Nonparametric Regression of Covariance Structures in Longitudinal Studies

Pan, Jianxin and Ye, Huajin and Li, Runze

2009

MIMS EPrint: **2009.46**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

ISSN 1749-9097

Nonparametric Regression of Covariance Structures in Longitudinal Studies

Jianxin Pan,*
The University of Manchester, UK

Huajun Ye
Peking University, China and *The University of Manchester, UK*

and Runze Li
Pennsylvania State University, USA

Summary. In this paper we propose a nonparametric data-driven approach to model covariance structures for longitudinal data. Based on a modified Cholesky decomposition, the within-subject covariance matrix is decomposed into a unit lower triangular matrix involving generalized autoregressive coefficients and a diagonal matrix involving innovation variances. Local polynomial smoothing estimation is proposed to model the nonparametric smoothing functions of the mean, generalized autoregressive coefficients and (log) innovation variances, simultaneously. We provide theoretical justification of consistency of the fitted smoothing curves in the mean, generalized autoregressive parameters and (log) innovation variances. Two real data sets are analyzed for illustration. Simulation studies are made to evaluate the efficacy of the proposed method.

Keywords: Covariance modelling; Local likelihood method; Longitudinal studies; Modified Cholesky decomposition; Modified cross validation with leave-one-subject-out; Nonparametric regression.

*Address for Correspondence: Jianxin Pan, School of Mathematics, The University of Manchester, Oxford Road, Manchester, M13 9PL UK.
Email: jianxin.pan@manchester.ac.uk

1 Introduction

Nonparametric modelling of covariance structures for longitudinal data has recently received an increasing attention. Nowadays it is well known that a good covariance estimation not only improves statistical inference of the mean but also characterizes covariance structures, which in some circumstances is of scientific importance. In the literature, Diggle and Verbyla (1998) studied nonparametric estimation of covariance structures using kernel-weighted local linear regression of sample variograms, which, though has a clear statistical interpretation, does not guarantee positive-definiteness of the estimated covariance matrices. Based on a modified Cholesky decomposition, Pourahmadi (1999; 2000) proposed parametric regressions for modelling covariance structures. Wu and Pourahmadi (2003) further considered local polynomial smoothing for certain sub-diagonals of the lower triangular matrix in the decomposition, by assuming other elements are zero. Their method is related to antedependence covariance structures (Zimmerman and Núñez-Antón, 2001) and is only applicable to balanced longitudinal data due to the involvement of the sample covariance matrix (Wu and Pourahmadi, 2003).

Local polynomial smoothing is one of the most commonly used nonparametric regression methods for modelling independent data, see Fan and Gijbels (1996) and Ruppert and Wand (1994). Recently, this technique has been applied to longitudinal data analysis, see Hoover, *et al.* (1998), Fan and Zhang (2000), Lin and Ying (2001) and references therein. This kind of work, however, ignores the within-subject correlation in longitudinal data. Based on local kernel smoothing, Lin and Carroll (2000) concluded that an independent covariance structure is actually the best one in the sense of the mean squared error, indicating that correct specification of within-cluster correlation has an adverse effect on the mean curve estimation (Lin, *et al.*, 2004). Chen and Jin (2005) explained this counter-intuitive phenomenon is due to a mismatch of 'local observations with global variances', and proposed an alternative approach based on the principle of 'local observations with local variances'. In other words, only covariances of local observations make a contribution to the mean curve estimation when using local polynomial smoothing. In this manner, correct specification of within-subject correlation does have a positive effect on the mean curve estimation.

In this paper a nonparametric data-driven approach is proposed to model the mean and covariance structures for longitudinal data. Based on a modified Cholesky decomposition, the within-subject covariance matrix is decomposed into a unit lower triangular matrix and a diagonal matrix (Pourahmadi, 1999). A smoothing function in time is then used to model the off-diagonal elements (also known as generalized autoregressive coefficients) of the lower triangular matrix, and the logarithm of the diagonal elements (also known as log-innovation variances) of the diagonal matrix. A local polynomial smoothing approach (Fan, *et al.*, 1998) is applied to fit smoothing curves in the mean, generalized autoregressive coefficients and log-innovation variances. In the spirit of 'local observations with local variances' (Chen and Jin, 2005), the mean curve estimation takes into account the within-subject correlation for longitudinal data. We show that the fitted smoothing curves have very nice theoretical properties. We also propose a modified cross-validation criterion with leave-one-subject-out to select the best value of the bandwidth. For illustration, real data analysis and simulation studies are conducted, showing the proposed approach works well for modelling the mean and covariance structures, simultaneously.

This paper is organized as follows. In Section 2, we propose smoothing curve models for the mean, generalized autoregressive parameters and log-innovation variances. In Section 3, we develop a local polynomial based approach for modelling the mean and covariance structures for longitudinal data, and then discuss the issue of bandwidth selection. In Section 4, asymptotic properties of the proposed estimation approaches are studied, with technical details included in the Appendices. In Section 5 the proposed approach is illustrated by real data analysis and simulation studies. Some further discussions are provided in Section 6.

2 Nonparametric models for mean-covariance structures

Let y_{ij} be the jth of m_i measurements on the ith of n subjects. Assume that t_{ij} is the time at which the measurement y_{ij} is made. Denote by $y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$ and $t_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})'$ be the $(m_i \times 1)$ vector of responses and measuring times of the ith subject, respectively. Suppose $\mathrm{E}(y_i) = \mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})'$ and $\mathrm{Var}(y_i) = \Sigma_i$ are the $(m_i \times 1)$ mean vector and $(m_i \times m_i)$ positive definite variance-covariance matrix of y_i, respectively. Throughout this paper, we assume y_i follows a normal distribution, i.e., $y_i \sim \mathcal{N}_{m_i}(\mu_i, \Sigma_i)$ for $i = 1, 2, \ldots, n$.

Accordingly, there exits a unique lower triangular matrix T_i with 1's as diagonal entries and a unique diagonal matrix D_i with positive diagonals such that $T_i \Sigma_i T_i' = D_i$. This modified Cholesky decomposition has a clear statistical interpretation, that is, the below-diagonal entries of T_i are the negatives of the autoregressive coefficients, ϕ_{ijk}, in the autoregressive model

$$\hat{y}_{ij} = \mu_{ij} + \sum_{k=1}^{j-1} \phi_{ijk}(y_{ik} - \mu_{ik}). \tag{2.1}$$

In other words, (2.1) is the linear least squares predictor of y_{ij} based on its predecessors $y_{i(j-1)}, \ldots, y_{i1}$. Note that as $j = 1$ the notation $\sum_{k=1}^{0}$ means zero. It can also be shown that the diagonal entries of D_i are the prediction error/innovation variances $\sigma_{ij}^2 = \mathrm{var}(\varepsilon_{ij})$ where $\varepsilon_{ij} = y_{ij} - \hat{y}_{ij}$ and $\hat{y}_{ij}$ are given in (2.1) $(1 \le j \le m_i; 1 \le i \le n)$ (Pourahmadi, 1999). Obviously, we have $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$ if $j \ne k$. We refer to ϕ_{ijk} as generalized autoregressive parameters and σ_{ij}^2 as innovation variances. It follows immediately that $\Sigma_i^{-1} = T_i' D_i^{-1} T_i$.

We propose to use three nonparametric smoothing functions in time to model the unconstrained parameters μ_{ij}, ϕ_{ijk} and $\log \sigma_{ij}^2$ as follows

$$\mu_{ij} = f(t_{ij}), \qquad \phi_{ijk} = g(t_{ij}, t_{ik}) \quad \text{and} \quad \log \sigma_{ij}^2 = q(t_{ij}), \tag{2.2}$$

where the unknown functions $f(t)$ and $q(t)$ are univariate but $g(t,s)$ may be two-dimensional. For simplicity, we only assume that ϕ_{ijk} depends on the two time points t_{ij} and t_{ik} through $g(\cdot,\cdot)$. When longitudinal data are stationary, the function g reduces to one-dimensional function in time lag, that is, $g(t,s) \equiv g(t-s)$. Clearly, the models in (2.2) are very broad and contain many commonly used covariance structures, such as compound symmetric, AR(1), antedependence structure and others. We refer to the models in (2.2) as nonparametric mean-covariance models.

Pourahmadi (1999) and Pan and MacKenzie (2003) considered parametric modelling of μ_{ij}, ϕ_{ijk} and $\log\sigma_{ij}^2$ in (2.2) for balanced and unbalanced longitudinal data, respectively. Ye and Pan (2006) discussed parametric modelling of covariance structures for unbalanced longitudinal data within the framework of generalized estimation equations. Wu and Pourahmadi (2003) studied local polynomial smoothing for a few of subdiagonals of the lower triangular matrix T_i by assuming others are zero. This differs from our smoothing approach in the sense that we model the whole correlation function surface rather than only a few of the profile functions on the surface. As long as the estimated smoothing functions $\hat{g}(t,s)$ and $\hat{q}(t)$ can be obtained, the estimators $\hat{T}_i$ and $\hat{D}_i$ of the matrices T_i and D_i can be constructed according to (2.2). The estimators of the covariance matrices are then formed by $\hat{\Sigma}_i = \hat{T}_i^{-1}\hat{D}_i\hat{T}_i^{\prime -1}$, implying that the positive definiteness of $\hat{\Sigma}_i$ is assured. Furthermore, baseline covariates of interest may be included in the models in (2.2) as a parametric part, thus creating semiparametric models for the mean and covariance structures. In this paper we focus on the nonparametric regression models. The extension to semiparametric models is straightforward and not considered here.

3 Local polynomial smoothing of mean-covariance structures

3.1 Local maximum likelihood estimation

Local maximum likelihood estimation was proposed by Fan, *et al.* (1998), which can be viewed as a nonparametric counterpart of the widely used parametric maximum likelihood technique. Suppose that the ith univariate observation y_i and its measurement time t_i have a contribution $\ell\{b(t_i), y_i\}$ to the log-likelihood, where $b(\cdot)$ is an unknown smooth function of interest and is assumed to have a continuous $(p+1)$th derivative at the time point t_0. Then for t_i in a neighborhood of t_0, a polynomial of degree p is used to approximate $b(t_i)$ via Taylor expansion:

$$b(t_i) \approx b(t_0) + (t_i - t_0)b'(t_0) + \cdots + (t_i - t_0)^p \frac{b^{(p)}(t_0)}{p!} \equiv x_i'\beta, \tag{3.1}$$

where $x_i = (1, (t_i - t_0), \ldots, (t_i - t_0)^p)'$ and $\beta = (\beta_0, \ldots, \beta_p)'$ with $\beta_v = b^{(v)}(t_0)/v!$ $(v = 0, \ldots, p)$. The local kernel-weighted log-likelihood is defined by

$$\ell_p(\beta, h, t_0) = \sum_{i=1}^{n} \ell\{x_i'\beta, y_i\} K_h(t_i - t_0), \tag{3.2}$$

where $K_h(t_i - t_0) = K\left((t_i - t_0)/h\right)/h$, $K(\cdot)$ is a known kernel function and h is a bandwidth parameter. The estimator of β that maximizes the local kernel-weighted log-likelihood, i.e.,

$$\hat{\beta}(t_0) = \arg\max_{\beta}\{\ell_p(\beta, h, t_0)\} \tag{3.3}$$

is called the local maximum likelihood estimator of β at t_0 (Fan, *et al.*,1998). When varying t_0, we obtain the estimated function $\hat{\beta}(t)$. Obviously, the first component $\hat{\beta}_0(t)$ is the local maximum likelihood estimation function of $b(t)$. For further details on local maximum likelihood estimation, one can refer to Fan *et al.* (1998).

