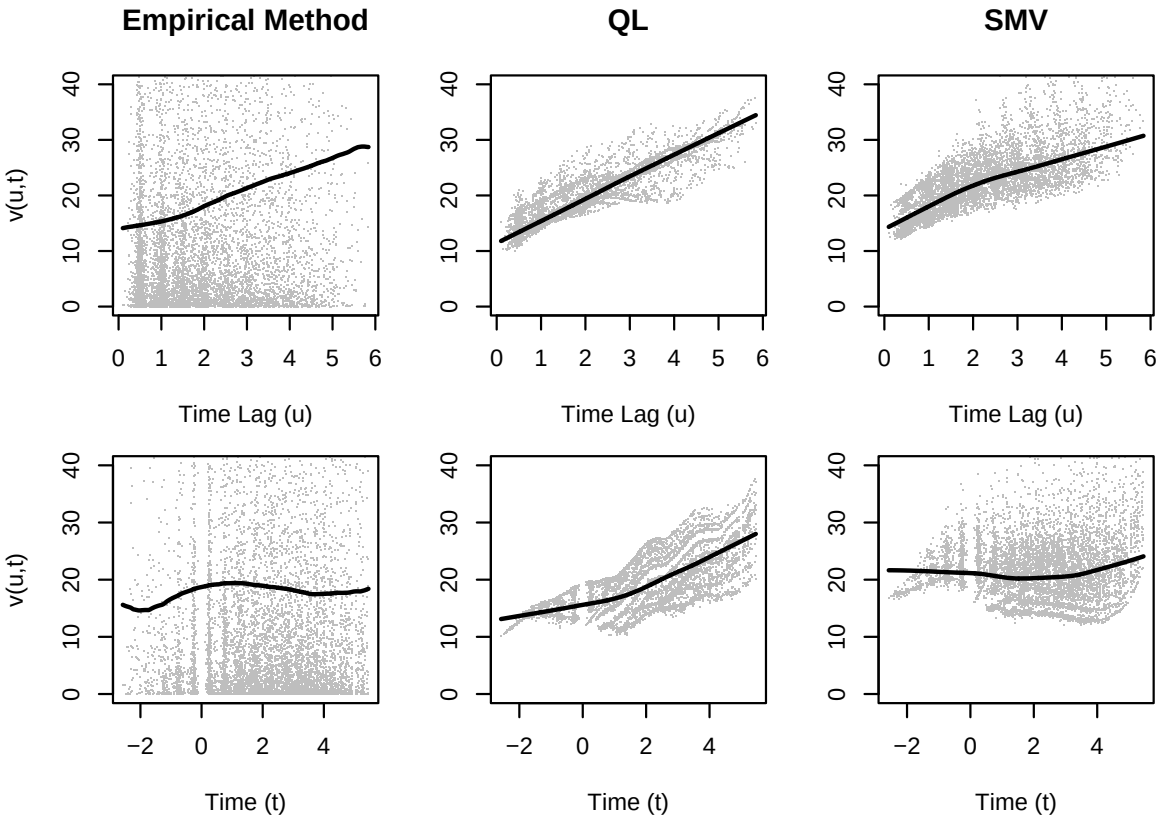

Figure 2: The sample variogram and its estimates by SMV and QL for CD4 data. Individual points represent the variogram at the sampling points and the solid lines are the LOWESS smoothers.

4.2 Simulation study

We conduct extensive numerical studies to assess the finite sample performance of the proposed method. We also test the asymptotic covariance formula in Theorem 2 and compare SMV with QL and the conventional GEE using working correlation. For each study, we simulate one thousand data sets.

Study 1. We first consider the following model

$$y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + f_0(t_{ij}) + e_{ij},\ i = 1, \cdots, m;\ j = 1, \cdots, n_i,$$

for $m = 100$. Note that we use ϵ_{ij} to denote the innovation and e_{ij} here for the random error associated with observation j for subject i. We use the sample scheme in Fan et al. (2007) such that the observation times are regularly scheduled but may be randomly missed in practice. More precisely, each individual has a set of scheduled time points $\{0, 1, 2, ..., 12\}$, and each scheduled time, except time 0, has a 20% probability of being skipped. The actual observation time is a random perturbation of a scheduled time: a uniform [0, 1] random variable is added to a nonskipped scheduled time. This results in different observed time points t_{ij} per subject, and then t_{ij} is transformed onto $[0, 1]$. Note that the longitudinal observations are highly irregular and some of $\{n_i\}$ are less than the number of the parameters in the same subject.

We take $x_{ij1} = t_{ij} + \delta_{ij}$, where δ_{ij} follows the standard normal distribution and let x_{ij2} follow a Bernoulli distribution with success probability 0.5. Note that x_{ij1} is time-varying. For the nonparametric function in the mean, we take $f_0(t) = \cos(t\pi)$. The error $(e_{i1}, \cdots, e_{in_i})'$ is generated to follow a multivariate normal distribution with mean 0 and covariance $\mathbf{\Sigma}_i$ satisfying $\mathbf{\Phi}_i\mathbf{\Sigma}_i\mathbf{\Phi}_i' = \mathbf{D}_i$, where $\mathbf{\Phi}_i$ and $\mathbf{D}_i$ are described in Section 2.1 with $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik})'$, $\mathbf{z}_{ij} = \mathbf{x}_{ij}$, and $f_1(t) = \sin(\pi t)$. We consider using AR(1) for $\mathbf{R}_i(\delta)$ in $\mathbf{W}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\delta)\mathbf{A}_i^{1/2}$, the working covariance structure of

$\boldsymbol{\epsilon}_i^2$. The compound symmetry (exchangeable) structure is also used. Because the results are similar, we omit the details. In each case, the parameter δ, measuring the correlation between $\epsilon_{ij}^2 = (y_{ij} - \hat{y}_{ij})^2$ and $\epsilon_{ik}^2 = (y_{ik} - \hat{y}_{ik})^2$, takes four different values, $\delta = 0, 0.2, 0.5, 0.8$, so that the effect of misspecification of $\mathbf{R}_i(\delta)$ on estimating $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$ can be studied. We take two specifications: Case (1) $\boldsymbol{\beta} = (1, 0.5)'$, $\boldsymbol{\gamma} = (0.2, 0.3)'$ and $\boldsymbol{\lambda} = (-0.5, 0.2)'$; Case (2) $\boldsymbol{\beta} = (1, 0)'$, $\boldsymbol{\gamma} = (0.2, 0)'$, and $\boldsymbol{\lambda} = (-0.5, 0)'$. For each setting, the expected sample size is about 1040. The number of the knots is taken to be $4 \approx 1040^{1/5}$. Numerical experiments show that the results are not very sensitive to the number of the knots.

