Article ID: 0253-2778(2013)08-0607-15

Joint semiparametric mean-covariance modeling by moving average Cholesky decomposition for longitudinal data

XING Xin, LIU Meimei, ZHANG Weiping

(*Dept. of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Modeling the mean and covariance simultaneously has recently received considerable attention when efficiently analyzing the longitudinal data. An unconstrained and statistically interpretable reparameterization of covariance matrix itself was presented by utilizing a novel Cholesky factor. The entries in such decomposition have moving average and log innovation interpretation and can thus be modeled as functions of covariates. With this decomposition and the consideration of model flexibility, new semiparametric models for jointly modeling the mean and covariance itself were proposed, rather than its inverse as commonly studied in literature. A spline based approach using generalized estimating equations was developed to estimate the parameters in the mean and the covariance. It was shown that the estimators for the parametric parts in both the mean and covariance are consistent and asymptotically normally distributed, and the nonparametric parts could be estimated at an optimal rate of convergence. Simulation studies and real data analysis illustrate that the proposed approach could yield highly reliable estimation of the mean and covariance matrix.

Key words: longitudinal data; semiparametric model; generalized estimating equation; modified Cholesky decomposition; moving average; B-spline

CLC number: O212.7 **Document code**: A doi: 10.3969/j.issn.0253-2778.2013.08.002

AMS Subject Classification (2000): 62J12; 62G20; 62F12

Citation: Xing Xin, Liu Meimei, Zhang Weiping. Joint semiparametric mean-covariance modeling by moving average Cholesky decomposition for longitudinal data[J]. Journal of University of Science and Technology of China, 2013,43(8):607-621.

纵向数据分析中使用滑动平均 Cholesky 分解对回归均值和协方差矩阵进行同时半参数建模

邢 昕,刘梅梅,张伟平

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:近年来,对纵向数据分析中回归均值和协方差矩阵同时进行建模研究得到越来越多的关注.为满足协

Received: 2012-12-13; **Revised**: 2013-05-17

Foundation item: Supported by the National Natural Science Foundation of China (11271347, 11171321).

Biography: XING Xin, male, born in 1987, master. Research field: Large-sample theory. E-mail: xingxin@mail.ustc.edu.cn

Corresponding author: ZHANG Weiping, PhD/associate Prof. E-mail: zwp@ustc.edu.cn

方差矩阵的正定性约束,文献中常考虑对其逆矩阵进行某种分解. 本文使用一种 Cholesky 分解方法对协方差矩阵本身进行分解,得到的参数没有取值限制且有着明确的统计意义. 具体地,分解后的参数可以视为滑动平均序列的系数和对数更新方差,且在整个实轴上取值无限制. 考虑到模型的稳健性和推断的有效性,提出了一种对回归均值和协方差矩阵同时进行半参数建模的方法,并利用广义估计方程和 B 样条给出了半参数模型的估计方法,得到了参数部分估计的渐近正态性以及非参数部分估计的最优收敛速度. 最后通过模拟和实例分析对所提方法进行了数值研究.

关键词:纵向数据;半参数模型;广义估计方程;修改的 Cholesky 分解;滑动平均;B 样条

0 Introduction

Longitudinal data is characterized by the fact that repeated observations for a subject tend to be correlated. This correlation presents additional opportunities and challenges for analysis. Regression models are widely used and provide general and versatile approaches to analyzing such data[1]. Liang et al.[2] introduced the generalized estimating equations (GEE) for fitting such repeatedly measured data. Although GEE could yield a consistent estimate of the mean parameter by using "working" correlation, such misspecified within-subject correlation may lead to a great loss of efficiency[3]. Therefore, modeling the mean and covariance simultaneously has received considerable interest when efficiently analyzing the longitudinal data. However, modeling the correlation matrix is more challenging than modeling the mean as there are usually more parameters in the correlation matrix and the positive definiteness of the matrix has to be assured. Pourahmadi[4] proposed an unconstrained and statistically interpretable reparameterization of precision matrix by modified Cholesky decomposition. This decomposition is attractive, as the entries in this decomposition can be interpreted as autoregressive parameters and log innovation variances in a time series context. Regression models can then be applied to these entries in a manner similar to the mean models, thus permitting parsimonious characterization of the covariance structure just like the mean. See Refs. [5-8] for further details.

In many applications, the problem of estimating the covariance matrix and precision matrix are usually considered separately, since inversion may be computationally costly, noisy and does not preserve some structural characteristics. Regression methods have been extensively studied by decomposing the precision matrix. But when the covariance matrix itself, rather than the precision matrix, is of interest, the modified Cholesky factor of covariance matrix also has a natural regression interpretation, i. e., the entries in this decomposition have moving average interpretation (see Ref. [9]), and therefore all Cholesky-based regularization methods can be applied to the covariance matrix itself instead of its inverse. Recently, Zhang et al.[10] proposed a BIC based variable selection technique by decomposing the covariance itself rather than the precision matrix. However, it is necessary to relax the parametric and normality assumption proposed in Ref. [10] as model misspecification may result in biased estimation, a problem even more severe than misspecification of the covariance. An attractive alternative is the semiparametric regression model, which provides an excellent trade-off between model interpretability and flexibility. Such a model is also called a partially linear model (PLM), since it relates the response variable with key covariates linearly and with the rest of the covariates nonparametrically. A comprehensive theory about PLM has been well explored (see e. g. Refs. [11-16] and others). For modeling the covariance matrix, Wu et al.[17] proposed nonparametric estimations of the covariance matrix, but their method only focused on balanced measurements, instead of dealing with

irregular observed measurements. Fan et al.[18-19] estimated the marginal variance via kernel smoothing and proposed a parametric model for the correlation matrix. Leng et al.[8] proposed semiparametric models in the mean-covariance modeling for longitudinal data based on Pourahmadi's autoregressive decomposition of precision matrix and regression splines[4].

In this paper, we utilize a new Cholesky factor model to analyze the within-subject variation by decomposing the covariance matrix itself rather than its inverse into a sequence of moving average coefficients and log innovation. Based on this decomposition, we propose semiparametric models for the mean and covariance itself simultaneously, and use the generalized estimating equation technique for parameter estimation. Apart from dealing with irregularly and unbalanced longitudinal data, our semiparametric models perform quite well both theoretically and computationally. The GEE estimators for the linear parts in the mean and covariance models are consistent and asymptotically normally distributed, and the nonparametric parts can also be estimated at the optimal convergence rate by taking advantage of regression splines.

The outline of this paper is as follows. Section 1 introduces the models and estimation methods. Section 2 provides the asymptotic properties of the proposed estimators. Extensive simulations and data analysis are discussed in Section 3. The proofs of the asymptotic results are given in the Appendix.

1　Models and estimation methods

1.1　Models

Denote the n_i repeatedly measured response of the ith subject by $y_i=(y_{i1},\cdots,y_{in_i})'$ and covariate vector as $x_i=(x'_{i1},\cdots,x'_{in_i})'$ respectively, whose components are observed at times $t_i=(t_{i1},\cdots,t_{in_i})'$ for $i=1,\cdots,m$. The total number of observation is $n=\sum_{i=1}^{m}n_i$. In a more general setting, t_{ij} does not have to be time, but can be any time-dependent covariate being modeled nonparametrically. Without loss of generality, we just assume that all the t_{ij} are scaled into the interval $[0,1]$. Furthermore, we assume $E(y_{ij}\mid x_{ij},t_{ij})=\mu_{ij}$ and $\mathrm{Var}(y_i\mid x_i,t_i)=\Sigma_i$, where x_{ij} is p-vector covariate.