3.2 Estimation of the mean function

Recall the ith subject responses $y_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})'$ and the measurement time $t_i = (t_{i1}, t_{i2}, \ldots, t_{im_i})'$. Since $y_i \sim \mathcal{N}_{m_i}(\mu_i, \Sigma_i)$, then the log-likelihood function ℓ, apart from a constant, can be written as

$$-2\ell = \sum_{i=1}^{n} \log|\Sigma_i| + \sum_{i=1}^{n} (y_i - f(t_i))' \Sigma_i^{-1} (y_i - f(t_i)), \tag{3.4}$$

where $\mu_i = f(t_i) = (f(t_{i1}), \ldots, f(t_{im_i}))'$ and $f(t)$ is the mean smooth function in (2.2). Assume that in a neighborhood of t_0 the mean function $f(t_{ij})$ has the Taylor expansion (3.1), i.e., $f(t_{ij}) = x_{ij}'\beta$ with $x_{ij} = (1, (t_{ij} - t_0), \ldots, (t_{ij} - t_0)^p)'$ and $\beta = (\beta_0, \ldots, \beta_p)'$ with $\beta_v = f^{(v)}(t_0)/v!$ $(v = 0, \ldots, p)$. Note that the first term in (3.4) does not make any contribution when estimating the mean function. In the spirit of Chen and Jin (2005) we propose to minimize the following partial local-weighted log-likelihood,

$$Q_1(\beta) = \sum_{i=1}^{n} (y_i - X_i'\beta)' K_{ih_1}^{1/2}(t_0) (I_i \Sigma_i I_i)^{-1} K_{ih_1}^{1/2}(t_0) (y_i - X_i'\beta) \tag{3.5}$$

with respect to β, where $X_i = (x_{i1}, \ldots, x_{im_i})'$, $K_{ih_1}(t_0) = \mathrm{diag}\{K_{h_1}(t_{i1} - t_0), \ldots, K_{h_1}(t_{im_i} - t_0)\}$ with the entries $K_{h_1}(t_{ij} - t_0) = K((t_{ij} - t_0)/h_1)/h_1$, and $I_i = \mathrm{diag}\{I(|t_{i1} - t_0| \leq h_1), \ldots, I(|t_{im_i} - t_0| \leq h_1)\}$ where $I(\cdot)$ is the indicator function and h_1 is the bandwidth. Typical choices of the kernel function $K(\cdot)$ include the Gaussian density $K(t) = (1/\sqrt{2\pi}) \exp\{-t^2/2\}$ and Epanechnikov kernel $K(t) = 0.75(1 - t^2) I(|t| \leq 1)$. It is noted that the matrix $I_i \Sigma_i I_i$ in (3.5) may be singular and so $(I_i \Sigma_i I_i)^{-1}$ represents the Moore-Penrose generalized inverse of the matrix $I_i \Sigma_i I_i$.

Note that the proposed criterion in (3.5) is different from the one proposed by Chen and Jin (2005) in that they assume the localization through I_i is only made to the correlation matrix with the variances being treated as nuisance parameters. In contrast, we propose to localize the variance-covariance matrices Σ_i when estimating the mean function. By minimizing $Q_1(\beta)$ over $\beta = (\beta_0, \ldots, \beta_p)'$, one can easily obtain the local maximum likelihood estimator

$$\hat{\beta}(t_0) = \left(\sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i' V_i^{-1} y_i \right), \tag{3.6}$$

where $V_i^{-1} = K_{ih_1}^{1/2}(t_0)(I_i\Sigma_i I_i)^{-1}K_{ih_1}^{1/2}(t_0)$. Denote $\hat{f}_\nu(t_0) = \nu!\hat{\beta}_\nu(t_0)$ as the estimator of the νth derivative $f^{(\nu)}(t_0)$ for $\nu = 0, 1, \ldots, p$. In particular, $\hat{f}_0(t_0)$ is the estimator of $f(t_0)$. By varying t_0 we obtain the estimator $\hat{f}_0(t) \equiv \hat{f}(t)$ of the mean function $f(t)$.

The calculation of the estimator $\hat{\beta}(t_0)$ in (3.6) assumes that the within-subject covariance matrices Σ_i are known. When unknown they can be replaced by the estimators $\hat{\Sigma}_i$, which are given in the next subsections.

3.3 Estimation of the generalized autoregressive parameter function

Applying the modified Cholesky decomposition to Σ_i, the log-likelihood function ℓ in (3.4) can be rewritten as

$$-2\ell = \sum_{i=1}^{n} \log|D_i| + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ij} - \sum_{k=1}^{j-1}\phi_{ijk}r_{ik}\right)^2 / \sigma_{ij}^2, \tag{3.7}$$

where $r_{ij} = y_{ij} - \mu_{ij}$ with $\mu_{ij} = f(t_{ij})$ and $D_i = \mathrm{diag}(\sigma_{i1}^2, \ldots, \sigma_{im_i}^2)$. The generalized autoregressive parameters ϕ_{ijk} are modelled in terms of a smooth function $g(t_{ij}, t_{ik})$, which may be one or two-dimensional, depending on whether or not the longitudinal data is stationary. Below we consider the general case. Similar to the mean function estimation, for the points (t_{ij}, t_{ik}) $(t_{ik} < t_{ij})$ in a neighborhood of (t_0, s_0) $(s_0 < t_0)$, the function $g(t, s)$ $(s < t)$ can be approximated by the Taylor expansion, for example, only through the linear term

$$g(t_{ij}, t_{ik}) \approx g(t_0, s_0) + (t_{ij} - t_0)g_t'(t_0, s_0) + (t_{ik} - s_0)g_s'(t_0, s_0) \equiv Z_{ijk}'\gamma, \tag{3.8}$$

where $g_t'(t_0, s_0) = \partial g(t, s)/\partial t|_{t=t_0, s=s_0}$ and $g_s'(t_0, s_0) = \partial g(t, s)/\partial s|_{t=t_0, s=s_0}$, whereas $Z_{ijk} = (1, (t_{ij} - t_0), (t_{ik} - s_0))'$ and $\gamma = (g(t_0, s_0), g_t'(t_0, s_0), g_s'(t_0, s_0))'$ are (3×1) vectors. The use of high-order Taylor expansion of $g(t, s)$ is straightforward but Z_{ijk} and γ are more than three-dimensional in this case.

Motivated by the local maximum likelihood estimation and using (3.7) and (3.8), we propose to minimize the partial local-weighted log-likelihood

$$Q_2(\gamma) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ij} - \sum_{k=1}^{j-1}(Z'_{ijk}\gamma)r_{ik}K_{h_2}^{1/2}(t_{ij}-t_0, t_{ik}-s_0)\right)^2 / \sigma_{ij}^2 \tag{3.9}$$

with respect to γ, where $K_{h_2}(t-t_0, s-s_0)$ is the two-dimensional kernel function, which is usually chosen as $K_{h_2}(t-t_0, s-s_0) = K(\{(t-t_0)^2 + (s-s_0)^2\}/h_2^2)/h_2$ where $K(\cdot)$ is the univariate kernel function. In (3.9), h_2 is the bandwidth parameter for estimating $g(\cdot,\cdot)$, which in general is different from h_1, the bandwidth parameter when estimating the mean function $f(t)$. The following proposition provides the solution to the equation $\partial Q_2(\gamma)/\partial\gamma = 0$.

PROPOSITION 3.1. *Let* $Z_{(i,j)} = \sum_{k=1}^{j-1} r_{ik}K_{h_2}^{1/2}(t_{ij}-t_0, t_{ik}-s_0)Z_{ijk}$ *and* $Z_{(i)} = (Z_{(i,1)}, \ldots, Z_{(i,m_i)})'$. *Minimizing the partial local-weighted log-likelihood* $Q_2(\gamma)$ *in (3.9) over* γ *leads to the local maximum likelihood estimator*

$$\hat{\gamma}(t_0, s_0) = \left(\sum_{i=1}^{n} Z'_{(i)}D_i^{-1}Z_{(i)}\right)^{-1}\left(\sum_{i=1}^{n} Z'_{(i)}D_i^{-1}r_i\right), \tag{3.10}$$

where $r_i = y_i - \mu_i \equiv (r_{i1}, \ldots, r_{im_i})'$ *is the* $(m_i \times 1)$ *residual vector.*

Proof: See Appendix A.

Proposition 3.1 provides an explicit solution of the local maximum likelihood estimator of the parameter γ at (t_0, s_0). When varying (t_0, s_0) over the range $\{(t,s) : s < t \text{ and } (t,s) \in \mathbb{R}^2\}$, we obtain the estimated function $\hat{\gamma}(t,s)$. The first component $\hat{\gamma}_0(t,s)$ is then the estimated function of the generalized autoregressive parameter function $g(t,s)$.

It is noted that when calculating the estimator of γ we assume that the mean and innovation variance are known. When unknown, they can be replaced by the corresponding estimators. For example, the mean parameter β can be replaced by its estimator (3.6) and the innovation variance parameter λ be replaced by the estimator given below.