For each simulated data set, we use $\delta = 0,\ 0.2,\ 0.5$ and 0.8 to study the robustness of the approach with regard to δ. Table 2 shows that our semiparametric methods literally yield unbiased estimates for the parameters. Additionally, the parameter δ used in the working covariance structure for the innovations has little effect on the estimation of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$, and the estimated mean square errors for f_0 and f_1, when the structure for $\mathbf{R}_i(\delta)$ is based on AR(1). These results imply that SMV is robust against misspecification of the structure of $\mathbf{R}_i(\delta)$. For Case (1), Figure 3 displays the true and fitted curves for nonparametric function f_0 and f_1 when $\mathbf{R}_i(\delta)$ is specified by AR(1) with $\delta = 0.2$. The three curves $\hat{f}_5, \hat{f}_{50}$ and $\hat{f}_{95}$ represent the fits which are 5%, 50% and 95% best in terms of the mean squared errors in 1000 runs, respectively. They show a close agreement with the true functions. To check the robustness of the choice of $\mathbf{W}$, we have also followed Ye and Pan (2006) by generating longitudinal data sets from normal mixture distributions. The results are robust to this deviation of the normality, and are qualitatively similar to those in Ye and Pan (2006). The detailed results can be found in the online supplemental materials.

Hereafter, we use AR(1) with $\delta = 0.2$ for $\mathbf{R}_i(\boldsymbol{\delta})$ in the following simulation when-

Table 2: Simulation results for Study 1 over 1000 replications. Presented are the sample means with sample standard errors in parentheses. $MSE(\hat{f}_i),\ i = 0, 1$ is the mean square error for the estimate f_i over all time points in the data.

	True	$\delta = 0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
β_1	1.0	$1.00_{(0.0276)}$	$1.00_{(0.0276)}$	$1.00_{(0.0276)}$	$1.00_{(0.0277)}$
β_2	0.5	$0.50_{(0.0579)}$	$0.50_{(0.0579)}$	$0.50_{(0.0579)}$	$0.50_{(0.0579)}$
γ_1	0.2	$0.20_{(0.0174)}$	$0.20_{(0.0174)}$	$0.20_{(0.0174)}$	$0.20_{(0.0174)}$
γ_2	0.3	$0.30_{(0.0525)}$	$0.30_{(0.0525)}$	$0.30_{(0.0526)}$	$0.30_{(0.0527)}$
λ_1	-0.5	$-0.50_{(0.0435)}$	$-0.50_{(0.0450)}$	$-0.50_{(0.0502)}$	$-0.50_{(0.0542)}$
λ_2	0.2	$0.20_{(0.0931)}$	$0.20_{(0.0974)}$	$0.20_{(0.1082)}$	$0.20_{(0.1161)}$
$MSE(\hat{f}_0)$		$0.0282_{(0.0371)}$	$0.0282_{(0.0371)}$	$0.0281_{(0.0371)}$	$0.0281_{(0.0371)}$
$MSE(\hat{f}_1)$		$0.0110_{(0.0081)}$	$0.0113_{(0.0084)}$	$0.0131_{(0.0097)}$	$0.0180_{(0.0130)}$
β_1	1.0	$1.00_{(0.0308)}$	$1.00_{(0.0308)}$	$1.00_{(0.0308)}$	$1.00_{(0.0308)}$
β_2	0	$0.00_{(0.0615)}$	$0.00_{(0.0616)}$	$0.00_{(0.0616)}$	$0.00_{(0.0618)}$
γ_1	0.2	$0.20_{(0.0181)}$	$0.20_{(0.0181)}$	$0.20_{(0.0181)}$	$0.20_{(0.0182)}$
γ_2	0	$0.00_{(0.0487)}$	$0.00_{(0.0487)}$	$0.00_{(0.0487)}$	$0.00_{(0.0488)}$
λ_1	-0.5	$-0.50_{(0.0453)}$	$-0.50_{(0.0465)}$	$-0.50_{(0.0514)}$	$-0.50_{(0.0551)}$
λ_2	0	$0.00_{(0.0873)}$	$0.00_{(0.0909)}$	$0.00_{(0.1017)}$	$0.00_{(0.1098)}$
$MSE(\hat{f}_0)$		$0.0184_{(0.0230)}$	$0.0183_{(0.0229)}$	$0.0184_{(0.0230)}$	$0.0184_{(0.0230)}$
$MSE(\hat{f}_1)$		$0.0112_{(0.0084)}$	$0.0114_{(0.0086)}$	$0.0132_{(0.0010)}$	$0.0182_{(0.0137)}$

ever our approach is used.

Study 2. This example is used to illustrate the performance of the asymptotic covariance formula in Theorem 2, following the simulation setup Case (1) in Study 1. In Table 3, "SD" represents the sample standard deviation of 1,000 estimates of $\boldsymbol{\beta}$ which can be viewed as the true standard deviation of the resulting estimates. "SE" represents the sample average of 1,000 estimated standard errors using formula (14), and "Std" represents the standard deviation of these 1,000 standard errors. Table 3 demonstrates that the standard error formula works well for different AR(1) correlation structures.

Table 3: Assessment of the standard errors using formula (14).

		$\delta = 0$	$\delta = 0.2$	$\delta = 0.5$	$\delta = 0.8$
β_1	SD	0.0276	0.0276	0.0276	0.0277
	$\text{SE}_{(\text{Std})}$	$0.0267_{(0.0027)}$	$0.0267_{(0.0027)}$	$0.0268_{(0.0027)}$	$0.0268_{(0.0028)}$
β_2	SD	0.0579	0.0579	0.0579	0.0579
	$\text{SE}_{(\text{Std})}$	$0.0553_{(0.0047)}$	$0.0553_{(0.0047)}$	$0.0554_{(0.0047)}$	$0.0555_{(0.0048)}$