To parameterise Σ_i, Pourahmadi[4] first proposed to decompose its inverse as $\Sigma_i^{-1}=T_i'D_i^{-1}T_i$. The lower triangular matrix T_i is unique with 1's on its diagonal and the below-diagonal entries of T_i are the negative autoregressive parameters ϕ_{ijk} satisfying

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

The diagonal entries of D_i are the innovation variances as $\sigma_{ij}^2=\mathrm{Var}(\varepsilon_{ij})$.

By letting $L_i=T_i^{-1}$, a lower triangular matrix with 1's on its diagonal, we can write $\Sigma_i=L_iD_iL_i'$. The entries l_{ijk} in L_i can be interpreted as the moving average coefficients in

$$y_{ij}-\mu_{ij}=\sum_{k=1}^{j-1}l_{ijk}\varepsilon_{ik}+\varepsilon_{ij} \tag{1}$$

where $\varepsilon_{i1}=y_{i1}-\mu_{i1}$ and $E(\varepsilon_i)=0$, $\mathrm{Cov}(\varepsilon_i)=D_i$, for $\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{in_i})'$. Note that the parameters l_{ijk} and $\log(\sigma_{ij}^2)$ are unconstrained and statistically meaningful, thus regularization methods can be applied to the covariance matrix itself instead of its inverse. For regular and balanced data, Rothman et al.[9] provided a banded estimator of the covariance matrix using such Cholesky decomposition in high dimensional cases, but didn't consider building regression models for statistical analysis, and they also commented that their method is not fully efficient.

Since the main difference between our decomposition and that in Ref.[4] is whether to use T_i or its inverse, it is helpful to examine these two decompositions for commonly used covariance matrices. If Σ is a compound symmetry $p\times p$ matrix given by $\sigma^2\{(1-\rho)I+\rho J\}$, where J is a matrix of ones, then the decomposition in Ref.[4] gives $\phi_{jk}=\rho\{1+(j-2)\rho\}^{-1}$, while for our decomposition $l_{jk}=\rho\{1+(k-1)\rho\}^{-1}$. If $\Sigma=$

$\sigma^2(\rho^{|i-j|})_{i,j=1}^p$ is an AR (1) matrix, then Pourahmadi's decomposition gives $\phi_{j,j-1}=\rho$, $\phi_{jk}=0$, $k<j-1$ and ours gives $l_{jk}=\rho^{|j-k|}$.

To avoid model misspecification as well as parsimoniously parameterize the mean-variance structure in terms of covariates, we impose the regression models

$$\left.\begin{aligned} g(\mu_{ij}) &= x_{ij}'\beta + f_0(t_{ij}),\\ l_{ijk} &= w_{ijk}'\gamma,\\ \log(\sigma_{ij}^2) &= z_{ij}'\lambda + f_1(t_{ij}) \end{aligned}\right\} \tag{2}$$

motivated by Refs. [4-6,8]. Here x_{ij}, w_{ijk} and z_{ij} are the $p\times 1$, $q\times 1$ and $d\times 1$ vectors of covariates respectively, and may contain baseline covariates, polynomials in time and their interactions as well. β is regression coefficients in the marginal mean model, γ and λ refer to dependence and variance parameters, and $f_0(\cdot)$ and $f_1(\cdot)$ are unknown smooth functions. The link function $g(\cdot)$ is assumed to be monotone and differentiable, log is the logarithmic function with base e. For convenience, we refer to these three regression models collectively as moving average models, and the regression models in Ref. [4] as autoregressive models.

1.2 Estimating equations

Following Refs. [12,20], we approximate f_0, f_1 by a regression spline, as splines can provide optimal rates of convergence for both the parametric and the nonparametric parts in the semiparametric model with a small number of knots[12,21]. Furthermore, we could utilize any computational algorithm developed for general linear models to fit the semiparametric extension of general linear models, since they treat the 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, and let $0=s_0<s_1<\cdots<s_{k_n}<s_{k_n+1}=1$ be a partition of the interval $[0,1]$. Using the s_i as 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. Let $f_0(t)$, $f_1(t)$ be approximated by $\pi'(t)\alpha$ and $\pi'(t)\tilde{\alpha}$, where $\pi(t)=(B_1(t),\cdots,B_K(t))'$ is the vector of basis functions and α, $\tilde{\alpha}\in\mathbb{R}^K$, note $\pi_{ij}=\pi(t_{ij})$, then the nonlinear regression models in (2) can be linearized as follows:

$$\left.\begin{aligned} g(\mu_{ij}) &= x_{ij}'\beta + \pi'(t_{ij})\alpha := b_{ij}'\theta,\\ \log(\sigma_{ij}^2) &= z_{ij}'\lambda + \pi'(t_{ij})\tilde{\alpha} := h_{ij}'\rho \end{aligned}\right\} \tag{3}$$

where $b_{ij}'=(x_{ij}',\ \pi_{ij}')$, $h_{ij}'=(z_{ij}',\ \pi_{ij}')$, and $\theta=(\beta',\alpha')'$, $\rho=(\lambda',\tilde{\alpha}')'$. Let $\mu_i=(\mu_{i1},\cdots,\mu_{in_i})'$, $g(\mu_i)=(g(\mu_{i1}),\cdots,g(\mu_{in_i}))'$, $B_i=(b_{i1},\cdots,b_{in_i})'=(x_i,\pi_i)$ and define t_i, z_i and H_i in a similar way for $i=1,\cdots,m$. With this notation, using the GEE method from Ref. [2], we can construct the estimating equations for θ,γ and ρ as follows.

$$\left.\begin{aligned} S_1(\theta) &= \sum_{i=1}^m B_i'\Delta_i(B_i\theta)\Sigma_i^{-1}(y_i-\mu_i(B_i\theta)) = 0,\\ S_2(\gamma) &= \sum_{i=1}^m \left(\frac{\partial\varepsilon_i'}{\partial\gamma}\right)D_i^{-1}\varepsilon_i = 0,\\ S_3(\rho) &= \sum_{i=1}^m H_i'D_i(H_i\rho)W_i^{-1}(\varepsilon_i^2-\sigma_i^2(H_i\rho)) = 0 \end{aligned}\right\} \tag{4}$$

where $\Delta_i=\Delta_i(B_i\theta)=\mathrm{diag}\{\dot{g}^{-1}(b_{ij}'\theta),\cdots,\dot{g}^{-1}(b_{in_i}'\theta)\}$ and $\dot{g}^{-1}(\cdot)$ is the derivative of the inverse function $g^{-1}(\cdot)$. When $j=1$ the notation $\sum_{k=1}^{0}$ means zero throughout this paper. Let $\varepsilon_i=(\varepsilon_{i1},\cdots,\varepsilon_{in_i})'$ with $\varepsilon_{ij}=r_{ij}-\sum_{k=1}^{j-1}l_{ijk}\varepsilon_{ik}$, we can see $\partial\varepsilon_i'/\partial\gamma$ is a $q\times n_i$ matrix with the first column zero and the jth ($j>2$) column

$$\partial\varepsilon_{ij}/\partial\gamma = -\sum_{k=1}^{j-1}[\varepsilon_{ik}w_{ijk}+l_{ijk}\partial\varepsilon_{ik}/\partial\gamma],$$

which indicates that ε_{ij} and $\partial\varepsilon_{ij}/\partial\gamma$ are defined recursively as is usual in moving average models. Additionally, W_i is the covariance matrix of ε_i^2, i.e. $W_i=\mathrm{Var}(\varepsilon_i^2)$. As in Refs. [7-8], a sandwich "working" covariance structure $W_i=A_i^{1/2}R_i(\delta)A_i^{1/2}$ is used to approximate the true W_i, where A_i is a diagonal matrix with $2\sigma_{ij}^4$ along the diagonal, and $R_i(\delta)$ mimics the correlation between ε_{ij}^2 and ε_{ik}^2 ($i\neq k$) by introducing the new parameter δ. Typical structures for $R_i(\delta)$ include compound symmetry and AR(1). As with the conventional generalized estimating equations for the mean, the

parameter δ may have little effect on the estimation of γ and ρ. Our real data analysis and simulation studies also confirm this point very well. It implies that the resulting estimators of parameters in the mean, moving average coefficients and innovation variances are robust against misspecification of $R_i(\delta)$, as is the efficiency of the estimation in the mean parameters only. At last, we define the solution of generalized estimating equations $\hat{\theta}$, $\hat{\gamma}$ and $\hat{\rho}$, as the generalized estimating equation estimators of θ, γ and ρ.