3.4 Estimation of the innovation variance function

When written as a function of the innovation variances σ_{ij}^2, the log-likelihood ℓ in (3.4) has the alternative form

$$-2\ell = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\log\sigma_{ij}^2 + (r_{ij} - \hat{r}_{ij})^2 \exp\left\{-\log\sigma_{ij}^2\right\}\right), \tag{3.11}$$

where $r_{ij} = y_{ij} - \mu_{ij}$ and $\hat{r}_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} r_{ik}$. By using the Taylor expansion in a neighborhood of t_0, the innovation variance function $q(t_{ij})$ can be approximated by $q(t_{ij}) = z_{ij}'\lambda$ where $z_{ij} = (1, (t_{ij} - t_0), \ldots, (t_{ij} - t_0)^d)'$ and $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_d)'$ with $\lambda_\nu = q^{(\nu)}(t_0)/\nu!$ $(\nu = 0, 1, \ldots, d)$. Therefore, similar to (3.2) we propose to minimize the following local-weighted function

$$Q_3(\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(z_{ij}'\lambda + (r_{ij} - \hat{r}_{ij})^2 \exp\left(-z_{ij}'\lambda\right)\right) K_{h_3}(t_{ij} - t_0) \tag{3.12}$$

with respect to λ, where $K_{h_3}(t_{ij} - t_0)$ is the standardized kernel function, i.e., $K_{h_3}(t_{ij} - t_0) = K((t_{ij} - t_0)/h_3)/h_3$, and h_3 is the bandwidth for estimating $q(t_0)$.

In contrast to $Q_1(\beta)$ and $Q_2(\gamma)$, minimizing $Q_3(\lambda)$ does not enjoy analytical solution as the score function $\partial Q_3(\lambda)/\partial\lambda$ is nonlinear in λ. Therefore, in order to calculate the estimator of λ, we have to look for an iterative solution, for example, using the New-Raphson algorithm. The following proposition gives the Newton-Raphson iterative solution for obtaining the local maximum likelihood estimator of the innovation variance parameter λ at t_0.

PROPOSITION 3.2. *Let* $V_i^{-1} = diag\{(r_{i1} - \hat{r}_{i1})^2/\sigma_{i1}^2, \ldots, (r_{im_i} - \hat{r}_{im_i})^2/\sigma_{im_i}^2\}$, $K_{ih_3}(t_0) = diag\{K_{h_3}(t_{i1} - t_0), \ldots, K_{h_3}(t_{im_i} - t_0)\}$ *and* $W_i^{-1} = V_i^{-1/2} K_{ih_3}(t_0) V_i^{-1/2}$ $(i = 1, 2, \ldots, n)$. *Then the Newton-Raphson solution for minimizing the local-weighted log-likelihood function* $Q_3(\lambda)$ *at* t_0, *can be iterated through*

$$\hat{\lambda}(t_0) = \left(\sum_{i=1}^{n} Z_i' W_i^{-1} Z_i\right)^{-1}\left(\sum_{i=1}^{n} Z_i' W_i^{-1} u_i\right), \tag{3.13}$$

where $Z_i = (z_{i1}, \ldots, z_{im_i})'$, $u_i = (logD_i + I_{m_i} - V_i)1_{m_i}$ *is the working response with* $logD_i = diag\{log\sigma^2_{i1}, \ldots, log\sigma^2_{im_i}\}$, *and* I_{m_i} *and* 1_{m_i} *are the* $(m_i \times m_i)$ *identity matrix and the* $(m_i \times 1)$ *vector of 1's, respectively.*

Proof: See Appendix A.

It is noted that when updating the estimator of λ we assume the mean and generalized autoregressive parameters are known. When unknown, they can be replaced by the corresponding estimators. In other words, the mean parameter β and the generalized autoregressive parameter γ can be replaced by their estimators in (3.6) and (3.10), respectively.

3.5 Bandwidth selection

It is well established that the choice of bandwidth parameters is more crucial than that of the kernel function. A selection criterion is needed to find the optimal value of bandwidth, which actually controls the 'local neighborhood' of observations. Obviously, the wider the local neighborhood the less rough the smoothing curve. Without loss of generality, we propose to use a constant bandwidth for each of the three smoothing functions. In the literature, cross-validation method was exclusively used to choose the optimal value of bandwidth. For longitudinal data, as suggested by Rice and Silverman (1991) we propose to use the cross-validation method with leave-one-subject-out, rather than leave-one-observation-out, to choose the optimal bandwidth. Similar strategies were considered by Hart and Wehrly (1993) where consistency of the principle was studied. Specifically, we use the modified cross-validation with leave-one-subject-out (MCVS) criteria for the mean and covariance smoothing functions as follows:

$$
\begin{aligned}
\mathrm{MCVS}(h_1) &= \sum_{i=1}^{n}\left(y_i-\hat{f}^{(-i)}(t_i)\right)'\Sigma_i^{-1}\left(y_i-\hat{f}^{(-i)}(t_i)\right) && (3.14)\\
\mathrm{MCVS}(h_2) &= \sum_{i=1}^{n}\left(r_i-\hat{r}_i^{(-i)}\right)'D_i^{-1}\left(r_i-\hat{r}_i^{(-i)}\right) && (3.15)\\
\mathrm{MCVS}(h_3) &= \sum_{i=1}^{n}\left(\varepsilon_i^2-\exp\left(\hat{q}^{(-i)}(t_i)\right)\right)'\Omega_i^{-1}\left(\varepsilon_i^2-\exp\left(\hat{q}^{(-i)}(t_i)\right)\right) && (3.16)
\end{aligned}
$$

where in (3.14) $\hat{f}^{(-i)}(t_i)$ denotes the estimator of $f(t)$ using the data when the ith subject is excluded. In (3.15) $\hat{r}_i^{(-i)} = (0, \hat{r}_{i2}^{(-i)}, \ldots, \hat{r}_{im_i}^{(-i)})'$ with $\hat{r}_{ij}^{(-i)} = \sum_{k=1}^{j-1}\hat{g}^{(-i)}(t_{ij}, t_{ik})(y_{ik} - f(t_{ik}))$ where $\hat{g}^{(-i)}(t_{ij}, t_{ik})$ is the estimator of $g(t,s)$ under the reduced dataset. In (3.16) $\varepsilon_i^2 = (\varepsilon_{i1}^2, \ldots, \varepsilon_{im_i}^2)'$ with $\varepsilon_{ij} = r_{ij} - \hat{r}_{ij}$ and $\hat{q}^{(-i)}(t_i)$ is the estimator of $q(t)$ using the data without the ith subject. It can be shown that $D_i = \mathrm{diag}\left\{\sigma_{i1}^2, \ldots, \sigma_{im_i}^2\right\}$ is actually the covariance matrix of the residuals r_i. It is noted that in (3.16) Ω_i is the covariance matrix of ε_i^2 and is given by $\Omega_i = \mathrm{Var}(\varepsilon_i^2) = 2\mathrm{diag}\{\sigma_{i1}^4, \ldots, \sigma_{im_i}^4\}$. In fact, as $y_i \sim \mathcal{N}_{m_i}(\mu_i, \Sigma_i)$ we have $\varepsilon_i \sim \mathcal{N}_{m_i}(0, D_i)$ according to the definition of $\varepsilon_i = T_i(y_i - \mu_i)$, see Section 2. Since D_i is diagonal, we know that ε_{ij} and $\varepsilon_{ik} (j \neq k)$ are independent and so do ε_{ij}^2 and $\varepsilon_{ik}^2 (j \neq k)$. On the other hand, as $\varepsilon_{ij}^2/\sigma_{ij}^2 \sim \chi_1^2$ it is obvious that $\mathrm{Var}(\varepsilon_{ij}^2) = 2\sigma_{ij}^4$ so that we have $\Omega_i = 2\mathrm{diag}\{\sigma_{i1}^4, \ldots, \sigma_{im_i}^4\}$. In contrast to the responses y_i (3.14), r_i and ε_i^2 play a role of responses in $\mathrm{MCVS}(h_2)$ and $\mathrm{MCVS}(h_3)$, respectively.

We then minimize $\mathrm{MCVS}(h_l)$ to obtain the optimal value of the bandwidth, that is, $\hat{h}_l = \arg\min_{h_l>0}\{\mathrm{MCVS}(h_l)\}$ for $l = 1, 2, 3$. Note that $\mathrm{MCVS}(h_1)$ and $\mathrm{MCVS}(h_2)$ can be easily calculated by using the estimators in (3.6) and (3.10) as the score functions of $Q_1(\beta)$ and $Q_2(\gamma)$ are linear in β and γ, respectively. But the calculation of $\mathrm{MCVS}(h_3)$ is difficult because the score function of $Q_3(\lambda)$ is nonlinear in λ. As a result, it may be very time-consuming for finding the optimal value that minimizes $\mathrm{MCVS}(h_3)$ in the whole range of $h_3 > 0$. Instead, we propose to pre-select a subarea of the bandwidth and then minimize the MCVS criterion in the subarea. This strategy is applied to the bandwidth selection in the numerical studies which will be presented later, and works reasonably well.

3.6 Backfitting algorithm

Based on the principles described above, we propose a backfitting algorithm to calculate the estimators of the smoothing functions in the mean, generalized autoregressive coefficients and log-innovation variances.

Step 1. Given starting values of $\hat{g}^{(0)}(t,s)$ and $\hat{q}^{(0)}(t)$ of the functions $g(t,s)$ and $q(t)$, the local weighted least square estimator of the mean function $\hat{f}^{(0)}(t)$ can be extracted from the first component of the vector $\hat{\beta}^{(0)}(t)$ given by

$$\hat{\beta}^{(0)}(t) = \left(\sum_{i=1}^{n} X_i' \hat{V}_i^{(0)-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' \hat{V}_i^{(0)-1} y_i\right)$$

where $\hat{V}_i^{(0)-1} = K_{\hat{h}_1^{(0)}}^{1/2}(t_i - t)\hat{\Sigma}_i^{(0)-1} K_{\hat{h}_1^{(0)}}^{1/2}(t_i - t)$ and $\hat{\Sigma}_i^{(0)}$ is formed by $\hat{g}^{(0)}(t,s)$ and $\hat{q}^{(0)}(t)$. The bandwidth estimator $\hat{h}_1^{(0)}$ is obtained by minimizing the MCVS(h_1) in (3.14).

Step 2. Denote $\hat{r}_{ij}^{(0)} = y_{ij} - \hat{f}^{(0)}(t_{ij})$ $(i = 1, \ldots, n,\ j = 1, \ldots, m_i)$. Form $\hat{u}_i^{(0)}$, $\hat{Z}_{(i)}^{(0)}$ and $\hat{W}_i^{(0)}$ $(i = 1, \ldots, n)$ defined in Propositions 3.1 and 3.2. Then the local weighted least square estimators $\hat{g}^{(1)}(t,s)$ and $\hat{q}^{(1)}(t)$ of the functions $g(t,s)$ and $q(t)$ in the generalized autoregressive parameters and log-innovation variances are the first components of $\hat{\gamma}^{(1)}(t)$ and $\hat{\lambda}^{(1)}(t)$, respectively, where

$$\begin{aligned}
\hat{\gamma}^{(1)}(t) &= \left(\sum_{i=1}^{n} \hat{Z}_{(i)}'^{(0)} \hat{D}_i^{(0)-1} \hat{Z}_{(i)}^{(0)}\right)^{-1} \left(\sum_{i=1}^{n} \hat{Z}_{(i)}'^{(0)} \hat{D}_i^{(0)-1} \hat{r}_i^{(0)}\right) \\
\hat{\lambda}^{(1)}(t) &= \left(\sum_{i=1}^{n} Z_i' \hat{W}_i^{(0)-1} Z_i\right)^{-1} \left(\sum_{i=1}^{n} Z_i' \hat{W}_i^{(0)-1} \hat{u}_i^{(0)}\right)
\end{aligned}$$

The bandwidth estimators $\hat{h}_2^{(0)}$ and $\hat{h}_3^{(0)}$ are obtained by minimizing MCVS(h_2) in (3.15) and MCVS(h_3) in (3.16), respectively.