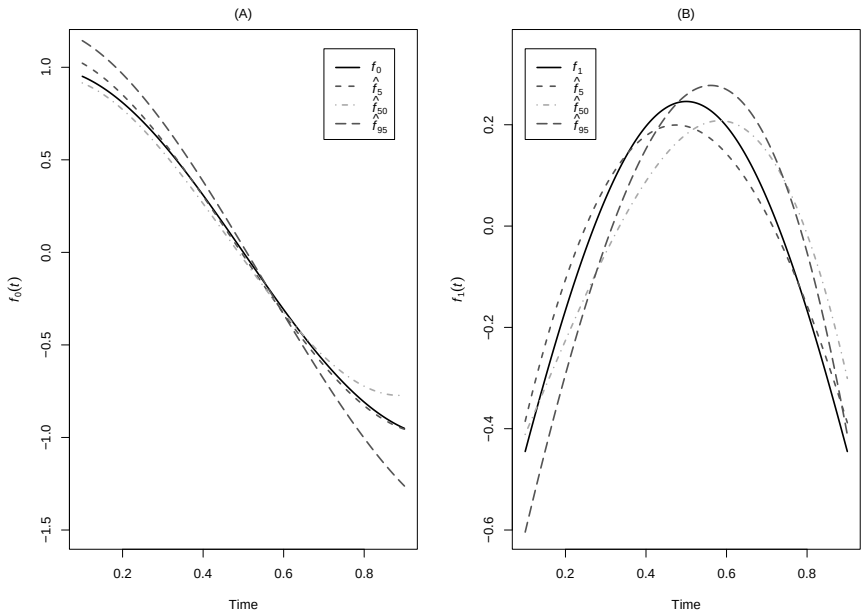

Figure 3: Nonparametric function f_0 and f_1 and their fitted curves $\hat{f}_5, \hat{f}_{50}, \hat{f}_{95}$, for AR(1) structure with $\delta = 0.2$.

Study 3. In this example, we assess the effect of misspecification of the working covariance structure $\mathbf{\Sigma}_i$ on the estimation of $\boldsymbol{\beta}$ and f_0. For comparison, we apply GEE with independent, exchangeable and AR(1) working correlation. The results are summarized in Table 4. Not surprisingly, all methods give almost unbiased estimates for $\boldsymbol{\beta}$. However, the standard errors of our semiparametric method is much smaller than those of the other methods, implying that our estimator is more efficient. Furthermore, for the nonparametric part f_0 in the mean, our semiparametric approach gives estimates with significantly smaller mean square errors. Taken together, the study shows that the SMV approach is more accurate in estimating the mean.

Table 4: Simulation results for Study 3 with standard errors in parentheses.

		SMV	Generalized Estimating Equations		
	True		Independence	Exchangeable	AR(1)
β_1	1.0	$1.00_{(8.5\times10^{-4})}$	$1.00_{(3.03\times10^{-3})}$	$1.00_{(2.18\times10^{-3})}$	$1.00_{(1.71\times10^{-3})}$
β_2	0.5	$0.50_{(1.76\times10^{-3})}$	$0.50_{(5.88\times10^{-3})}$	$0.51_{(4.27\times10^{-3})}$	$0.50_{(3.21\times10^{-3})}$
$MSE(f_0)$		$0.0282_{(0.0371)}$	$0.1037_{(0.1328)}$	$0.0887_{(0.1196)}$	$0.0917_{(0.1217)}$

Study 4. We use this example to compare SMV and QL for estimating the covariance matrix. The mean model, taken from Example 5.1 in Fan and Wu (2008), has a parametric form as $y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + e_{ij}$, where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3) = (2, 1.5, 3)$, x_{ijk} marginally follows standard normal distribution and the correlation between the ith and the jth covariate in $\mathbf{x}$ is $0.5^{|i-j|}$. We use a linear model for the mean since the main difference between SMV and QL lies in covariance estimation. We consider two scenarios for generating the covariance matrix, first of which favors QL and the second SMV. Especially, for QL, we take $h_0(t) = 0.5\exp(t/12 + \sin(t/12))$ as the marginal variance function and the true correlation matrix as AR(1) with parameter $a = 0,\ 0.3,\ 0.6$ or 0.9. For SMV, we take h_0 as the innovation function and $\mathbf{w}_{ijk} = (1, t_{ij} - t_{ik})$ as in Study 1, but fix $\gamma_1 = 0$ while allowing γ_2 to be $0.3,\ 0.6$ or 0.9. Note that the numerical values of γ_2 are not directly comparable to the values of a, because one is defined on the autoregressive structure and the other is on the correlation. We remark that when $a = 0$, both the variance-correlation and the modified Cholesky decompositions are appropriate. The observation times are generated according to Study 1 and we take $m = 200$ as in Fan and Wu (2008). For QL, we use the AR(1) working correlation structure. Thus, if the data generating process favors QL, we would fit a correctly specified model using QL but a misspecified model using SMV.

Since both approaches give literally unbiased estimates for $\boldsymbol{\beta}$, here we only report the sample standard errors of the estimated $\boldsymbol{\beta}$ under these two methods, together with those y using GEE with working independence (GEE). To compare accuracy in estimating the covariance matrix, we use the entropy loss $L(\Sigma, \hat{\Sigma}) = m^{-1}\sum_{i=1}^{m}\{trace(\Sigma_i\hat{\Sigma}_i^{-1}) - \log|\Sigma_i\hat{\Sigma}_i^{-1}| - n_i\}$, where Σ_i is the true covariance matrix and $\hat{\Sigma}_i$ is its estimator. The results are summarized in Table 5. When the model

favors Fan et al. (i.e. $a \neq 0$), QL gives more accurate estimates of $\boldsymbol{\beta}$ and the covariance structure, judged by the smaller sample standard deviations of $\boldsymbol{\beta}$ and the smaller means of the entropy loss. On the other hand, SMV performs better if the data is generated from the second scenario (i.e. $\gamma_2 \neq 0$). Both SMV and QL ourperform the GEE estimator with working independence, even for misspecified models. Interestingly, when both approaches are appropriate ($a = 0$), our approach performs better with smaller standard errors and entropy losses. This may be due to the fact that in estimating the marginal variance, Fan et al. (2007) used local constant estimate.

Table 5: Simulation results comparing SMV and QL . For $\boldsymbol{\beta}$, the sample standard errors (multplied by a factor of 1000) are reported, while for $L(\Sigma, \hat{\Sigma})$, both the mean and the sample standard errors (in parentheses), multplied by a factor of 100, are reported.