1.3 Main algorithm

An application of the quasi-Fisher scoring algorithm on Eq. (4) can directly yield the numerical solutions iteratively for θ, γ and ρ by fixing the other parameters respectively. More specifically, the algorithm works as follows.

① Set $k=0$, initialize the parameters as $\theta^{(0)}$, $\gamma^{(0)}$ and $\rho^{(0)}$.

② Compute Σ_i using $\gamma^{(k)}$ and $\rho^{(k)}$. Then given Σ_i, update θ by

$$\theta^{(k+1)} = \theta^{(k)} + [\sum_{i=1}^{m} B_i' \Delta_i (B_i\theta) \Sigma_i^{-1} \Delta_i (B_i\theta) B_i]^{-1} \cdot \sum_{i=1}^{m} B_i' \Delta_i \Sigma_i^{-1} (y_i - \mu_i) \mid_{\theta=\theta^{(k)}} \tag{5}$$

③ Given $\theta=\theta^{(k+1)}$ and $\rho=\rho^{(k)}$, update γ via

$$\gamma^{(k+1)} = \gamma^{(k)} - \left[\{ \sum_{i=1}^{m} G_i(\gamma) \}^{-1} \sum_{i=1}^{m} \left(\frac{\partial \varepsilon_i'}{\partial \gamma} \right) D_i^{-1} \varepsilon_i \right] \Big|_{\gamma=\gamma^{(k)}} \tag{6}$$

where

$$G_i(\gamma) = \sum_{j=2}^{n_i} \frac{1}{\sigma_{ij}^2} (W_{ij} + \sum_{k=1}^{j-1} a_{ijk} W_{ik}) \cdot D_i (W_{ij} + \sum_{k=1}^{j-1} a_{ijk} W_{ik})';$$

$W_{ij} = (w_{ij1}, \cdots, w_{ij(j-1)}, 0, \cdots, 0)$ is a $q \times n_i$ matrix and a_{ijk} is the (j,k)th element of L_i^{-1}. The $d \times n_i$ matrix $W_{i1}=0$.

④ Given $\theta = \theta^{(k+1)}$ and $\gamma = \gamma^{(k+1)}$, update ρ using

$$\rho^{(k+1)} = \rho^{(k)} + [\sum_{i=1}^{m} H_i' D_i W_i^{-1} D_i H_i]^{-1} \cdot \sum_{i=1}^{m} H_i' D_i W_i^{-1} (\varepsilon_i^2 - \sigma_i^2) \mid_{\rho=\rho^{(k)}} \tag{7}$$

⑤ Set $k \leftarrow k+1$ and repeat Steps ②～⑤ until a prespecified convergence criterion is met.

Note that a proper initial value is vital to making the algorithm converge well. A good starting value for θ is to choose $\Sigma_i^{(0)}$ as an identity matrix. According to the theory of GEE, this initial value of $\Sigma_i^{(0)}$ guarantees the consistency of the initial estimators in the mean, which in return guarantees consistency of the moving average parameters and innovative parameters after the first iteration. We update θ, γ and ρ iteratively until a pre-chosen convergence criterion is met. In our numerical study part, convergence was usually obtained within several iterations of this algorithm when choosing the Euclidean norm of the successive difference less than 10^{-6} as the convergence criterion.

1.4 Knots selection

Knots selection is important in spline smoothing. Because the number of distinct knots k_n has to increase with m for asymptotic consistency, but too many knots would also increase the variance of estimators. Similar to Refs. [8, 22], we use the sample quartiles of $\{t_{ij}, i=1,\cdots,m; j = 1, \cdots, n_i\}$ as knots. For example, if we use four internal knots, they are taken to be the four 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 $n_t^{1/5}$, where n_t is the number of distinct values in $\{t_{ij}, i = 1, \cdots, m; j = 1, \cdots, n_i\}$. This particular choice is consistent with the asymptotic theory of Section 2 and for the purpose of simplicity, it works well in a wide variety of problems according to our experience. Data-adaptive methods such as cross-validation can also be used for knots selection but are computationally more demanding, which is beyond the scope of the article.

2 Asymptotic properties

In this paper, for a vector a, its Euclidean norm is denoted by $\| a \|$, and for any square

matrix A, $\| A \|$ denotes its modulus of the largest singular value of A. To study the rates of convergence for $\hat{\beta},\hat{\gamma},\hat{\lambda}$ and $\hat{f}_0$, $\hat{f}_1$, we first give a set of regularity conditions and explanations. If the estimating Eq. (4) has multiple solutions, then only a sequence of consistent estimators $(\hat{\theta},\hat{\gamma},\hat{\rho})$ is considered in this section. A sequence $(\hat{\theta},\hat{\gamma},\hat{\rho})$ is said to be a consistent sequence if $(\hat{\theta}',\hat{\gamma}',\hat{\lambda}')' \to (\theta_0',\gamma_0',\lambda_0')'$ and $\sup_t | \pi'(t)\hat{\alpha} - f_0(t) | \to 0$, $\sup_t |\tilde{\pi}'(t)\hat{\tilde{\alpha}} - f_1(t)| \to 0$ in probability as $n \to \infty$.

The following assumptions are required for our asymptotic results:

(A Ⅰ) The number of independent subjects n goes to infinity and $\max_i n_i$ is bounded. We also require the dimensions p, q, d of covariates x_{ij}, w_{ijk} and z_{ij} are fixed, and assume without loss of generality that the $t_{ij}'s$ are all scaled into the interval $[0, 1]$. The first four moments of y_{ij} exist.

(A Ⅱ) The sth derivatives of f_0 and f_1 are bounded for some $s \geqslant 2$.

(A Ⅲ) The covariates w_{ijk} and matrices W_i^{-1} are all bounded. The function $g^{-1}(\cdot)$ has locally bounded second derivatives.

(A Ⅳ) The parametric space Θ is a compact subset of $\mathbb{R}^{p+q+d}$, and the true parameter value $(\beta_0',\gamma_0',\lambda_0')'$ is in the interior of the parameter space Θ.

The assumptions (A Ⅰ)～(A Ⅳ) are standard. The existence of first four moments of the response is needed for consistently estimating the parameters in the variance. The smoothness conditions (A Ⅱ) determine the rate of convergence of the spline estimates. Condition (A Ⅲ) is satisfied as t is bounded. Assumption (A Ⅳ) is routinely made in linear models.