Step 3. Use the resulting estimators $\hat{g}^{(1)}(t,s)$ and $\hat{q}^{(1)}(t)$ to replace $\hat{g}^{(0)}(t,s)$ and $\hat{q}^{(0)}(t)$, respectively, and then repeat Step 1 and Step 2 above until convergence.

A convenient starting value for $\hat{g}^{(0)}(t,s)$ and $\hat{q}^{(0)}(t)$ may be taken as $\hat{g}^{(0)}(t,s) = 0$ $(s \le t)$ and $\hat{q}^{(0)}(t) = 0$. In other words, the $(m_i \times m_i)$ identity matrix can be chosen as the starting value for estimating the covariance matrix Σ_i.

4 Asymptotic properties

In this section we discuss the asymptotic properties of the nonparametric smoothing estimators in (3.6), (3.10) and (3.13). Without loss of generality, we only consider the scenario of stationary longitudinal data. In other words, the lower triangular elements ϕ_{ijk} in T_i only depend on the lag in time through a univariate smoothing function $g(t^*)$ where $t^* \equiv t - s$ with $s < t$.

We naturally assume that all time points t_{ij} are positive, i.e., $t_{ij} > 0$. If some $t_{ij} < 0$ (e.g., after centralization), a shift of the times is necessary to ensure all t_{ij} are positive. Below we use the principle of a counting process (Lin and Ying, 2001) to obtain asymptotic results of the nonparametric smoothing estimators. The counting process which characterizes the distribution of the time points t_{ij}, is defined by

$$N_i(t) = \sum_{j=1}^{m_i} I(t_{ij} \le t) \tag{4.1}$$

where $I(\cdot)$ is the indicator function. Assume that the response function $y_i(t)$ and the mean function $f(t)$ are observed at each of the jump points of the counting processes $N_i(t)$ and $N(t)$, respectively. Using the notation of the counting process we rewrite the local least squares (3.5) as:

$$\sum_{i=1}^{n} \int_0^{\infty} \left(y_i(t) - X_i'(t)\beta\right)' K_{h_1}^{1/2}(t-t_0)(I_i\Sigma_i I_i)^{-1} K_{h_1}^{1/2}(t-t_0)\left(y_i(t) - X_i'(t)\beta\right) dN_i(t) \tag{4.2}$$

Suppose that $\varphi_i(t)$ and $\varphi(t)$ are the intensity functions of the processes $N_i(t)$ and $N(t)$, respectively. The following mild regularity conditions are necessary for the technical proofs provided in Appendices B and C.

(i) *The smooth functions $f(t)$, $g(t^*)$, and $q(t)$ are all bounded and have continuous $(p+1)th$, $(b+1)th$ and $(d+1)th$ derivatives, respectively, where p, b and d are fixed integers.*

(ii) *The elements of the covariance matrices $\Sigma_i(t)$ =Var($Y_i|t$) and of the inverses $\Sigma_i^{-1}(t)$ are all bounded and continuous in time. Furthermore, the third and fourth moments of $Y_i(t)$ exist and are all bounded.*

(iii) *The intensity functions $\varphi_i(t)$ and $\varphi(t)$ are positive, bounded and differentiable.*

(iv) *The kernel function $K(\cdot)$ is a bounded density function satisfying $\int_R K(u)du = 1$, $\int_R u^{2i}K^j(u)du < +\infty$, and $\int_R u^{2i-1}K^j(u)du = 0$ ($j = 1, 2$; $i = 1, 2, 3, ...$).*

(v) *The functional spaces $\mathcal{C}$, $\mathcal{B}$ and $\mathcal{V}$ for the local parameters $\beta(t)$, $\gamma(t)$ and $\lambda(t)$ are all compact. In other words, each space above is closed and bounded.*

(vi) *The bandwidth parameters, h_ν, satisfy $h_\nu \to 0$ and $nh_\nu \to \infty$ as $n \to \infty$ ($\nu = 1, 2, 3$).*

We also denote

$$c_K = \int_R u^2K(u)du \quad \text{and} \quad d_K = \int_R K^2(u)du.$$

Note that the conditions imposed on $K(\cdot)$, which are for convenience of technical proofs, are very regular but can be relaxed further. Let $\epsilon(t)$ represent the stochastic process of the random error having the observed realization $\epsilon_i = (\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{im_i}))'$. It is actually the random error in the nonparametric regression model $y_i = f(t_i) + \epsilon_i$ ($i = 1, 2, \ldots, n$). A sequence of function estimator $\hat{f}(t)$ is said to be a consistent estimator of the function $f(t)$ if for any t the sequence $\hat{f}(t)$ converges to $f(t)$ in probability, that is $\hat{f}(t) \xrightarrow{P} f(t)$ as $n \to \infty$. The theorems below provide the consistencies of the mean and covariance smooth function estimators when the local polynomial estimation method is used.

Theorem 1. Suppose the regularity conditions (i)-(vi) above are satisfied. When using the local maximum likelihood estimation method in (3.6) to model the mean function $f(t)$, the resulting estimator $\hat{f}(t)$ has the conditional mean square errors:

$$E[(\hat{f}(t)-f(t))^2|t_1,\ldots,t_n]=\frac{1}{4}h_1^4(f''(t))^2c_K^2+\frac{Var(\epsilon(t))d_K}{(nh_1)\varphi(t)}+o_p(h_1^4+\frac{1}{nh_1}) \tag{4.3}$$

Therefore, $\hat{f}(t)$ converges to $f(t)$ in probability, that is, $\hat{f}(t)\xrightarrow{P}f(t)$ as $n\to\infty$.

Proof: See Appendix B.

Theorem 2. Suppose the regularity conditions (i)-(vi) above are satisfied. When using the local maximum likelihood estimation methods in (3.10) and (3.13) to model the generalized autoregressive coefficients function $g(t^*)$ and the log-innovation variances function $q(t)$, the resulting estimators $\hat{g}(t^*)$ and $\hat{q}(t)$ converge to $g(t^*)$ and $q(t)$ in probability, respectively, that is, $\hat{g}(t^*)\xrightarrow{P}g(t^*)$ and $\hat{q}(t)\xrightarrow{P}q(t)$ as $n\to\infty$.

Proof: See Appendix C.

5 Numerical studies

5.1 Cattle data

We reanalyze Kenward (1987)'s Cattle data in which 60 cattle were assigned randomly to two treatment groups A and B. Half the cattle received treatment A and the other half received treatment B. The cattle were weighted 11 times over 133-day period at 0, 14, 28, 42, 56, 70, 84, 98, 112, 126 and 133 in days. The objective of the study was to investigate treatment effects on intestinal parasites.

This data set was analyzed by many authors, including Pourahmadi (1999), Wu and Pourahmadi (2003), and Pan and Mackenzie (2003). For illustration, the proposed non-parametric estimation method is used to model the mean and covariance structures for the treatment A data only. Without loss of generality, we assume a univariate smooth function $g(t^*_{ijk})$ for modelling the generalized autoregressive parameters ϕ_{ijk} where $t^*_{ijk} \equiv t_{ij} - t_{ik}$. In other words, we assume a stationary correlation structure for the data.

Figure 1 is about here.

The bandwidth parameter estimators, obtained by using the MCVS method, are $h_1 = 0.80$, $h_2 = 0.80$ and $h_3 = 0.55$, respectively. Note that the cattle data are balanced, and so in Figure 1 we plot the sample generalized autoregressive coefficients and sample log-innovation variances against time/lag. We also display the fitted curves for the mean and covariance structures using the proposed nonparametric smoothing methods. Figure 1 shows that the estimated smoothing curves fit the data well, though for the log-innovation variances there is a slight departure from the samples at the right corner. This may be due to the fact that the log transformation made to the innovation variances can be further improved. The estimated smoothing curves are very similar to the ones obtained by Pan and MacKenzie (2003), where three polynomials in time, one of degree 8 and two cubics, were used to model the mean parameters, the generalized autoregressive parameters and the log-innovation variances. The nonparametric estimators of the mean and covariance structures are much more parsimonious and can be regarded as a guide to the formulation of parametric models.

Table 1 gives the correlation matrix obtained by the resulting nonparametric smoothing estimators. We also compare our estimators to those obtained by Wu and Pourahmadi (2003), where they only model the first two subdiagonals of $\hat{T}$ and assume others to be zero. Table 1 shows some differences in the estimated correlation matrix obtained by the two approaches.

Table 1 is about here.

5.2 CD4+ Cell data

CD4+ data comprise CD4+ cell counts for 369 HIV-infected men (Diggle, *et al.*, 2002). Altogether there are 2376 values of CD4+ cell counts, with several repeated measurements being made for each individual at different times covering a period of approximately eight and a half years. For further details on design and medical implications of the study, one can refer to Diggle, *et al.* (2002).

Modelling of the CD4+ data is really challenging as the data are highly unbalanced. In the literature, various estimation methods were proposed to model the mean and covariance structures of the data, see Diggle and Verbyla (1998), Ye and Pan (2006) and references therein. Below we reanalyze the CD4+ data using our proposed local maximum likelihood estimation methods. The resulting nonparametric estimators of the mean, generalized autoregressive parameters, and log-innovation variances are presented in Figure 2. Here we assume a univariate smoothing function $g(t^*_{ijk})$ where $t^*_{ijk} \equiv t_{ij} - t_{ik}$ when modelling the generalized autoregressive parameters ϕ_{ijk}. In other words, the within-subject correlation is assumed to be stationary.

Figure 2 is about here.