	β_1			β_2			β_2			$L(\Sigma, \hat{\Sigma})$	
	QL	SMV	GEE	QL	SMV	GEE	QL	SMV	GEE	QL	SMV
$a = 0$	30.9	29.5	31.7	32.8	31.8	34.7	28.2	26.9	33.0	$22.3_{(7.6)}$	$11.7_{(2.0)}$
$a = 0.3$	23.4	26.7	31.8	26.5	29.7	35.7	23.1	25.4	30.2	$3.8_{(1.8)}$	$96.6_{(4.2)}$
$a = 0.6$	17.4	22.8	32.3	19.2	25.4	35.0	17.3	22.6	31.6	$4.3_{(2.2)}$	$225.4_{(6.2)}$
$a = 0.9$	8.7	14.6	32.6	9.8	15.8	35.3	8.49	14.9	32.8	$12.5_{(5.3)}$	$592.1_{(9.9)}$
$\gamma_2 = 0.3$	30.2	27.3	32.8	34.6	30.7	36.8	32.9	27.0	32.6	$66.7_{(11.0)}$	$11.8_{(2.1)}$
$\gamma_2 = 0.6$	32.0	26.6	36.3	33.1	27.4	41.3	30.3	25.2	37.5	$240.9_{(9.9)}$	$11.6_{(2.1)}$
$\gamma_2 = 0.9$	42.3	23.3	47.2	46.6	25.7	54.1	42.9	22.7	46.5	$568.8_{(24.1)}$	$12.8_{(2.0)}$

5 Discussion

We have proposed semiparametric mean-covariance models for longitudinal data analysis using the modified Cholesky decomposition. This decomposition is adopted such that partly linear regression models can be applied to the autoregressive coefficients and log innovation variances. On the one hand, our approach extends the semiparametric model for the mean in longitudinal analysis. On the other hand, our approach relaxes the parametric assumption made by Ye and Pan (2006). We have provided theoretical justification for the proposed approach and compared our method with the variance-correlation decomposition of Fan et al. (2007) and the conventional GEE.

In practice, the true underlying covariance structure is unknown. It is important to develop tools to check whether the imposed covariance decomposition is reasonable. We have used variogram for this purpose in CD4 data analysis. More work needs to be done in this direction.

Our results are valid under the usual assumptions that the number of observations m goes to infinity while the number of repeated measurements n_i is bounded. When m or $\max\{n_i\}$ or both go to infinity, the asymptotic results in Xie and Yang (2003) may be extended to our setup as the model in (4) can be seen as a semiparametric generalization of the GEE. However, our setup is more challenging due to simultaneous models for the mean and the covariance. Although limited simulation (not shown) indicates that our method still works, a more rigorous study deserves more attention.

In this work, we only consider the classical setup when $\mathbf{x}$ are $\mathbf{z}$ are finite dimensional. For a diverging number of covariates in the mean, Lam and Fan (2008) studied the profile likelihood estimator. For our semiparametric approach, it will be interesting to investigate the statistical properties with diverging numbers of parameters both in the mean and the variance. Finally, it would be interesting to extend the semiparametric approach to nonnormal longitudinal data analysis.

Appendix: sketch of proofs

Proofs of Lemma 2 and Theorem 2 are available in the supplemental materials. The following lemma is stated for easy reference (Schumaker, 1981, Theorem 12.7).

Lemma 1. *Under Assumptions (A1)-(A2), there exists constants C_0 and C_1 such that $\sup_{t\in[0,1]} |f_0(t) - \boldsymbol{\pi}'(t)\boldsymbol{\alpha}_0| \le C_0 k_n^{-s}$, and $\sup_{t\in[0,1]} |f_1(t) - \boldsymbol{\pi}'(t)\tilde{\boldsymbol{\alpha}}_0| \le C_1 {k_n}^{-s}$.*

Proof of Theorem 1. Equation (12) can be obtained directly from He et al (2005).

Here we only give a proof of equation (13) when all $\mathbf{W}_i$ are known, denoted by $\mathbf{W}_{0i}$. Similar asymptotic results hold when all $\mathbf{W}_{0i}$ are replaced by consistent estimates. Let

$$\mathbf{T}_m = \begin{pmatrix} \mathbf{A}_m^{-1/2} & -\mathbf{A}_m^{-1/2}\mathbf{H}'\mathbf{W}^0\mathbf{M}(\mathbf{M}'\mathbf{W}^0\mathbf{M})^{-1} \\ \mathbf{0} & k_n^{1/2}\mathbf{Q}_m^{-1} \end{pmatrix},$$

where $\mathbf{A}_m = \mathbf{H}^{*\prime}\mathbf{W}^0\mathbf{H}^* = \sum_{i=1}^m \mathbf{H}_i^{*\prime}\mathbf{W}_i^0\mathbf{H}_i^*$, $\mathbf{Q}_m^2 = k_n\mathbf{M}'\mathbf{W}^0\mathbf{M}$. Obviously, condition (A6) implies that $\mathbf{A}_m/m \to \mathbf{A} > 0$ in probability for some positive matrix $\mathbf{A}$ as $m \to \infty$. From the definition of $\mathbf{T}_m$, it is easy to know that $\mathbf{T}_m \sum_{i=1}^m \mathbf{H}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}\mathbf{D}_{0i}\mathbf{H}_i\mathbf{T}_m' = \mathbf{I}_{d+K}$, where $\mathbf{I}_{d+K}$ is $(d+K)\times(d+K)$ identity matrix. We further denote

$$\boldsymbol{\zeta}(\boldsymbol{\rho}) = \begin{pmatrix} \boldsymbol{\zeta}_1 \\ \boldsymbol{\zeta}_2 \end{pmatrix} = (\mathbf{T}_m')^{-1}(\boldsymbol{\rho}-\boldsymbol{\rho}_0) = \begin{pmatrix} \mathbf{A}_m^{1/2}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0) \\ k_n^{-1/2}\mathbf{Q}_m(\tilde{\boldsymbol{\alpha}}-\tilde{\boldsymbol{\alpha}}_0) + k_n^{1/2}\mathbf{Q}_m^{-1}\mathbf{M}'\mathbf{W}^0\mathbf{H}(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0) \end{pmatrix}$$

and

$$\hat{\boldsymbol{\zeta}} = \begin{pmatrix} \hat{\boldsymbol{\zeta}}_1 \\ \hat{\boldsymbol{\zeta}}_2 \end{pmatrix} = \boldsymbol{\zeta}(\hat{\boldsymbol{\lambda}}, \widehat{\tilde{\boldsymbol{\alpha}}}). \tag{A.1}$$