To study the asymptotic properties of estimators, we assume the dependence between x_{ijk}, z_{ijk} and t_{ij} as follows:

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

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

$$j = 1,\cdots,n_i;\ i = 1,\cdots,m;$$

where δ_{ijk} and $\tilde{\delta}_{ijl}$ are mean zero random variables independent of the corresponding random errors and of one another. Let Λ_n and $\tilde{\Lambda}_n$ be the $n \times p$ and $n \times d$ matrices with $n = \sum n_i$ whose kth column are $\delta_k = (\delta_{11k}, \cdots, \delta_{1n_1k}, \cdots, \delta_{mn_mk})'$ and a similar definition to $\tilde{\delta}_k$. We make the following assumptions:

(A Ⅴ) ① $(E\Lambda_n = 0,\ \sup_n \frac{1}{n} E \| \Lambda_n \|^2 < \infty)$, and so as to $\tilde{\Lambda}_n$.

② $k_n M' \Sigma^0 M$ and $k_n M' W^0 M$ are nonsingular for a sufficiently large n, and the eigenvalues of $k_n M' \Sigma^0 M/n$ and $k_n M' W^0 M/n$ are bounded away from 0 and infinity, where $M = (\pi_1', \cdots, \pi_m')$, $\Sigma^0 = \mathrm{diag}\{\Sigma_1^0, \cdots, \Sigma_m^0\}$ with

$$\Sigma_i^0 = \Delta_{0i}\Sigma_{0i}^{-1}\Delta_{0i} = \Delta_i(X_i\beta_0 + f_0(t_i))\Sigma_{0i}^{-1}(\gamma_0,\lambda_0,f_1)\Delta_i(X_i\beta_0 + f_0(t_i))$$

and W^0 is defined in a similar way.

We take the number of knots k_n as the integer part of $N^{1/(2s+1)}$, where s is defined in (A Ⅱ) and taken as ② in this work. For this knot number, Condition ② of (A Ⅴ) is expected to hold as this is a property of the B-spline basis functions[12].

The asymptotic properties of $(\hat{\beta}_m, \hat{\gamma}_m, \hat{\lambda}_m)$ involve computation of the covariance matrix $\Omega_m = (\delta_m^{kl})_{k,l=1,2,3}$ of $\frac{1}{\sqrt{m}}(S_{10}', S_{20}', S_{30}')'$, where S_{10}, S_{20} and S_{30} are defined by

$$\left.\begin{aligned} S_{10} &= S_1(\beta_0;\gamma_0,\lambda_0) = \sum_{i=1}^m X_i^{*'}\Delta_{0i}\Sigma_{0i}^{-1}(y_i - \mu_{0i}),\\ S_{20} &= S_2(\gamma_0;\beta_0,\lambda_0) = \sum_{i=1}^m \left(\frac{\partial \varepsilon_{0i}'}{\partial \gamma}\right) D_{0i}^{-1}\varepsilon_{0i},\\ S_{30} &= S_3(\lambda_0;\beta_0,\gamma_0) = \sum_{i=1}^m Z_i^{*'}D_{0i}W_{0i}^{-1}(\varepsilon_{0i}^2 - \sigma_{0i}^2) \end{aligned}\right\} \tag{10}$$

where $X^* = (I - P)X$, $P = M(M'\Sigma^0 M)^{-1}M'\Sigma^0$ and a similar definition to Z^*, $\mu_{0i} = x_i'\beta_0 + f_0(t_i)$, $\varepsilon_{0i} = L_i^{-1}(y_i - \mu_{0i})$ and $\log(\sigma_{0i}^2) = z_i'\lambda_0 + f_1(t_i)$.

We also assume the covariance matrix Ω_m satisfying the following property:

(A Ⅵ) The covariance matrix Ω_m is positive definite, and there exists a positive definite matrix Ω such that

$$\lim_{m\to\infty}\Omega_m = \Omega = \begin{pmatrix} \omega^{11} & \omega^{12} & \omega^{13} \\ \omega^{21} & \omega^{22} & \omega^{23} \\ \omega^{31} & \omega^{32} & \omega^{33} \end{pmatrix} \tag{11}$$

Additionally, $\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} G_{0i}(\gamma_0) = \omega^* > 0$.

Theorem 2.1 If Assumptions (AⅠ)～(AⅥ) hold and the number of knots $k_n = O(n^{1/(2s+1)})$, then

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

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

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

As pointed out in Ref. [22], (12) and (13) imply that $\int(\hat{f}_i(t)-f_i(t))^2\mathrm{d}t = O_p(n^{-2s/(2s+1)})$, $i=0,1$, under general conditions (see, e. g., Lemmas 8 and 9 in Ref. [23]). This is the optimal rate of convergence for estimating f_0, f_1 under the smoothness Assumption (AⅡ).

Theorem 2.2 Under Assumptions (AⅠ)～(AⅥ), the generalized estimating equation estimator $(\hat{\beta}'_m, \hat{\gamma}'_m, \hat{\lambda}'_m)'$ is $\sqrt{m}$-consistent and asymptotically normal, that is

$$\hat{\beta}_m = \beta_0 + (X^{*'}\Sigma^0 X^*)^{-1}S_{10} + o_p\left(\frac{1}{\sqrt{m}}\right) \tag{14}$$

$$\hat{\gamma}_m = \gamma_0 + (\sum_{i=1}^{m} G_{0i})^{-1}S_{20} + o_p\left(\frac{1}{\sqrt{m}}\right) \tag{15}$$

$$\hat{\lambda}_m = \lambda_0 + (Z^{*'}W^0 Z^*)^{-1}S_{30} + o_p\left(\frac{1}{\sqrt{m}}\right) \tag{16}$$

Consequently,

$$\sqrt{m}\begin{pmatrix} \hat{\beta}_m - \beta_0 \\ \hat{\gamma}_m - \gamma_0 \\ \hat{\lambda}_m - \lambda_0 \end{pmatrix} \to N(0, Q^{-1}\Omega Q^{-1})$$

in distribution as $m\to\infty$, and the diagonal block matrix $Q=\mathrm{diag}(\omega^{11},\omega^*,\omega^{33})$.

From Theorem 2.2, we see that the asymptotic variance could reduce to a diagonal matrix for the normally distributed response variables. For statistical inference, we use a robust estimator of the covariance matrix of $\hat{\beta}_m$ named the "sandwich" estimator of $\mathrm{Cov}(\hat{\beta}_m)$ as follows:

$$\mathrm{Cov}(\hat{\beta}_m) = M_0^{-1}M_1M_0^{-1} \tag{17}$$

where

$$M_0 = \sum_{i=1}^{m} X_i^{*'}\hat{\Delta}\hat{\Sigma}_i^{-1}\hat{\Delta}X_i^*,$$

$$M_1 = \sum_{i=1}^{m} X_i^{*'}\hat{\Delta}\hat{\Sigma}_i^{-1}(y_i-\hat{\mu}_i)(y_i-\hat{\mu}_i)'\hat{\Sigma}_i^{-1}\hat{\Delta}X_i^*.$$

The estimated covariance matrices of $\hat{\gamma}_m$ and $\hat{\lambda}_m$ can be obtained in a similar way.

3 Numerical studies

For brevity, we refer to our proposed approach as semiparametric modeling by moving average (SMA) decomposition approach, and the methods in Ref. [8] as semiparametric modeling by autoregressive (SAR) decomposition approach. In this section, we first study the performance of our approach through extensive simulations. Finally, we apply our approach to the CD4+ cell number dataset, and the comparison between our approach and the conventional GEE and SAR is conducted.