The estimated bandwidth parameters obtained by the MCVS method are $h_1 = 0.68$, $h_2 = 0.02$, and $h_3 = 0.17$. Figure 2 shows that the smoothing function for modelling the generalized autoregressive parameters displays a cubic polynomial pattern, confirming the discovery made by Ye and Pan (2006). The resulting mean trajectory is similar to the one obtained by Diggle *et al.* (2002) who used smoothing spline to model the mean by assuming certain covariance structures. From Figure 2 we can also see that the mean smoothing function is roughly a constant between around 1000 cells and the time of seroconversion and then decreases afterwards, indicating that the CD4+ cell loss may be more rapid after seroconversion. For the log-innovation variances, it seems that the fitted smooth curve is a higher order polynomial, in contrast to the cubic polynomial suggested by Ye and Pan (2006). But if we ignore some small wiggly part the cubic polynomial curve may be reasonable for modelling the log-innovation variances. Note that when modelling the mean and log-innovation variances, the curve patterns near the end of time may not be reliable as there are fewer observations there for highly unbalanced longitudinal data.

5.3 Monte Carlo simulation

In this section a simulation study is made to assess the performance of the proposed approach. For each subject we assume that there are five repeated measurements and the corresponding time points at which the observations are made are randomly chosen from the uniform distribution over the interval $[-2, 2]$. Denote $t_i = (t_{i1}, \ldots, t_{i5})$ as the time points and $t_i^* = \{t_{ij} - t_{ik} : j = 2, \ldots, 5; k = 1, j-1\}$ as the lag in time $(i = 1, \ldots, n)$ for the ith subject. We choose three underlying target function curves for the mean, generalized autoregressive parameters and log-innovation variances as follow.

$$
\begin{aligned}
f(t_i) &= 7 \times \left[\exp\left\{-(t_i+1)^2\right\} + \exp\left\{-(t_i-1)^2\right\}\right] - 5.5 && (5.1)\\
g(t_i^*) &= 0.9 \times \sin\left(\frac{\pi}{8}(4 - t_i^*)\right) && (5.2)\\
q(t_i) &= 2 - \frac{1}{3} \times t_i^2 && (5.3)
\end{aligned}
$$

Based on (5.1)-(5.3) we then use $f(t_i)$ to form the mean vectors μ_i and use $g(t_i^*)$ and $q(t_i)$ to construct the covariance matrices Σ_i. We generate random vectors $y_i \sim \mathcal{N}_5(\mu_i, \Sigma_i)$ $(i = 1, \ldots, n)$. We choose three sample sizes $n = 250, 500$ and 1000 in our simulation studies to monitor the effects of the sample size. For each subject the generated data have five correlated observations, but the corresponding observation times may be very irregular so that the generated longitudinal data are unbalanced.

For each sample size above, we perform 100 simulations for the target function curves and calculate the average of estimated curves in the simulations, which are displayed in Figure 3.

Figure 3 is about here.

Here the MCVS criterion is used to find the optimal value of bandwidth parameter for each simulated data set. From Figure 3 we see that, on average, the local kernel weighted likelihood estimation for modelling of the mean and covariance structures performs very well, though for log-innovation variances there is a small discrepancy between the target curve and simulated curve when the sample size is 500.

An argument may be that, the average simulation curves may not be always good enough to reflect the efficacy of the methodology. As an alternative we give some typical simulation replicates, for example, the most wiggly, the median and the smoothest ones. Figure 4 provides such information when the sample size is $n = 250$, where the solid curve is the truth, and the dotted curve represents the most wiggly replicate in the 100 simulated data, corresponding to the optimal bandwidths $h_1 = 0.12, h_2 = 0.01$ and $h_3 = 0.10$ for smoothing the mean, the generalized autoregressive parameters and log-innovation variances, respectively. The dash-dotted curve represents the replicate close to the median of the estimated smoothing parameters, that is, $h_1 = 0.22, h_2 = 0.04$ and $h_3 = 0.12$. The dashed curve in contrast is the smoothest replicate in the 100 simulated data, which has the optimal bandwidths $h_1 = 0.42, h_2 = 0.06$ and $h_3 = 0.40$ for the estimated functions $\hat{f}(t)$, $\hat{g}(t^*)$ and $q(t)$, respectively. It is noted that when modelling the log-innovation variances the median replicate increases rapidly near the right boundary of time, see Figure 4. This phenomenon is the well-known boundary problem. In other words, nonparametric smoothing methods especially kernel methods may not perform very well in the boundary area due to the sparse data problem.

Figure 4 is about here.

6 Discussion

More and more evidence show that statistical inference for longitudinal data may not be efficient when misspecification of the within-subject covariance structures occurs. In some circumstances parameter estimators in the mean could be very biased when covariance structures are misspecified. Therefore statistical modelling for covariance structures has attracted more attention in the past decade. In this paper we propose a nonparametric local kernel weighted likelihood based approach to model the mean and covariance structures, simultaneously. The modelling approach involves the use of the modified Cholesky decomposition, which decomposes the within-subject covariances into the generalized autoregressive parameters and the innovation variances. These reparameterized covariance parameters, together with the mean, are in turn parsimoniously modelled by three nonparametric smooth functions of time/lag. Local kernel-weighted likelihood estimation methods are then used to estimate the nonparametric smooth functions. Numerical results show that the proposed approach performs very well.

We propose to use a modified cross-validation with leave-one-subject-out criterion to find the optimal value of the bandwidth parameter. Numerical results confirm the efficacy of this bandwidth selector. In practice, however, it might be more attractive if we have a bandwidth selection criterion that can adaptively find the location-varying bandwidth parameter estimator. This deserves a further investigation. Furthermore, it is very common that certain baseline covariates, along with the time, might be of interest, so the models in (2.2) could include an additional parametric part. This actually forms semiparametric models for modelling the mean, generalized autoregressive parameters and log-innovation variances. Within this framework, the assumption of homogeneous covariances across subjects become testable (Pan and MacKenzie, 2003). More sophisticated models including varying-coefficients models and functional models are under investigation in our studies.

Acknowledgements

The work was supported by grants from the Engineering and Physical Sciences Research Council and an Overseas Research Scholarship of the U.K.

Appendix A. *Proof of Propositions 3.1 and 3.2*

Proof of Proposition 3.1: Consider the first-order derivative of $Q_2(\gamma)$ with respect to γ

$$\begin{aligned}\frac{\partial Q_2(\gamma)}{\partial \gamma} &= -2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ij} - \sum_{k=1}^{j-1} Z'_{ijk}\gamma r_{ik}K_{h_2,ijk}^{1/2}\right)\left(\sum_{k=1}^{j-1} r_{ik}K_{h_2,ijk}^{1/2}Z_{ijk}\right)/\sigma_{ij}^2 \\ &= -2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(r_{ij} - Z'_{(i,j)}\gamma\right)Z_{(i,j)}/\sigma_{ij}^2 \\ &= -2\sum_{i=1}^{n} Z'_{(i)}D_i^{-1}\left(r_i - Z_{(i)}\gamma\right)\end{aligned}$$

where $K_{h_2,ijk} = K_{h_2}(t_{ij} - t_0, t_{ik} - s_0)$. Therefore, the equation $\partial Q_2(\gamma)/\partial\gamma = 0$ has the solution (3.10), which is obviously the minimizer of the partial local-weighted log-likelihood $Q_2(\gamma)$. The proof is complete.

Proof of Proposition 3.2: First, the first-order derivative of $Q_3(\lambda)$ is equal to

$$\begin{aligned}\frac{\partial Q_3(\lambda)}{\partial \lambda} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1 - (r_{ij} - \hat{r}_{ij})^2 \exp\left\{-z'_{ij}\lambda\right\}\right) z_{ij}K_{h_3}(t_{ij} - t_0) \\ &= \sum_{i=1}^{n} Z'_iW_i^{-1}(V_i - I_{m_i})1_{m_i}\end{aligned}$$

where the second equality is rewritten in matrix notation. Then, the second-derivative of $Q_3(\lambda)$ can be written as

$$\begin{aligned}\frac{\partial^2 Q_3(\lambda)}{\partial\lambda\partial\lambda'} &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}(r_{ij} - \hat{r}_{ij})^2 \exp\{-z'_{ij}\lambda\}z_{ij}z'_{ij}K_{h_3}(t_{ij} - t_0) \\ &= \sum_{i=1}^{n} Z'_iW_i^{-1}Z_i\end{aligned}$$

Therefore, given a starting value of λ the Newton-Raphson one-step iteration for the solution at t_0 is obtained by

$$\begin{aligned}\hat{\lambda}(t_0) &= \lambda + \left(-\sum_{i=1}^{n} Z_i' W_i^{-1} Z_i\right)^{-1} \left(\sum_{i=1}^{n} Z_i' W_i^{-1} (V_i - I_{m_i}) 1_{m_i}\right) \\ &= \left(\sum_{i=1}^{n} Z_i' W_i^{-1} Z_i\right)^{-1} \left(\sum_{i=1}^{n} Z_i' W_i^{-1} \{Z_i \lambda - (V_i - I_{m_i}) 1_{m_i}\}\right) \\ &= \left(\sum_{i=1}^{n} Z_i' W_i^{-1} Z_i\right)^{-1} \left(\sum_{i=1}^{n} Z_i' W_i^{-1} (\log D_i + I_{m_i} - V_i) 1_{m_i}\right) \\ &= \left(\sum_{i=1}^{n} Z_i' W_i^{-1} Z_i\right)^{-1} \left(\sum_{i=1}^{n} Z_i' W_i^{-1} u_i\right)\end{aligned}$$

and the proof is complete.

Appendix B. *Proof of Theorem 1*

Recall that the local weighted least square estimator of the mean parameters is

$$\hat{\beta}(t) = \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) X_i(t)\right)^{-1} \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) y_i\right) \tag{B.1}$$

where $V_i^{-1} = K_{ih_1}^{1/2}(t)(I_i \Sigma_i I_i)^{-1} K_{ih_1}^{1/2}(t)$. We denote $(I_i \Sigma_i I_i)^{-1} = (\sigma_i^{jk})_{j,k=1,\ldots,m_i}$. Note that $(I_i \Sigma_i I_i)^{-1}$ denotes the Moore-Penrose generalized inverse of the matrix $I_i \Sigma_i I_i$. We then conclude that $|\sigma_i^{jk}| < \infty$ under regularity condition (ii) . Without loss of generality, we only provide the proof for the local linear regression estimation. In other words, $x_{ij}(t) = (1, (t_{ij} - t))'$ and $\beta(t) = (\beta^0(t), \beta^1(t))'$ and so $\hat{f}(t) = e_1' \hat{\beta}(t)$ with $e_1 = (1, 0)'$.