From Lemma 1, it is easy to know that for sufficiently large n,

$$\frac{1}{n}\sum_{i=1}^m\sum_{j=1}^{n_i}\left\{\hat{f}_1(t_{ij}) - f_1(t_{ij})\right\}^2 \;\le\; \frac{2}{n}\sum_{i=1}^m\sum_{j=1}^{n_i}(\boldsymbol{\pi}_{ij}'(\widehat{\tilde{\boldsymbol{\alpha}}}-\tilde{\boldsymbol{\alpha}}_0))^2 + 2C_0k_n^{-2s},$$

and $\|\mathbf{A}_m^{-1/2}(\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0)\| \le \|\hat{\boldsymbol{\zeta}}\|$,

$$\begin{aligned}\Big[\frac{1}{n}\sum_{i=1}^m\sum_{j=1}^{n_i}(\boldsymbol{\pi}_{ij}'(\widehat{\tilde{\boldsymbol{\alpha}}}-\tilde{\boldsymbol{\alpha}}_0))^2\Big]^{1/2} &= n^{-1/2}\|\mathbf{M}(\widehat{\tilde{\boldsymbol{\alpha}}}-\tilde{\boldsymbol{\alpha}}_0)\| \le Cn^{-1/2}\|k_n^{-1/2}\mathbf{Q}_n(\widehat{\tilde{\boldsymbol{\alpha}}}-\tilde{\boldsymbol{\alpha}}_0)\| \\ &\le\; Cn^{-1/2}\|\hat{\boldsymbol{\zeta}}\| + C\boldsymbol{\lambda}_n^{-1/2}\|\hat{\boldsymbol{\lambda}}-\boldsymbol{\lambda}_0\|\sup_{\|\mathbf{a}\|=1,\|\mathbf{b}\|=1}|n^{-1}\mathbf{a}'\mathbf{M}'\mathbf{W}^0\mathbf{H}bk_n^{1/2}|,\end{aligned}$$

where λ_n is the minimum eigenvalue of $k_n\mathbf{M}'\mathbf{W}^0\mathbf{M}/n$. Then by Lemma 6.2 of He and Shi (1996) it suffices to show that $\|\hat{\boldsymbol{\zeta}}\| = O_p(k_n^{1/2})$. To do so, let $\mathbf{R}_{mi} = \boldsymbol{\pi}_i\tilde{\boldsymbol{\alpha}}_0 - f_1(t_i)$, $\boldsymbol{\eta}_i^0 = \mathbf{H}_i\boldsymbol{\lambda}_0 + f_1(t_i)$, and $\boldsymbol{\varsigma}_i = \tilde{\mathbf{H}}_i\boldsymbol{\zeta} + \mathbf{R}_{mi}$, where $\tilde{\mathbf{H}}_i = \mathbf{H}_i\mathbf{T}_m' = (\mathbf{H}_i^*\mathbf{A}_m^{-1/2}, \boldsymbol{\pi}_i\mathbf{Q}_m^{-1}k_n^{1/2})$. Then it's easy to see that $\mathbf{H}_i\boldsymbol{\zeta} = \boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i$, $\boldsymbol{\sigma}_i^2 = \exp(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)$, and the third estimating equation of (4) can be rewritten as

$$\mathbf{S}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) = \sum_{i=1}^m \mathbf{H}_i'\mathbf{D}_i(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \exp(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)) = 0. \tag{A.2}$$

Multiply $\mathbf{T}_m$ to equation (A.2) and we get

$$\boldsymbol{\Psi}(\zeta) = \mathbf{T}_m \mathbf{S}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) = \sum_{i=1}^{m} \tilde{\mathbf{H}}_i' \mathbf{D}_i(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \exp(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)) = 0. \tag{A.3}$$

It is easy to known that both (A.2) and (A.3) give the same root for $\boldsymbol{\zeta}$ by conditions (A4) and (A5). Let $\mathbf{a} \in \mathbb{R}^{d+\tilde{K}}$ such that $\mathbf{a}'\mathbf{a} = 1$. We expand $\mathbf{a}'\boldsymbol{\Psi}(\boldsymbol{\zeta})$ in a Taylor series,

$$\begin{aligned}
\mathbf{a}'\boldsymbol{\Psi}(\boldsymbol{\zeta}) &= \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i' \mathbf{D}_i(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \exp(\boldsymbol{\eta}_i^0 + \boldsymbol{\varsigma}_i)) \\
&= \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i' \mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_{0i}^2) - \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i' \mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}\mathbf{D}_{0i}\boldsymbol{\varsigma}_i \\
&+ \sum_{i=1}^{m} \boldsymbol{\varsigma}_i' \frac{\partial \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_i}{\partial \boldsymbol{\varsigma}_i}\Big|_{\boldsymbol{\varsigma}_i=0} \mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_{0i}^2) + \mathbf{R}_m^*(\boldsymbol{\varsigma}^*),
\end{aligned} \tag{A.4}$$

where $\mathbf{R}_m^*(\boldsymbol{\varsigma}^*) = \sum_{i=1}^m \mathbf{R}_{mi}^*(\boldsymbol{\varsigma}_i^*)$ and $\mathbf{R}_{mi}^*(\boldsymbol{\varsigma}_i^*) = \frac{1}{2}\boldsymbol{\varsigma}_i'[\partial^2 \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_i\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_i^2)/\partial\boldsymbol{\varsigma}_i\partial\boldsymbol{\varsigma}_i'|_{\boldsymbol{\varsigma}_i=\boldsymbol{\varsigma}_i^*}]\boldsymbol{\varsigma}_i$ for $\boldsymbol{\varsigma}_i^* = \boldsymbol{\eta}_i^0 + \tau_i\boldsymbol{\varsigma}_i$ $(i = 1, \cdots, m)$ with $0 < \tau_i < 1$. Further, let $\boldsymbol{\Phi}(\boldsymbol{\zeta}) = \sum_{i=1}^m \tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2) - \boldsymbol{\zeta}$, where $\boldsymbol{\epsilon}_{0i} = \mathbf{y}_i - \boldsymbol{\mu}_{0i}$. Denote the solution of $\boldsymbol{\Phi}$ as $\tilde{\boldsymbol{\zeta}}$, that is,

$$\tilde{\boldsymbol{\zeta}} = \begin{pmatrix} \tilde{\boldsymbol{\zeta}}_1 \\ \tilde{\boldsymbol{\zeta}}_2 \end{pmatrix} = \sum_{i=1}^{m} \tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2). \tag{A.5}$$