3.1 Simulation study

In this section we investigate the finite sample performance of our proposed statistical estimation and inference methods. For each setup, we generate 1 000 data sets, and consider subject sizes $m=100$, 200 or 500, respectively. Each subject is supposed to be measured n_i times with $n_i-1\sim$ Binomial(12, 0.8), and the measurement time $t'_{ij}s$ are generated from the uniform distribution, leading to different numbers of repeated measurements n_i for each subject, which is an unbalanced longitudinal dataset.

Study 1 We first consider the following mean model to investigate the finite sample performance of our proposed approach and the impact of working covariance parameter δ:

$$\left.\begin{aligned} y_{ij} &= x_{ij1}\beta_1 + x_{ij2}\beta_2 + f_0(t_{ij}) + e_{ij}, \\ i &= 1,\cdots,m;\ j = 1,\cdots,n_i \end{aligned}\right\} \tag{18}$$

where $x_{ij1}=t_{ij}+\delta_{ij}$, δ_{ij} follows the standard normal distribution and x_{ij2} follows a Bernoulli distribution

with success probability 0.5. As for the nonparametric function in the mean, we take $f_0 = \cos(\pi t)$, and the errors $(e_{ij1}, \cdots, e_{ijn_i})'$ follow a multivariate normal distribution with mean 0 and covariance Σ_i satisfying $\Sigma_i = L_i D_i L_i'$, where L_i and D_i are modeled by Eq. (2) with $w_{ijk} = (1, t_{ij} - t_{ik})'$, $z_{ij} = x_{ij}$, and $f_1(t) = \sin(\pi t)$. We utilize AR(1) structure for $R_i(\delta)$ in $W_i = A_i^{1/2} R_i(\delta) A_i^{1/2}$, the working covariance of ε_i^2. For each simulated data set, we use $\delta = 0$, 0.2, 0.5 and 0.8 to study the robustness of our approach with regard to δ. Since the compound symmetry (exchangeable) structure has the similar results as AR(1), the details are omitted here. Also, we set the true value of these parameters as $\beta = (1, 0.5)'$, $\gamma = (0.2, 0.3)'$ and $\lambda = (-0.5, 0.2)$; the expected sample size is about 1 060, so the number of knots in B-spline is taken to be $4 \approx 1\,060^{1/5}$.

Tab. 1 shows that our SMA method yields unbiased estimators of the parametric parts in both the mean and covariance models. Meanwhile, we can see the parameter δ used in the working covariance structure for the innovations has little effect on either the estimation of β, γ, λ or the mean square errors in f_0 and f_1. These results confirm that our SMA is robust against misspecification of the structure of $R_i(\delta)$. Fig. 1 demonstrates the true and fitted curves for nonparametric functions f_0 and f_1 when $R_i(\delta)$ is specified by AR(1) with $\delta = 0.2$. The three curves $\hat{f}_5$, $\hat{f}_{50}$ and $\hat{f}_{90}$ represent the fits which are 5%, 50% and 90% best in terms of the mean squared errors in 1 000 runs, respectively. All these curves show a good agreement in fitting the true nonparametric functions.

Study 2 With the simulation setup in Study 1, we verify the performance of the asymptotic covariance Eq. (17) in Theorem 2.2. We refer to the sample standard deviation of 1 000 estimates as *SD*, which can be viewed as the true standard deviation of the resulting estimates. Meanwhile, we define *SE* as the sample average of 1 000 estimated standard errors using Eq. (17), and *std* as the standard deviation of these 1 000 standard errors. Tab. 2 shows that the standard error formula works consistently with *SD* under AR(1) correlation structures with different $\delta = 0$, 0.2, 0.5, 0.8.

Tab. 1 Simulation results for Study 1 over 1 000 replications. The estimates of parametric parts in both the mean and covariance models with sample standard errors in parentheses

	true	$\delta=0$	$\delta=0.2$	$\delta=0.5$	$\delta=0.8$
β_1	1.0	0.999 2 (0.035 7)	1.001 1 (0.035 8)	1.002 6 (0.035 4)	0.994 2 (0.038 7)
β_2	0.5	0.501 7 (0.067 9)	0.499 5 (0.069 6)	0.499 2 (0.071 8)	0.491 1 (0.078 6)
γ_1	0.2	0.196 7 (0.024 3)	0.196 1 (0.023 2)	0.196 4 (0.024 4)	0.197 5 (0.022 9)
γ_2	0.3	0.305 4 (0.055 4)	0.308 0 (0.054 9)	0.307 2 (0.059)	0.302 2 (0.054)
λ_1	−0.5	−0.504 8 (0.047 9)	−0.505 2 (0.048 5)	−0.509 4 (0.046 6)	−0.502 3 (0.045 4)
λ_2	0.2	0.199 2 (0.095 7)	0.197 6 (0.091 4)	0.202 4 (0.095 0)	0.203 7 (0.091 5)
MSE($\hat{f}_0$)	—	0.023 0 (0.021 8)	0.023 3 (0.021 7)	0.023 1 (0.021 5)	0.022 1 (0.021 8)
MSE($\hat{f}_1$)	—	0.020 9 (0.011 6)	0.021 1 (0.011 4)	0.020 9 (0.011 3)	0.020 8 (0.011 3)

【Note】 MSE($\hat{f}_i$), $i = 0, 1$, is the mean square error for the estimate f_i over all time points in the data

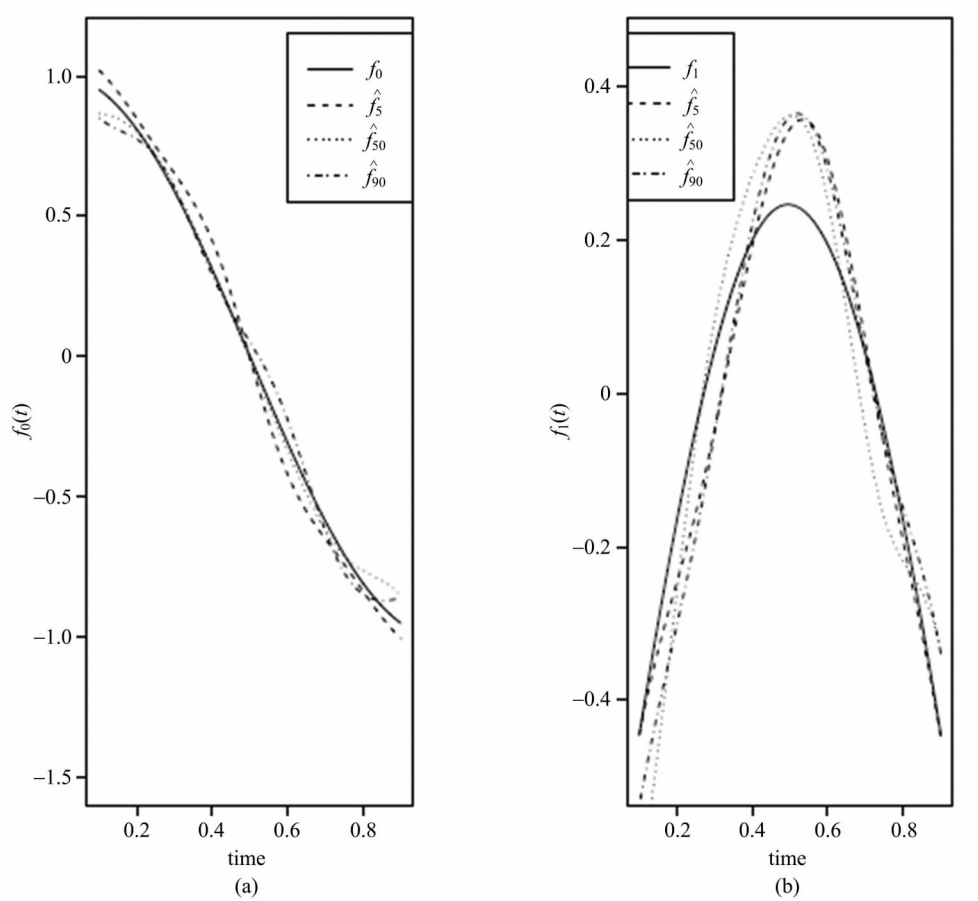


Fig. 1　Nonparametric function f_0 and f_1 and their fitted curves $\hat{f}_5$, $\hat{f}_{50}$, $\hat{f}_{90}$, for AR(1) structure with $\delta = 0.2$