Note that $E\hat{f}(t) = e_1' \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) X_i(t)\right)^{-1} \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) f(t_i)\right)$ where $f(t_i) = (f(t_{i1}), ..., f(t_{im_i}))'$ with

$$f(t_{ij}) = f(t) + (t_{ij} - t) f'(t) + (t_{ij} - t)^2 \frac{1}{2} f''(t) + R_{ij}(t) = x_{ij}'(t) \beta(t) + Q_{ij}(t) + R_{ij}(t),$$

$Q_{ij}(t) = (t_{ij} - t)^2 f''(t)/2$, and $R_{ij}(t)$ is the remainder. The bias of the mean estimator can thus be approximated by

$$E[\hat{f}(t) | t_1, \ldots, t_n] - f(t) \doteq e_1' \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) X_i(t)\right)^{-1} \left(\sum_{i=1}^{n} X_i'(t) V_i^{-1}(t) Q_i(t)\right) \tag{B.2}$$

where $t_i = (t_{i1}, ..., t_{im_i})'$ and $Q_i(t) = (Q_{i1}(t), ..., Q_{im_i}(t))'$ for $i = 1, 2, ..., n$.

Based on (B.2) we first consider the asymptotic behavior of $V_i^{-1}(t)$. From its definition we know that the (j,k)th element of $V_i^{-1}(t)$ is $K_{h_1}^{1/2}(t_{ij}-t)\sigma_i^{jk}K_{h_1}^{1/2}(t_{ik}-t)$ $(j,k=1,\ldots,m_i)$. When the bandwidth h_1 is sufficiently small, $K_{h_1}^{1/2}(t_{ij}-t)\sigma_i^{jk}K_{h_1}^{1/2}(t_{ik}-t)$ is almost negligible where $j \neq k$ and $|\sigma_i^{jk}| < \infty$. Therefore, $V_i^{-1}(t)$ can be approximately viewed as diagonal matrices, that is, $V_i^{-1}(t) \doteq \mathrm{diag}\{K_{h_1}(t_{i1}-t)\sigma_i^{11}, \ldots, K_{h_1}(t_{im_i}-t)\sigma_i^{m_im_i}\}$, as long as the bandwidth h_1 is small enough.

We continue by considering the asymptotic behavior of $S_n = n^{-1}\sum_{i=1}^n X_i'(t)V_i^{-1}(t)X_i(t)$. When h_1 is small enough, S_n can be approximated by

$$\begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{pmatrix} \quad \text{where } S_{n,l} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sigma_i^{jj}K_{h_1}(t_{ij}-t)(t_{ij}-t)^l \text{ for } l=0,1,2.$$

Below we apply the counting process $N_i(t)$ in (4.1) to the estimation procedure. Obviously, the response process $y_i(t)$ is observed at the jump points of $N_i(t)$. Let $\varphi(t)$ be the intensity function of the counting process $N(t)$ and denote $\tau(t_{ij}) = \sigma_i^{jj}$. Then we have

$$S_{n,l} = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\tau(t_{ij})K_{h_1}(t_{ij}-t)(t_{ij}-t)^l = n^{-1}\sum_{i=1}^{n}\int_0^\infty \tau(s)K_{h_1}(s-t)(s-t)^l dN_i(s)$$

for $l = 0, 1, 2$. Under the regularity conditions, we obtain $S_{n,l} \to E(S_{n,l})$ as $n \to \infty$. On the other hand, the expectation $E(S_{n,l})$ is give by

$$\begin{aligned}
E(S_{n,l}) &= n^{-1}\sum_{i=1}^{n}E\int_0^\infty \tau(s)\frac{K((s-t)/h_1)}{h_1}(s-t)^l dN_i(s) \\
&= \int_0^\infty \tau(s)\frac{K((s-t)/h_1)}{h_1}(s-t)^l\varphi(s)ds \\
&= \int_{-t/h_1}^\infty \tau(t+uh_1)K(u)u^l h_1^l\varphi(t+uh_1)du \\
&\doteq h_1^l\int_{-\infty}^{+\infty}(\tau(t)\varphi(t)+o_p(1))K(u)u^l du \\
&= h_1^l\left\{\tau(t)\varphi(t)\int_{-\infty}^{+\infty}u^lK(u)du + o_p(1)\int_{-\infty}^{+\infty}u^lK(u)du\right\} \\
&= h_1^l\tau(t)\varphi(t)\int_{-\infty}^{+\infty}u^lK(u)du\,(1+o_p(1))
\end{aligned}$$

Using the condition (iv) we then have the asymptotic behavior of S_n being

$$S_n = \begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2} \end{pmatrix} = \begin{pmatrix} \tau(t)\varphi(t) + o_p(1) & 0 \\ 0 & h_1^2\tau(t)\varphi(t)c_K + o_p(h_1^2) \end{pmatrix}.$$

In a similar manner we can show that $n^{-1}\left(\sum_{i=1}^n X_i'(t)V_i^{-1}(t)Q_i(t)\right)$ satisfies

$$\begin{pmatrix} n^{-1}\displaystyle\sum_{i=1}^n\sum_{j=1}^{m_i}\tau(t_{ij})K_{h_1}(t_{ij}-t)(f^{''}(t)/2)(t_{ij}-t)^2 \\ \\ n^{-1}\displaystyle\sum_{i=1}^n\sum_{j=1}^{m_i}\tau(t_{ij})K_{h_1}(t_{ij}-t)(f^{''}(t)/2)(t_{ij}-t)^3 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}h_1^2\tau(t)\varphi(t)f^{''}(t)c_K + o_p(h_1^2) \\ \\ 0 \end{pmatrix}.$$

Based on (B.2), the asymptotic behavior of the bias of $\hat{f}(t)$ is given by

$$\text{Bias}(\hat{f}(t)|t_1,\ldots,t_n) = \{E[\hat{f}(t)|t_1,\ldots,t_n] - f(t)\} = \frac{1}{2}h_1^2 f^{''}(t)c_K + o_p(h_1^2) \qquad (B.3)$$

as long as n is large enough. Finally, we consider the asymptotic behavior of the variances of $\hat{f}(t)$. Obviously,

$$\text{Var}(\hat{f}(t)|t_1,\ldots,t_n) = \frac{1}{n}e_1'S_n^{-1}\left(n^{-1}\sum_{i=1}^n X_i'(t)V_i^{-1}(t)\Sigma_i V_i^{-1}(t)X_i(t)\right)S_n^{-1}e_1. \qquad (B.4)$$

Below we study the asymptotic behavior of $S_n^* = n^{-1}\sum_{i=1}^n X_i'(t)V_i^{-1}(t)\Sigma_i V_i^{-1}(t)X_i(t)$. Note that the (j,k)th off-diagonal elements of $V_i^{-1}(t)\Sigma_i V_i^{-1}(t)$ must be of the form

$$\sum_{s'=1}^{m_i}\sum_{s=1}^{m_i} K_{h_1}^{1/2}(t_{ij}-t)\sigma_i^{js}K_{h_1}^{1/2}(t_{is}-t)\text{Cov}(y_i(t_{is}),y_i(t_{is'}))K_{h_1}^{1/2}(t_{is'}-t)\sigma_i^{s'k}K_{h_1}^{1/2}(t_{ik}-t),$$

which is negligible as $j \neq k$ and h_1 is sufficiently small. In other words, $V_i^{-1}(t)\Sigma_i V_i^{-1}(t)$ can be approximately viewed as diagonal matrices when the bandwidth h_1 is small enough. Therefore, the elements of S_n^* can be approximated by

$$S_{n,l}^* = n^{-1}\sum_{i=1}^n\sum_{j=1}^{n_i}(t_{ij}-t)^l K_{h_1}^2(t_{ij}-t)(\sigma_i^{jj})^2\text{Var}(y_i(t_{ij})) \quad (\text{for } l = 0,1,2)$$

Similar to $S_{n,l}$ above, we can show that

$$S_{n,l}^* = h_1^{l-1}\tau^2(t)\varphi(t)\text{Var}(\epsilon(t))\int_{-\infty}^{+\infty}u^l K^2(u)du\,(1+o_p(1)). \quad (l = 0,1,2)$$

Based on (B.4) we have

$$\mathrm{Var}(\hat{f}(t)|t_1,\ldots,t_n) = (nh_1)^{-1}(\varphi(t))^{-1}\mathrm{Var}(\epsilon(t))d_K + o_p(\frac{1}{nh_1}). \qquad (B.5)$$

Combining this with (B.3) we obtain the local conditional MSE of the mean smooth estimator $\hat{f}(t)$ in (4.3). Furthermore, it follows that under the mild regular conditions we have $E[(\hat{f}(t) - f(t))^2|t_1,\ldots,t_n] \to 0$ as $h_1 \to 0$ and $nh_1 \to \infty$, implying $\hat{f}(t) \xrightarrow{P} f(t)$ as $n \to \infty$. The proof is complete.

Appendix C. *Proof of Theorem 2*

The proof of Theorem 1 requires an analytical form of the smooth estimator. This similar strategy, thus, can be used to show the consistency of the smooth estimator for the generalized autoregressive parameters, as it has a closed form. For the log-innovation variances, however, the smooth estimator does not have a simple analytical form and so the technical strategy shown above cannot be directly used. In a spirit of Chiu *et al.* (1996) we propose an alternative method to show the consistency of the smooth estimator for the log-innovation variances. The conclusion made here is actually $\hat{q}(t) \xrightarrow{a.s.} q(t)$ as $n \to \infty$, that is, $\hat{q}(t)$ converges almost surely to $q(t)$, implying $\hat{q}(t) \xrightarrow{P} q(t)$ as $n \to \infty$.