From these two expressions the difference between $\mathbf{a}'\boldsymbol{\Psi}(\boldsymbol{\zeta})$ and $\mathbf{a}'\boldsymbol{\Phi}(\boldsymbol{\zeta})$ can be expressed as

$$\begin{aligned}
\mathbf{a}'(\boldsymbol{\Psi}(\boldsymbol{\zeta}) - \boldsymbol{\Phi}(\boldsymbol{\zeta})) &= \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\epsilon}_{0i}^2) - \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}\mathbf{D}_{0i}\mathbf{R}_{mi} \\
&+ \sum_{i=1}^{m} \boldsymbol{\varsigma}_i' \frac{\partial \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_i}{\partial \boldsymbol{\varsigma}_i}\Big|_{\boldsymbol{\varsigma}_i=0} \mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_{0i}^2) + R_m^*(\boldsymbol{\varsigma}^*) \\
&=: I_{n0} - I_{n1} + I_{n2}(\zeta) + R_m^*(\boldsymbol{\varsigma}^*).
\end{aligned}$$

By Cauchy-Schwarz inequality, the definition of $\tilde{\mathbf{H}}$ and $k_n = O(n^{1/(2s+1)})$, we have

$$E(I_{n0})^2 \leq \sum_{i=1}^{m} \mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}\mathbf{D}_{0i}\tilde{\mathbf{H}}_i\mathbf{a} \sum_{i=1}^{m} E(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\epsilon}_{0i}^2)'\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\epsilon}_{0i}^2)$$

$$
\begin{aligned}
&= \sum_{i=1}^{m} E(\tilde{\mathbf{D}}(\boldsymbol{\nabla}\boldsymbol{\mu}_i)\boldsymbol{\epsilon}_{0i} + (\boldsymbol{\nabla}\boldsymbol{\mu}_i)^2)'\mathbf{W}_{0i}^{-1}(\tilde{\mathbf{D}}(\boldsymbol{\nabla}\boldsymbol{\mu}_i)\boldsymbol{\epsilon}_{0i} + (\boldsymbol{\nabla}\boldsymbol{\mu}_i)^2) \\
&= \sum_{i=1}^{m} trace\{\tilde{\mathbf{D}}(\boldsymbol{\nabla}\boldsymbol{\mu}_i)\mathbf{W}_{0i}^{-1}\tilde{\mathbf{D}}(\boldsymbol{\nabla}\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_{0i}\} + \sum_{i=1}^{m}\{(\boldsymbol{\nabla}\boldsymbol{\mu}_i)^2\}'\mathbf{W}_{0i}^{-1}(\boldsymbol{\nabla}\boldsymbol{\mu}_i)^2 \\
&\leq C\sum_{i=1}^{m}[\|\boldsymbol{\nabla}\boldsymbol{\mu}_i\|^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\boldsymbol{\nabla}\boldsymbol{\mu}_{ij}]^4 = O_p(k_n), \qquad \text{(A.6)}
\end{aligned}
$$

where $\boldsymbol{\nabla}\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1} - \boldsymbol{\mu}_{0i1}, \cdots, \boldsymbol{\mu}_{in_i} - \boldsymbol{\mu}_{0in_i})'$ and $\tilde{\mathbf{D}}(\boldsymbol{\nabla}\boldsymbol{\mu}_i) = diag\{\mu_{i1} - \mu_{0i1}, \cdots, \mu_{in_i} - \mu_{0in_i}\}$ with $\boldsymbol{\mu}_i = \mu(\mathbf{B}_i\hat{\boldsymbol{\theta}}) = g^{-1}(\mathbf{B}_i\hat{\boldsymbol{\theta}})$. The last inequality in (A.6) can be obtained easily by He et al (2005). Thus $|I_{n0}| = O_p(k_n^{1/2})$. For I_{n1}, Obviously,

$$
\begin{aligned}
|I_{n1}| &= \Big|\sum_{i=1}^{m}\mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_{0i}\mathbf{W}_{0i}^{-1}\mathbf{D}_{0i}\mathbf{R}_{mi}\Big| = |\mathbf{a}'\tilde{\mathbf{H}}\mathbf{W}^0\mathbf{R}_m| \\
&= \{\mathbf{a}'\tilde{\mathbf{H}}'\mathbf{W}^0\tilde{\mathbf{H}}\mathbf{a}\}^{1/2}\{\mathbf{R}_m'\boldsymbol{\Sigma}^0\mathbf{R}_m\}^{1/2} = O(n^{1/2}k_n^{-s}) = O(k_n^{1/2}),
\end{aligned}
$$

where $\mathbf{R}_m = (\mathbf{R}_{m1}, \cdots, \mathbf{R}_{mn_m})'$.

For $I_{n2}(\zeta)$, write

$$
I_{n2}(\boldsymbol{\zeta}) = \sum_{i=1}^{m}\boldsymbol{\zeta}'\tilde{\mathbf{H}}_i'\mathbf{G}_{0,i} + \sum_{i=1}^{m}\mathbf{R}_{mi}'\mathbf{G}_{0,i} =: I_{n2}^{(1)}(\boldsymbol{\zeta}) + I_{n2}^{(2)},
$$

where $\mathbf{G}_{0,i} = \frac{\partial\mathbf{a}'\tilde{\mathbf{H}}_i'\mathbf{D}_i}{\partial\boldsymbol{\varsigma}_i}\Big|_{\boldsymbol{\varsigma}_i=0}\mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_{0i}^2)$. Let $\tilde{\mathbf{e}}_i = \mathbf{W}_{0i}^{-1}(\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_{0i}^2) = (\tilde{e}_{i1}, \cdots, \tilde{ve}_{in_i})'$. It is easy to know that $\mathbf{G}_{0i} = \text{diag}\{\sigma_{0i1}^2\tilde{e}_{1i}, \cdots, \sigma_{0in_i}^2\tilde{e}_{in_i}\}\tilde{\mathbf{H}}_i\mathbf{a} =: \mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{a}$. Then by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\Big(I_{n2}^{(1)}\Big)^2 &= \left(\sum_{i=1}^{m}\boldsymbol{\zeta}'\tilde{\mathbf{H}}_i'\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{a}\right)^2 = \left(\sum_{k=1}^{\bar{d}}\boldsymbol{\xi}_k\sum_{i=1}^{m}\mathbf{1}_k'\tilde{\mathbf{H}}_i'\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{a}\right)^2 \\
&\leq \|\boldsymbol{\xi}\|^2\sum_{k=1}^{\bar{d}}\left(\sum_{i=1}^{m}\mathbf{1}_k'\tilde{\mathbf{H}}_i'\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{a}\right)^2 \leq \|\boldsymbol{\xi}\|^2\sum_{k,j=1}^{\bar{d}}\left(\sum_{i=1}^{m}\mathbf{1}_k'\tilde{\mathbf{H}}_i'\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{1}_j\right)^2,
\end{aligned}
$$