Tab. 2　Comparison of the standard errors using Eq. (17) with sample standard deviation

		$\delta=0$	$\delta=0.2$	$\delta=0.5$	$\delta=0.8$
β_1	*SD*	0.035 7	0.035 8	0.035 4	0.038 7
	SE	0.033 6	0.033 6	0.033 7	0.033 4
	std	(0.003 2)	(0.003 3)	(0.003 3)	(0.003 2)
β_2	*SD*	0.067 9	0.069 6	0.071 8	0.078 6
	SE	0.068 2	0.068 1	0.068 1	0.067 6
	std	(0.005 2)	(0.005 4)	(0.005 5)	(0.005 8)
γ_1	*SD*	0.024 3	0.023 2	0.024 4	0.022 9
	SE	0.023 0	0.023 0	0.023 0	0.023 0
	std	(0.001 3)	(0.001 3)	(0.001 3)	(0.001 3)
γ_2	*SD*	0.055 4	0.054 9	0.058 6	0.053 8
	SE	0.053 0	0.053 0	0.053 1	0.053 0
	std	(0.005 4)	(0.005 3)	(0.005 0)	(0.005 3)
λ_1	*SD*	0.047 9	0.048 5	0.046 6	0.045 4
	SE	0.044 0	0.045 2	0.051 0	0.054 7
	std	(0.004 9)	(0.005 1)	(0.005 7)	(0.006 5)
λ_2	*SD*	0.095 7	0.091 4	0.095 1	0.091 5
	SE	0.088 7	0.091 5	0.101 9	0.108 2
	std	(0.006 7)	(0.007 7)	(0.009 6)	(0.011 5)

Study 3　In this study, we compare SMA approach and Leng et al.'s SAR approach[8] under different data generating processes. The main measurements for comparison are the differences between the fitted mean $\hat{\mu}_i$ and the true mean μ_i, the fitted covariance matrix $\hat{\Sigma}_i$ and the true Σ_i. In particular, we define two relative errors as

$$\mathrm{err}(\hat{\mu}_i) = \| \hat{\mu}_i - \mu_i \| / \| \mu_i \|,$$

$$\mathrm{err}(\hat{\Sigma}_i) = \| \hat{\Sigma}_i - \Sigma_i \| / \| \Sigma_i \|.$$

We compute the averages of these two indexes for 1 000 replications with $n = 100$, 200 for each dataset.

Case Ⅰ. We take a similar model in Study 1 to generate data sets except that $\gamma = c(0.1, 0.2)$ and $w_{ijk} = (1, (t_{ij} - t_{ik})^2)'$. In this case, SAR model is mis-specified.

Case Ⅱ. We generate data from SAR model. A similar model structure as in Case Ⅰ is

implemented by changing the l_{ijk} in Eq. (2) to ϕ_{ijk}. In this case, our SMA model is mis-specified.

Case Ⅲ. The main difference in our SMA model and the SAR model in Ref. [8] is that we decompose the covariance matrix itself instead of its inverse, the precision matrix. Therefore, to compare these two methods when models are mis-specified for both approaches, we take the mean model as in Case Ⅰ, but the covariance matrix with the blocking structure[9]. The diagonal entry σ_{ij} of Σ_i satisfies $\log(\sigma_{ij}^2)=z_{ij}'\lambda+f_1(t_{ij})$ with the same settings in Study 1; for $\left\lfloor\frac{j}{2}\right\rfloor\leqslant k\leqslant j-1$ where $\lfloor a\rfloor$ denotes the largest integer less than or equal to a, the (j,k)th element in Σ_i equals to $\sigma_{ij}^2\cdot 0.5^{j-k}$; and other elements are zeros.

Tab. 3 provides the average errors for SMA model and SAR model under these three cases. For Case Ⅰ where the data are generated from our model, our approach is substantially better than the alternative one in all comparison criteria. For Case Ⅱ where the data are generated from the autoregressive decomposition model, our approach still works reasonably well. The error measurements by our approach only inflate slightly compared to the alternative approach that fully exploits the model information. Therefore, when the true covariance matrix follows the moving average structure, the errors in estimating μ and Σ both increase when incorrectly decomposing the covariance matrix using the autoregressive structure, and vice versa. The magnitude of inflation in the errors totally depends on the data generating process. However, model misspecification seems to affect the moving average decomposition to a less degree in this simulation. But in other simulations not reported here, it could affect the moving average decomposition to a great degree. For Case Ⅲ where a model is mis-specified for both approaches, our method works satisfactorily.

Tab. 3 Average of relative errors $\mathrm{err}(\hat{\mu})=\sum_{l=1}^{m}\mathrm{err}(\hat{\mu}_l)/m$ **and** $\mathrm{err}(\hat{\Sigma})=\sum_{l=1}^{m}\mathrm{err}(\hat{\Sigma}_l)/m$

fit / true	n	SMA		SAR	
		$\mathrm{err}(\hat{\mu})$	$\mathrm{err}(\hat{\Sigma})$	$\mathrm{err}(\hat{\mu})$	$\mathrm{err}(\hat{\Sigma})$
SMA	100	0.096 1	0.152 6	0.098 7	0.160 0
	200	0.069 1	0.103 0	0.068 9	0.121 4
SAR	100	0.100 3	0.157 5	0.098 7	0.139 8
	200	0.066 3	0.119 8	0.070 0	0.099 1
blocking covariance structure	100	0.106 6	0.372 6	0.107 2	0.405 3
	200	0.073 8	0.350 4	0.074 3	0.397 4

3.2 Real data analysis

In this part, we restudy the CD4+ cell data with our proposed estimation method. The HIV causes AIDS by reducing a person's ability to fight infection, which could decrease the number of CD4+ cells in infected individuals. Thus, an infected person's CD4+ cell number can be used to monitor disease progress. This dataset includes 2 376 values of CD4+ cell number for 369 infected men, and is highly unbalanced since each individual has a different number of repeated measurements and unequally spaced time points. Here, we use the square root transformation of the response by the suggestion in Ref. [11] to relate the CD4+ counts to 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}, and mental illness scores x_{ij5}.

The objectives of the longitudinal analysis in this dataset are to identify factors which influence CD4+ cell changes and the covariance structures for the CD4+ cell data. For the mean model, we consider

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

For the covariance structure, we take covariates for the moving average coefficients as $w_{ijk}=(1,t_{ij}-t_{ik},(t_{ij}-t_{ik})^2,(t_{ij}-t_{ik})^3)'$ similar to the points in Ref. [7], and for the log-innovation variances as $z_{ij}=x_{ij}$, which allows us to examine whether the innovations are dependent on the covariates. Here, the number of knots is taken to be $[(2376)^{1/5}]=7$, which is the optimal number of knots that promise the convergence for the

nonparametric parts f_0, f_1 in Theorem 2.1.