First, the local kernel-weighted log-likelihood contributed by the ith subject at the time t can be written as

$$-2\ell\left(y_i; Z_i'\lambda(t)\right) = \sum_{j=1}^{m_i}\left(z_{ij}'\lambda(t) + (r_{ij}-\hat{r}_{ij})^2\exp\left\{-z_{ij}'\lambda(t)\right\}\right)K_{h_3}(t_{ij}-t) \qquad (C.1)$$

where $z_{ij} = (1,(t_{ij}-t))'$ and $\lambda(t) = (\lambda^0(t),\lambda^1(t))'$. We assume $\lambda_0(t) = (\lambda_0^0(t),\lambda_0^1(t))'$ is the true value of $\lambda(t)$. Note that $q(t) \equiv \lambda^0(t) = e_1'\lambda(t)$ and the true value is $q_0(t) \equiv \lambda_0^0(t) = e_1'\lambda_0(t)$. Since $E(r_{ij}-\hat{r}_{ij})^2 = \sigma_{ij}^2 = \exp\{\log\sigma_{ij}^2\} = \exp\{q(t_{ij})\}$, the expectation of

$-2\log \ell(y_i; Z_i\lambda(t))$ at $\lambda(t) = \lambda_0(t)$ is give by

$$\begin{aligned}
& E_0\left[-2\log \ell\left(y_i; Z_i'\lambda(t)\right)\right] \\
= & \sum_{j=1}^{m_i}\left[z_{ij}'\lambda(t) + \exp\left\{q_0(t_{ij})\right\}\cdot \exp\left\{-z_{ij}'\lambda(t)\right\}\right] K_{h_3}(t_{ij}-t) \\
\doteq & \sum_{j=1}^{m_i}\left[z_{ij}'\lambda(t) + \exp\left\{z_{ij}'(\lambda_0(t)-\lambda(t)) + \frac{(t_{ij}-t)^2}{2}q_0^{''}(t)\right\}\right] K_{h_3}(t_{ij}-t)
\end{aligned}$$

where $q_0(t_{ij}) \doteq q_0(t) + q_0'(t)(t_{ij}-t) + (1/2)q_0^{''}(t)(t_{ij}-t)^2$ is applied. Furthermore, noting that $z_{ij}'\lambda(t) = \lambda^0(t) + (t_{ij}-t)\,\lambda^1(t)$ and using the principle of a counting process we have

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n} E_0\left[-2\log \ell\left(y_i; Z_i\lambda(t)\right)\right] \\
\doteq\ & n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Big[\lambda^0(t) + (t_{ij}-t)\,\lambda^1(t) \\
& + \exp\Big\{\left(\lambda_0^0(t)-\lambda^0(t)\right) + (t_{ij}-t)(\lambda_0^1(t)-\lambda^1(t)) + \frac{1}{2}(t_{ij}-t)^2 q_0^{''}(t)\Big\}\Big] K_{h_3}(t_{ij}-t) \\
=\ & n^{-1}\sum_{i=1}^{n}\int_0^{\infty}\Big[\lambda^0(t) + (s-t)\lambda^1(t) \\
& + \exp\Big\{\left(\lambda_0^0(t)-\lambda^0(t)\right) + (s-t)(\lambda_0^1(t)-\lambda^1(t)) + \frac{1}{2}(s-t)^2 q_0^{''}(t)\Big\}\Big] K_{h_3}(s-t)dN_i(s)
\end{aligned}$$

Let $u = (s-t)/h_3$ and denote $G(uh_3) = uh_3(\lambda_0^1(t)-\lambda^1(t)) + \frac{1}{2}u^2h_3^2q_0^{''}(t)$. The last equality above can be further expressed as

$$\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\int_{-t/h_3}^{+\infty}\Big[\lambda^0(t) + h_3u\lambda^1(t) + \exp\Big\{\left(\lambda_0^0(t)-\lambda^0(t)\right) + G(uh_3)\Big\}\Big]\frac{K(u)}{h_3}dN_i(t+uh_3) \\
= & \int_{-\infty}^{+\infty}\Big[q(t) + h_3u\lambda^1(t) + \exp\Big\{\left(q_0(t)-q(t)\right) + G(uh_3)\Big\}\Big] K(u)(\varphi(t)+o_p(1))du
\end{aligned}$$

as long as n is large enough and h_3 is sufficiently small. Applying Taylor expansion to the function $\exp\{G(uh_3)\}$ and noting $G(uh_3) = uh_3\{(\lambda_0^1(t)-\lambda^1(t)) + uh_3q_0''(t)/2\}$, we can show that, under the regularity conditions,

$$\int_{-\infty}^{+\infty}\exp\{G(uh_3)\}K(u)du \to 1 \ \text{ as } \ h_3 \to 0.$$

Therefore, we have showed that

$$C_0(q(t)) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E_0\left[-2\log \ell\left(y_i; Z_i'\lambda(t)\right)\right] = \left(q(t) + \exp\left\{q_0(t)-q(t)\right\}\right)\varphi(t) \qquad (C.2)$$

as long as the bandwidth h_3 is small enough. Clearly, the finite limit function $C_0(q(t))$ achieves its minimum at $q(t) = q_0(t)$, that is, $C_0(q(t)) > C_0(q_0(t))$ for any $q(t) \neq q_0(t)$.

We then consider $V_0\left[-2\log\ell(y_i; Z_i'\lambda(t))\right]$, the variance of $-2\log\ell\left(y_i; Z_i'\lambda(t)\right)$ at $\lambda(t) = \lambda_0(t)$. Note that $\mathrm{Var}\{(r_{ij} - \hat{r}_{ij})^2\} = 2\sigma_{ij}^4 = 2\exp\{2q(t_{ij})\}$. Based on (C.1) and using the regularity condition assumptions, it is easy to show that there exists a constant κ_0 independent of i such that $V_0\left[-2\log\ell(y_i; Z_i'\lambda(t))\right] < \kappa_0$. Therefore, we have

$$\sum_{i=1}^{\infty} \frac{V_0\left[-2\log\ell(y_i; Z_i'\lambda(t))\right]}{i^2} < +\infty$$

By Kolmogorov's strong law of large numbers we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left[-2\log\ell(y_i; Z_i'\lambda(t))\right] - \frac{1}{n}\sum_{i=1}^{n} E_0\left[-2\log\ell(y_i; Z_i'\lambda(t))\right] \overset{a.s}{\to} 0 \qquad (C.3)$$

as $n \to \infty$. Combining (C.2) and (C.3), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\left[-2\log\ell(y_i; Z_i'\lambda(t))\right] \overset{a.s}{\to} C_0(q(t)) \quad (\text{as } n \to \infty) \qquad (C.4)$$

We are now ready to show $\hat{q}_n(t) \overset{a.s}{\to} q_0(t)$ $(n \to \infty)$. In fact, if it does not hold there must exist a subsequence $\{m\} \subset \{n\}$ such that $\hat{q}_m(t) \overset{a.s}{\to} \tilde{q}(t) \neq q_0(t)$ as $m \to \infty$. Let $\hat{\lambda}_m(t) = (\hat{q}_m(t), \hat{q}_m'(t))'$ and $\tilde{\lambda}(t) = (\tilde{q}(t), \tilde{q}'(t))'$. According to the definition of $\hat{\lambda}_m(t)$ we know

$$\frac{1}{m}\sum_{i=1}^{m}\left[-2\log\ell(y_i; Z_i'\hat{\lambda}_m(t))\right] \leq \frac{1}{m}\sum_{i=1}^{m}\left[-2\log\ell(y_i; Z_i'\lambda_0(t))\right] \qquad (C.5)$$

where $\lambda_0(t)$ is the true value of $\lambda(t)$ at t. Based on (C.4) and (C.5), we obtain

$$C_0(\tilde{q}(t)) \leq C_0(q_0(t)) \qquad (C.6)$$

due to the fact that the convergence is in uniform. It is clear that (C.6) contradicts the conclusion drawn from (C.2). Therefore we have $\hat{q}_n(t) \overset{a.s.}{\to} q_0(t)$ as $n \to \infty$, implying $\hat{q}_n(t) \overset{P}{\to} q_0(t)$ as $n \to \infty$. Similarly, it can be shown that $\hat{g}_n(t^*) \overset{P}{\to} g_0(t^*)$ as $n \to \infty$ but the details are omitted here. The proof is complete.

References

Chen, K. and Jin, Z. (2005) Local polynomial regression analysis of clustered data. *Biometrika*, **92**, 59-74.

Chiu, T.Y.M. Leonard, T. and Tsui, K.W. (1996). The matrix-logarithm covariance model. *J. Am. Stat. Assoc.*, **91**, 198-210.

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

Diggle, P.J. and Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, **54**, 401-415.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* Chapman and Hall, London.

Fan, J.Q. Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *J.R.Statist.Soc. B*, **60**, 591-608.

Fan, J. and Zhang, J.T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc. B.*, **62**, 303-322.

Hart, J.D. and Wehrly, T.E. (1993). Consistency of Cross-Validation when the data are curves. *Stoch. Process. Applic.*, **45**, 351-361.

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

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl. Statist.* **36**, 296-308.

Lin, X. H. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Ass.*, **95**, 520-534.

Lin, X. H., Wang, N., Welsh, A. and Carroll, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered data. *Biometrika*, **91**, 177-193.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussions). *J. Am. Statist. Ass.*, **96**, 103-126.

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

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.

Rice, J. A. and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J.R.Statist.Soc. B*, **53**, 233-243.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Stat.*, **22**, 1346-1370.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, **90** 831-844.

Zimmerman, D.L. and Núñez-Antón (2001). Parametric modelling of growth curve data :An overview. *Sociedad de Estadística e Investigación Operativa Test*, Vol **10**, 1-73.

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

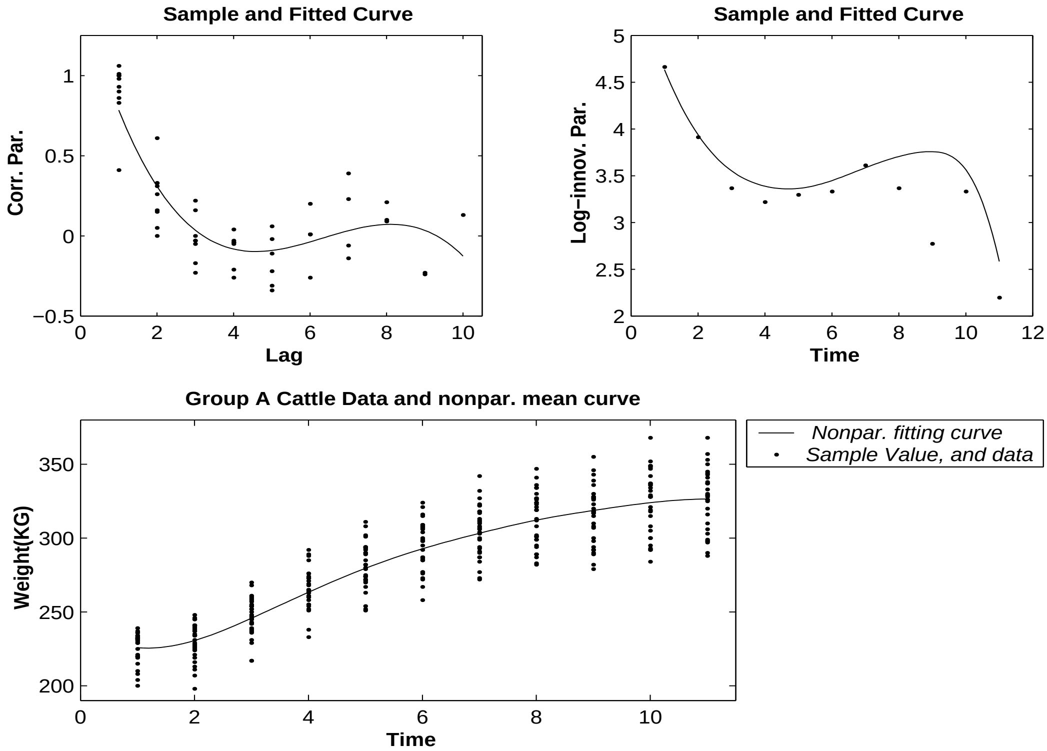

Figure 1: Nonparametric smooth curves for modelling the mean, generalized autoregressive parameters and log-innovation variances for Cattle data in the treatment A, using the local kernel weighted likelihood estimation method.