where $\bar{d} = d + K$ and $\mathbf{1}_k = (0, \cdots, 0, 1, 0, \cdots, 0)'$ is a $\bar{d}$ vector with 1 as its kth element and 0 elsewhere. By conditions (A1),(A3), (A5) and (A6), we have

$$
E(I_{n2}^{(1)})^2 \leq C\|\boldsymbol{\zeta}\|^2\sum_{k,j=1}^{\bar{d}}\sum_{i=1}^{m}E\left(\mathbf{1}_k'\tilde{\mathbf{H}}_i'\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i\mathbf{1}_j\right)^2
$$

$$
\begin{aligned}
&\leq C\|\boldsymbol{\zeta}\|^2 \sum_{k,j=1}^{\bar{d}} \sum_{i=1}^{m} \mathbf{1}_k' \tilde{\mathbf{H}}_i' \tilde{\mathbf{H}}_i \mathbf{1}_k E\|\mathbf{A}(\tilde{\mathbf{e}}_i)\tilde{\mathbf{H}}_i \mathbf{1}_j\|^2 \\
&\leq C\|\boldsymbol{\zeta}\|^2 \sup_i \sum_{k=1}^{\bar{d}} \mathbf{1}_k' \tilde{\mathbf{H}}_i' \tilde{\mathbf{H}}_i \mathbf{1}_k \sum_{i=1}^{m} \sum_{k=1}^{\bar{d}} \mathbf{1}_k' \tilde{\mathbf{H}}_i' \tilde{\mathbf{H}}_i \mathbf{1}_k O(k_n) \\
&= C\|\boldsymbol{\zeta}\|^2 \sup_i trace\{\tilde{\mathbf{H}}_i \tilde{\mathbf{H}}_i'\} trace\{\sum_{i=1}^{m} \tilde{\mathbf{H}}_i \tilde{\mathbf{H}}_i'\} O(k_n) \\
&\leq C\|\boldsymbol{\zeta}\|^2 k_n \sup_i trace\{\mathbf{H}_i^* \mathbf{A}_m^{-1} \mathbf{H}_i^{*\prime} + k_n \boldsymbol{\pi}_i \mathbf{Q}_m^{-2} \boldsymbol{\pi}_i'\} O(k_n) \\
&= O(k_n^3 \|\boldsymbol{\zeta}\|^2 / n),
\end{aligned}
$$

where the constant C, independent of n, may vary from line to line. Therefore, for sufficiently large L, we have $\sup_{\|\boldsymbol{\zeta}\| \leq L k_n^{1/2},\ \mathbf{a}'\mathbf{a}=1} |I_{n2}^{(1)}(\boldsymbol{\zeta})| = O_p(n^{-1/2} k_n^2)$. Similarly, we can show that $\sup_{\mathbf{a}'\mathbf{a}=1} |I_{n2}^{(2)}| = O(k_n^{1-s})$. Combining these two results and noting that $k_n = O(n^{1/(2s+1)})$, we obtain $\sup_{\mathbf{a}'\mathbf{a}=1} |I_{n2}| = O_p(k_n^{1/2})$.

For $R_m^*(\boldsymbol{\varsigma}^*)$, let $\mathbf{F}_i^* = (\partial^2 \mathbf{a}' \tilde{\mathbf{H}}_i' \mathbf{D}_i \mathbf{W}_{0i}^{-1} (\boldsymbol{\epsilon}_i^2 - \boldsymbol{\sigma}_i^2) / \partial \boldsymbol{\varsigma}_i \partial \boldsymbol{\varsigma}_i') | \boldsymbol{\varsigma}_i = \boldsymbol{\varsigma}_i^*$. We see that

$$
\begin{aligned}
R_m^*(\boldsymbol{\varsigma}^*) &= \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{\zeta}' \tilde{\mathbf{H}}_i' \mathbf{F}_i^* \tilde{\mathbf{H}}_i \boldsymbol{\xi} + \sum_{i=1}^{m} \mathbf{R}_{mi}' \mathbf{F}_i^* \tilde{\mathbf{H}}_i \boldsymbol{\zeta} + \frac{1}{2} \sum_{i=1}^{m} \mathbf{R}_{mi}' \mathbf{F}_i^* \mathbf{R}_{mi} \\
&= I_{n3}^{(1)}(\boldsymbol{\zeta}) + I_{n3}^{(2)}(\boldsymbol{\zeta}) + I_{n3}^{(3)}(\boldsymbol{\zeta}).
\end{aligned}
$$

By (A3), (A5) and (A6), we have that $\sup_{\mathbf{a}'\mathbf{a}=1} \|\mathbf{F}_i^*\| = O_p(n^{-1/2} k_n^{1/2})$. Hence

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\zeta}\| \leq L k_n^{1/2}, \mathbf{a}'\mathbf{a}=1} |I_{n3}^{(1)}(\boldsymbol{\xi})| = O_p(n^{-1/2} k_n^{5/2}), \quad \sup_{\|\boldsymbol{\zeta}\| \leq L k_n^{1/2}, \mathbf{a}'\mathbf{a}=1} |I_{n3}^{(2)}(\boldsymbol{\xi})| = O_p(k_n^{3/2-s}), \\
&\sup_{\|v\zeta\| \leq L k_n^{1/2}, \mathbf{a}'\mathbf{a}=1} |I_{n3}^{(3)}(\boldsymbol{\xi})| = O_p(n^{1/2} k_n^{1/2-2s}), \quad \sup_{\|\boldsymbol{\zeta}\| \leq L k_n^{1/2}, \mathbf{a}'\mathbf{a}=1} |R_n^*(\boldsymbol{\varsigma}^*)| = O_p(k_n^{1/2}).
\end{aligned}
$$