Tab. 4 shows the estimating results for β by our SMA, where the working structure $R_i(\delta)$ is used as AR(1) with $\delta=0.2$ 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. Additionally, we also compare the SAR method in Ref. [8] with our current SMA model. The results show that both SMA and SAR provide estimators with generally smaller standard errors compared with conventional GEE. SMA is in accord with SAR that smoking and drugs are highly significant variables, while mental illness score is marginally significant. But in the GEE results, the significance of smoking is missed under the AR(1) covariance structure; drug use is also missed under either AR(1) or exchangeable variance structure; furthermore, the estimators under the independent working correlation indicate that mental score is not significant, which contradicts the results using other working correlations.

Tab. 4　CD4+ cell data

	SMA	SAR	GEE		
			independence	AR(1)	exchangeable
β_1	0.027 4 (0.033 8)	0.005 (0.030)	0.015 (0.035)	0.016 (0.034)	0.002 (0.032)
β_2	0.620 2 (0.136 1)	0.768 (0.130)	0.981 (0.184)	0.262 (0.190)	0.596 (0.136)
β_3	0.849 1 (0.320 3)	0.821 (0.345)	1.075 (0.528)	0.471 (0.350)	0.494 (0.358)
β_4	0.049 3 (0.036 4)	0.044 (0.038)	−0.064 (0.059)	0.050 (0.041)	0.060 (0.043)
β_5	−0.034 9 (0.013 57)	−0.030 (0.014)	−0.031 (0.021)	−0.046 (0.014)	−0.048 (0.015)

【Note】 The estimates of parametric parts in the mean model based on square root CD4+ cell numbers, with standard errors in parentheses

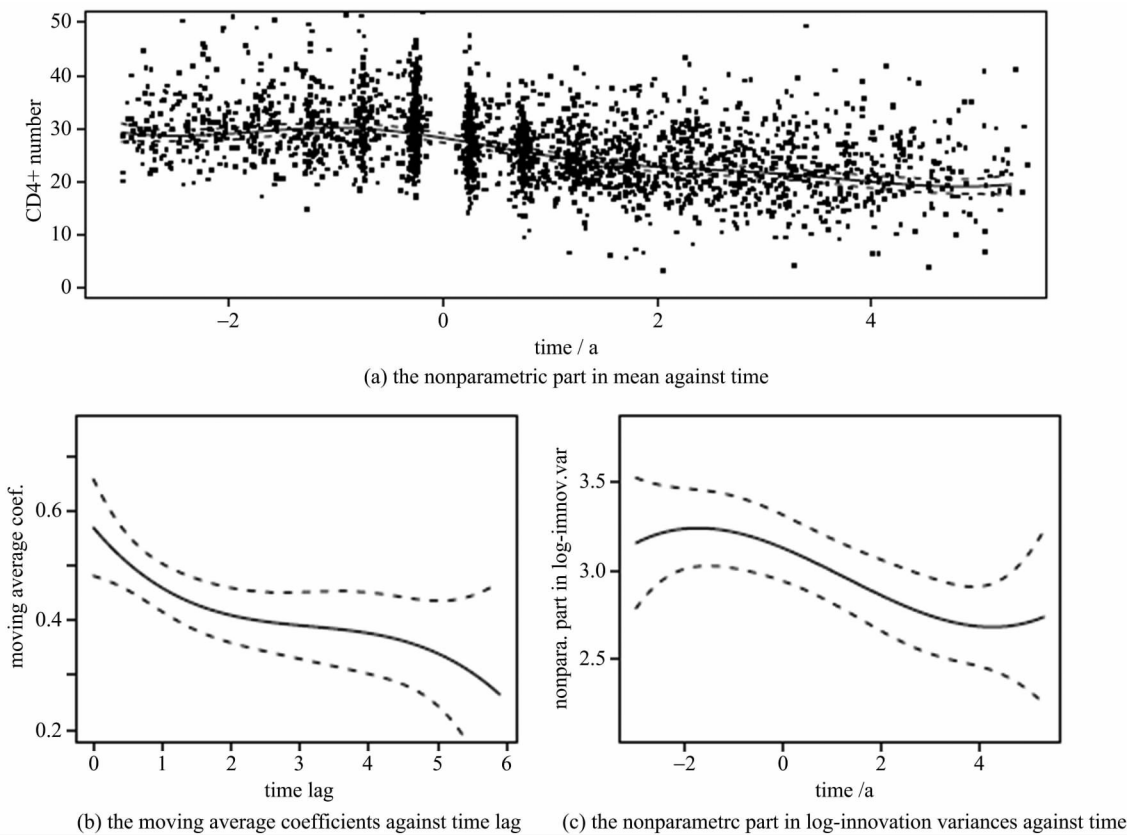
For the moving average and log innovation parameters, our model yields estimators with standard errors in parentheses as $\gamma_1=0.5877_{(0.0443)}$, $\gamma_2=-0.1622_{(0.0668)}$, $\gamma_3=0.0467_{(0.0296)}$, $\gamma_4=-0.0047_{(0.0037)}$, $\lambda_1=-0.0003_{(0.0070)}$, $\lambda_2=0.0842_{(0.0284)}$, $\lambda_3=0.0423_{(0.0872)}$, $\lambda_4=0.009_{(0.0128)}$, $\lambda_5=-0.0061_{(0.0037)}$. Then we can calculate the covariance matrix using $\Sigma_i=L_iD_iL_i'$.

Fig. 2 illustrates the fitted curves for f_0, moving average coefficients, and f_1 as a function of time and time lag. Here $R(\delta)$ in the working covariance structure of log-innovation variances is specified by AR(1) with $\delta=0.2$. Fig. 2(a) shows the nonparametric part of mean function changes slowly during the time since seroconversion, which indicates that the trajectory of the mean curve is consistent. Fig. 2(b) displays the estimated moving average parameters l against the time lag between measurements in the same subject, and as the figure shows, is a cubic polynmial, which decrease more clearly in the time lag less than two years, then slightly as the time lag becomes larger. As for the changes of innovation against time, it seems more fluctuating based on Fig. 2(c).

We also compare our method with SAR approach in terms of prediction. Using leave-one-out method, we split the data into two parts, the first part is used for training data sets to fit the model, and the second part which only has one sample is called the testing data set. We repeat the process 369 times to make sure each subject could be treated as testing data. To justify whether the models are appropriate, we apply the predictive mean errors defined as $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2/n$, and predictive covariance error $\sum_{i=1}^{m}\|\Sigma_i-\hat{\Sigma}_i\|/m$. Here $\Sigma_i=(Y_i-\tilde{\mu}_i(\theta))\cdot(Y_i-\tilde{\mu}_i(\theta))'$ and $\hat{\Sigma}_i=\hat{L}_i\hat{D}_i\hat{L}'_i$. Then we have the average mean errors for SMA is 37.113, and SAR is 36.936; the average covariance errors for SMA is 220.761 and SAR is 229.069. From this we can see both SMA and SAR approaches are reliable in the mean model, and our method in estimating the covariance matrix outperforms the other.

4　Discussion

In joint semiparametric modeling of mean and covariance, we proposed using a new Cholesky



All the dashed curves represent asymptotic 95% confidence intervals

Fig. 2 The fitted curves with confidence intervals for the CD4+ cell data

decomposition with moving average interpretation to reparameterize the covariance matrix itself instead of its inverse by Leng et al.[8]. Obviously, our work provides an alternative approach to analyzing covariance matrix. The main advantage of our work is that, estimating the covariance matrix directly is computationally efficient, and preserves the structural characteristics. Especially when the covariance matrix itself has a certain structure, like banding or blocking, directly decomposing the covariance matrix itself instead of its inverse could retain such structural characteristics, thus achieving a more efficient estimation.