Table 1: Cattle data in the treatment A. The estimated correlation matrix by Wu and Pourahmadi (2003) are given above the main diagonal. The fitted correlations, below the main diagonal, are obtained by smoothing the lower triangular entries of $\hat{T}$ and the diagonals of $\hat{D}$ using the local polynomial method.

Time point	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.81	0.75	0.70	0.66	0.63	0.60	0.58	0.57	0.55	0.55
2	0.74	1.00	0.91	0.85	0.81	0.77	0.73	0.71	0.70	0.67	0.66
3	0.76	0.86	1.00	0.92	0.88	0.84	0.79	0.77	0.76	0.73	0.72
4	0.72	0.85	0.91	1.00	0.93	0.89	0.84	0.82	0.81	0.77	0.77
5	0.67	0.81	0.89	0.93	1.00	0.93	0.88	0.86	0.85	0.81	0.81
6	0.62	0.76	0.84	0.90	0.94	1.00	0.92	0.90	0.89	0.85	0.84
7	0.56	0.70	0.78	0.85	0.89	0.93	1.00	0.94	0.93	0.90	0.89
8	0.53	0.65	0.73	0.79	0.84	0.89	0.93	1.00	0.96	0.93	0.92
9	0.51	0.61	0.68	0.74	0.79	0.85	0.90	0.94	1.00	0.95	0.94
10	0.49	0.59	0.65	0.71	0.75	0.81	0.86	0.91	0.94	1.00	0.98
11	0.45	0.57	0.63	0.68	0.73	0.78	0.84	0.90	0.95	0.98	1.00

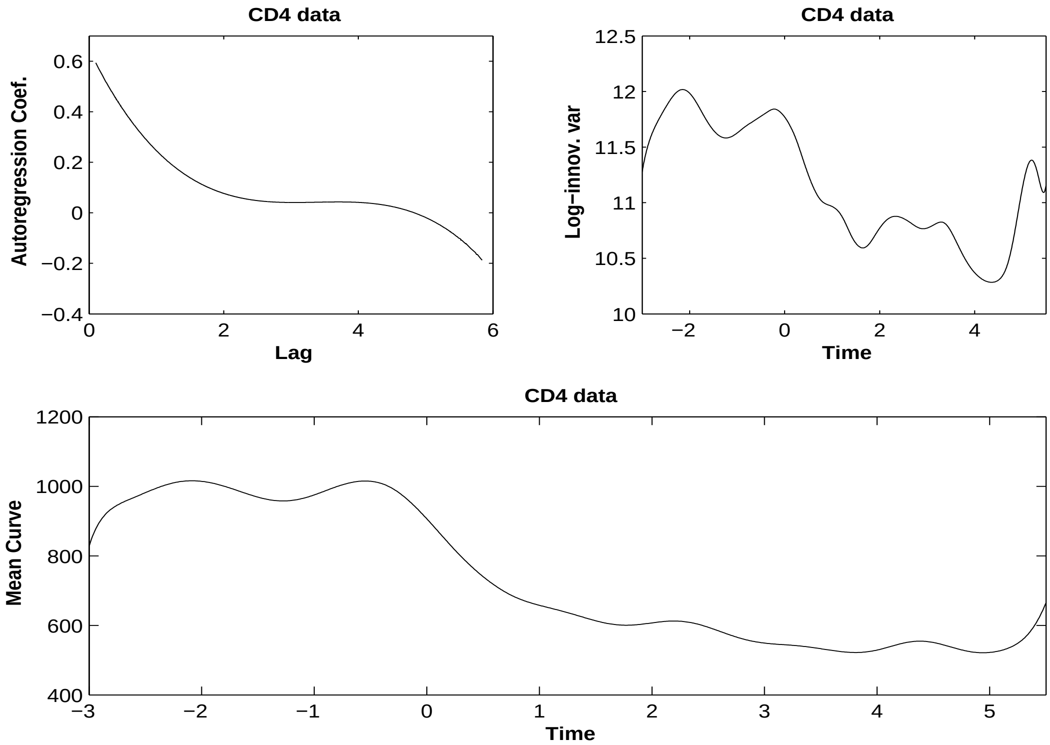

Figure 2: Nonparametric smooth curves for modelling the mean, generalized autoregressive parameters and log-innovation variances for CD4+ cell data, using the local kernel weighted likelihood estimation method.

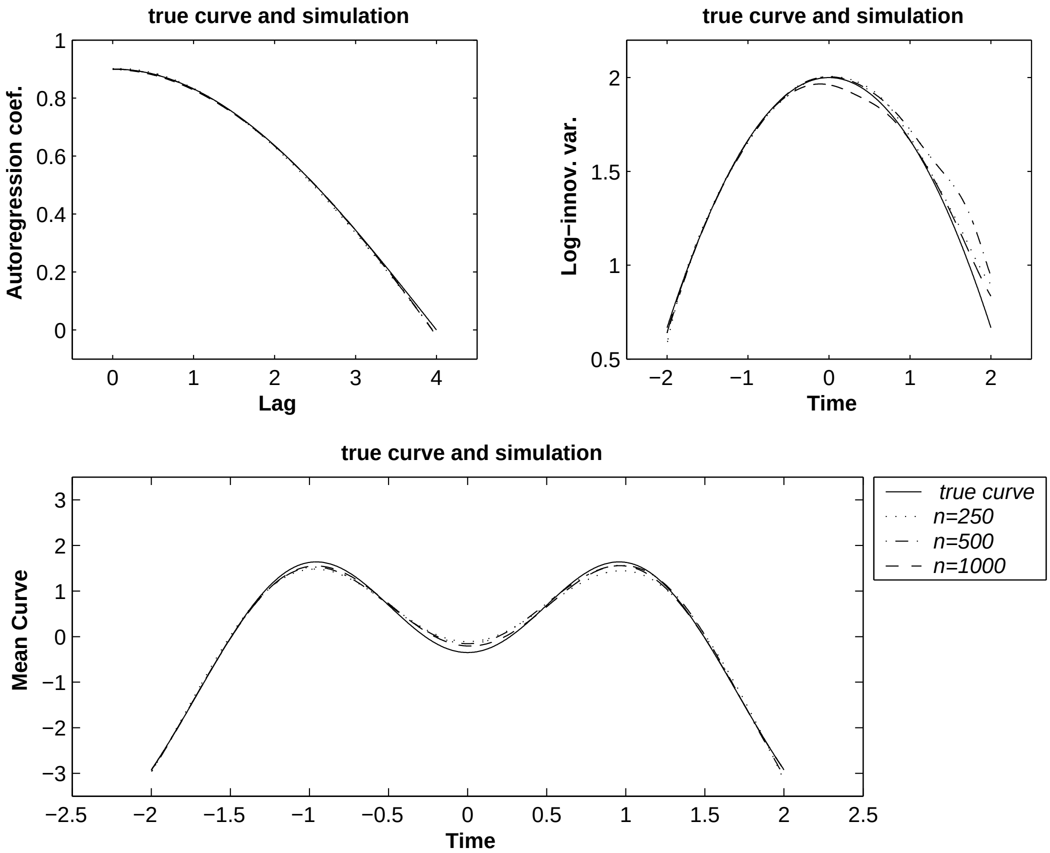

Figure 3: Local kernel weighted likelihood based estimators of the mean, generalized autoregressive parameters and log-innovation variances for simulated data with three sample sizes($\cdots\cdots$, $n = 250$; $\cdot - \cdot-$, $n = 500$; - - - -, $n = 1000$). The solid curves represents the true curves.

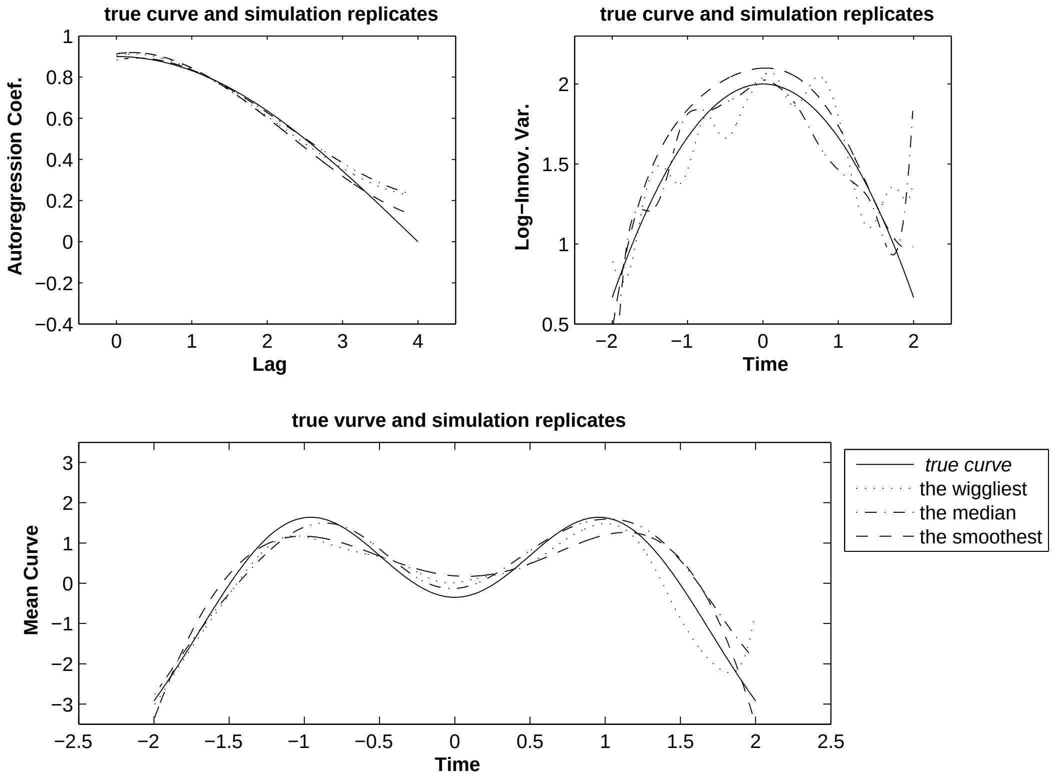

Figure 4: The smoothest, the most wiggly replicates and the replicate close to the median of the smoothing parameter estimators for simulated data with sample size $n = 250$. The solid curves represents the true curves.