Putting all the approximations together, we have $\sup_{\|\boldsymbol{\zeta}\| \leq L k_n^{1/2}} \|\boldsymbol{\Psi}(\boldsymbol{\zeta}) - \boldsymbol{\Phi}(\boldsymbol{\zeta})\| = O_p(k_n^{1/2})$, and for sufficient large C, direct calculations give

$$
\begin{aligned}
E\|\tilde{\boldsymbol{\zeta}}\|^2 &= \sum_{i=1}^{m} E[(\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2)' \mathbf{W}_{0i}^{-1} \mathbf{D}_{0i} \mathbf{H}_i^* \mathbf{A}_m^{-1} \mathbf{H}_i^{*\prime} \mathbf{D}_{0i} \mathbf{W}_{0i}^{-1} (\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2) \\
&\quad + k_n (\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2)' \mathbf{W}_{0i}^{-1} \mathbf{D}_{0i} \boldsymbol{\pi}_i \mathbf{Q}_m^{-2} \boldsymbol{\pi}_i' \mathbf{D}_{0i} \mathbf{W}_{0i}^{-1} (\boldsymbol{\epsilon}_{0i}^2 - \boldsymbol{\sigma}_{0i}^2)] \\
&\leq C trace\{\mathbf{H}^* \mathbf{A}_m^{-1} \mathbf{H}^{*\prime} + k_n \mathbf{M} \mathbf{Q}_m^{-2} \mathbf{M}'\} = O(k_n).
\end{aligned}
$$

Therefore, $\sup_{\|\boldsymbol{\zeta}\|\le Lk_n^{1/2}} \|\boldsymbol{\Psi}(\boldsymbol{\zeta})-\boldsymbol{\zeta}\| \le \sup_{\|\boldsymbol{\zeta}\|\le Lk_n^{1/2}} \|\boldsymbol{\Psi}(\boldsymbol{\zeta})-\boldsymbol{\Phi}(\boldsymbol{\zeta})\| + \|\tilde{\boldsymbol{\zeta}}\| = LO_p(k_n^{1/2}) + O_p(k_n^{1/2})$, which implies that $\sup_{\|\boldsymbol{\zeta}\|\le Lk_n^{1/2}} \|\boldsymbol{\Psi}(\boldsymbol{\zeta})-\boldsymbol{\zeta}\| \le Lk_n^{1/2}$ in probability for sufficiently large L. Thus Brouwer's fixed-point theorem ensures that the map $\boldsymbol{\zeta} \mapsto \boldsymbol{\zeta}-\boldsymbol{\Psi}(\boldsymbol{\zeta})$ has a fixed point $\hat{\boldsymbol{\zeta}}$ that is a zero of $\boldsymbol{\Psi}(\boldsymbol{\zeta})$ with $\|\hat{\boldsymbol{\zeta}}\| = O_p(k_n^{1/2})$.

Lemma 2. *Under conditions (A1)-(A6), Let* $(\hat{\boldsymbol{\beta}}_m', \hat{\boldsymbol{\alpha}}_m', \hat{\boldsymbol{\gamma}}_m', \hat{\boldsymbol{\lambda}}_m', \widehat{\widetilde{\boldsymbol{\alpha}}}_m')'$ *be the root of generalized estimating equation (4), then*

$$\|\hat{\boldsymbol{\xi}}_1 - \tilde{\boldsymbol{\xi}}_1\| = o_p(1), \quad \|\sqrt{m}(\hat{\boldsymbol{\gamma}}_m - \boldsymbol{\gamma}_0) - \tilde{\boldsymbol{\gamma}}\| = o_p(1), \quad \|\hat{\boldsymbol{\zeta}}_1 - \tilde{\boldsymbol{\zeta}}_1\| = o_p(1).$$

where $\hat{\boldsymbol{\xi}}_1 = \mathbf{C}_m^{1/2}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)$, $\tilde{\boldsymbol{\xi}}_1 = \mathbf{C}_m^{1/2}\tilde{\mathbf{S}}_1$ *with* $\mathbf{C}_m = \mathbf{X}^{*\prime}\boldsymbol{\Sigma}^0\mathbf{X}^*$; $\tilde{\boldsymbol{\gamma}} = [\sum_{i=1}^m \mathbf{V}_i^{0\prime}\mathbf{D}_{0i}^{-1}\mathbf{V}_i^0/m]^{-1}\frac{1}{\sqrt{m}}\tilde{\mathbf{S}}_2$; $\hat{\boldsymbol{\zeta}}_1$ *and* $\tilde{\boldsymbol{\zeta}}_1$ *are given by (A.1)and (A.5) respectively.*

References

Diggle, P. J., Heagerty P. J., Liang, K. Y. and Zeger, S. L. (2002). Analysis of Longitudinal Data. Oxford University Press. Oxford.

Diggle, P. J. and Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, 54, 403–15.

Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147, 186-197.

Fan, J., Huang, T. and Li, R. Z. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of American Statistical Association*, 35, 632–641.

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of American Statistical Association*, 103, 1520-1533.

Härdle, W., Liang, H. and Gao, J. (2000). Partially Linear Models. Springer -Verlag, Germany.

He, X., Fung, W.K., and Zhu, Z.Y. (2005). Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, 472, 1176–1184.

He, X. and Shi, P. (1996). Bivariate tensor-product B-Splines in a partly linear model, *Journal of Multivariate Analysis*, 58, 162–181.

He, X., Zhu, Z. Y. and Fung, W. K. (2002). Estimating in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89, 579–590.

Heckman, N. E. (1986). Spline smoothing in a partly linear model, *Journal of the Royal Statistical Society, Ser. B*, 48, 244–248.

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 models. *Biometrika*, 73, 13–22.

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations, *Journal of the American Statistical Association*, 96, 1045–1056.

Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems, *Journal of the Royal Statistical Society, Series B*, 68, 69-88.

Pan, J. and Mackenzie, G. (2003). Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, 90, 239–244.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47, 825–839.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86, 677–90.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87, 425–35.

Schumaker, L. L. (1981). Spline Functions. Wiley. New York.

Speckman, P. (1988). Kernel smoothing in partly linear models. *Journal of the Royal Statistical Society, B*, 50, 413-436.

Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13, 689–705.

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.

Wang Y. G. and Carey, V. (2003). Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance. *Biometrika*, 90, 29–41.

Welsh, A. H., Lin, X. and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, 97, 482–493.

Wu, H. and Zhang, J. (2006). Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches. John Wiley and Sons, New York.

Wu, W. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 831–844

Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics*, 31, 310-347.

Ye, H. and Pan, J. (2006). Modelling of covariates structures in generalized estimating equations for longitudinal data. *Biometrika*, 93, 927–941.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50, 689–699.