The preference of our model over that by Leng et al.[8] is likely dependent on data. In practice, we may rely on a combination of graphical tools such as regressograms[4], which is suitable for balanced data sets, and numerical tools such as cross validation to choose an appropriate factorization and parametrization. More work needs to be done in this direction. If a clear trend is spotted in the sample regressogram, the corresponding factorization may be preferred. Quantitatively, we can always employ cross validation for comparing the predictive performance for estimating the mean and the observed covariance. A more accurate prediction is an indication to use the corresponding decomposition.

In this paper, we only consider the classical setup when the covariates are finite dimensional and continuous responses. It will be interesting to investigate the statistical properties with diverging numbers of parameters both in the mean and the variance and extend the approach to categorical longitudinal data analysis.

Appendix

The following lemma, which follows easily from Ref. [24, Theorem 12.7], is stated for easy

reference.

Lemma A1　Under Assumptions (A Ⅰ) and (A Ⅱ), there exist constants C_0 and C_1 such that

$$\sup_{t\in[0,1]} | f_0(t)-\pi'(t)\alpha_0 | \leqslant C_0 k_n^{-s},$$
$$\sup_{t\in[0,1]} | f_1(t)-\pi'(t)\tilde{\alpha}_0 | \leqslant C_1 k_n^{-s}.$$

Proof of Theorem 2.1　Here we only prove Eq. (13), and assume all W_i are known as W_{0i}, Similar asymptotic results hold when all W_{0i} are replaced by consistent estimates. The proof of Eq. (12) could be obtained from Ref. [22]. Throughout the following proof, suppose that C, C_0, C_1 always stand for positive constants and they may denote different values even within the same expression. From Lemma A1, for sufficiently large m, we can easily get

$$\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\hat{f}_1(t_{ij})-f_1(t_{ij})\}^2 =$$
$$\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\pi'_{ij}\hat{\tilde{\alpha}}-\pi'_{ij}\tilde{\alpha}_0+\pi'_{ij}\tilde{\alpha}_0-f_1(t_{ij})\}^2 \leqslant$$
$$\frac{2}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\pi'_{ij}(\hat{\tilde{\alpha}}-\tilde{\alpha}_0)\}^2+2C_0k_n^{-2s} \qquad (A1)$$

Let

$$T_n=\begin{pmatrix} A_n^{-1/2} & -A_n^{-1/2}Z'W^0M(M'W^0M)^{-1} \\ 0 & k_n^{1/2}Q_n^{-1}\end{pmatrix},$$

where

$$A_n = Z^{*'}W^0Z^* = \sum_{i=1}^{m}Z_i^{*'}W_i^0Z_i^*,$$
$$Z^* = (I-P)Z,$$
$$P = M(M'W^0M)^{-1}M'W^0,$$
$$Q_n^2 = k_nM'W^0M.$$

Direct calculation shows that

$$T_nH'W^0HT'_n = T_n\sum_{i=1}^{m}H'_iD_{0i}W_{0i}^{-1}D_{0i}H_iT'_n = I_{d+K},$$

where I_{d+K} stands for a $(d+K)\times(d+K)$ identity matrix. Further more, let

$$\zeta(\rho)=\begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}=(T'_n)^{-1}(\rho-\rho_0)=$$
$$\begin{pmatrix} A_n^{1/2}(\lambda-\lambda_0) \\ k_n^{-1/2}Q_n(\tilde{\alpha}-\tilde{\alpha}_0)+k_n^{1/2}Q_n^{-1}M'W^0Z(\lambda-\lambda_0)\end{pmatrix}$$

and $\hat{\zeta}'=\zeta(\hat{\lambda},\hat{\tilde{\alpha}})=(\hat{\zeta}'_1,\ \hat{\zeta}'_2)'$, then by Assumptions (A Ⅲ), (A Ⅴ) and the fact that $\| \hat{\zeta}_2 \| \leqslant \| \hat{\zeta} \|$ we have

$$\left[\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{\pi'_{ij}(\hat{\tilde{\alpha}}-\tilde{\alpha}_0)\}^2\right]^{1/2} =$$
$$n^{-1/2}\| M(\hat{\tilde{\alpha}}-\tilde{\alpha}_0)\| \leqslant$$
$$Cn^{-1/2}\| k_n^{-1/2}Q_n(\tilde{\alpha}-\tilde{\alpha}_0)\| \qquad (A2)$$
$$\leqslant Cn^{-1/2}\{\| \hat{\zeta}\| + \| k_n^{1/2}Q_m^{-1}M'W^0Z(\hat{\lambda}-\lambda_0)\|\} \leqslant$$
$$Cn^{-1/2}\| \hat{\zeta}\| + C\lambda_n^{-1/2}\| \hat{\lambda}-\lambda_0\| \cdot$$
$$\sup_{\|a\|=1,\|b\|=1} | n^{-1}a'M'W^0Zbk_n^{1/2} | \qquad (A3)$$

where λ_n is the minimum eigenvalue of $k_nM'W^0M/n$. Then by Ref. [12, Lemma 6.2], it suffices to verify $\| \hat{\zeta}\| = O_p(k_n^{1/2})$. The rest of the proof follows the same arguments as those of Ref. [8].

Computation of the Hessian matrix and its expectation

Since

$$\frac{\partial S_2}{\partial\gamma'} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\sigma_{ij}^2}\left[\frac{\partial\varepsilon_{ij}}{\partial\gamma}\left(\frac{\partial\varepsilon_{ij}}{\partial\gamma}\right)'+\frac{\partial^2\varepsilon_{ij}}{\partial\gamma\partial\gamma'}\varepsilon_{ij}\right]$$

where

$$\frac{\partial^2\varepsilon_{ij}}{\partial\gamma\partial\gamma'} = -\sum_{k=1}^{j-1}\left[w_{ijk}\frac{\partial\varepsilon_{ik}}{\partial\gamma'}+\frac{\partial\varepsilon_{ik}}{\partial\gamma}w'_{ijk}+l_{ijk}\frac{\partial^2\varepsilon_{ik}}{\partial\gamma\partial\gamma'}\right],$$

it is easy to show that

$$G_i(\gamma) = E\frac{\partial S_2}{\partial\gamma} = \sum_{j=1}^{n_i}\frac{1}{\sigma_{ij}^2}E\frac{\partial\varepsilon_{ij}}{\partial\gamma}\left(\frac{\partial\varepsilon_{ij}}{\partial\gamma}\right)' \qquad (A4)$$

Noting that

$$\partial\varepsilon_{ij}/\partial\gamma = -\sum_{k=1}^{j-1}[\varepsilon_{ik}w_{ijk}+l_{ijk}\partial\varepsilon_{ik}/\partial\gamma]$$

can be re-expressed as

$$\frac{\partial\varepsilon_{ij}}{\partial\gamma} = -W_{ij}\varepsilon_i - \sum_{k=1}^{j-1}a_{ijk}W_{ik}\varepsilon_i,\ j=2,\cdots,n_i \qquad (A5)$$

where

$$\varepsilon_i = (\varepsilon_{i1},\cdots,\varepsilon_{in_i})^{\mathrm{T}},$$
$$W_{ij} = (w_{ij1},\cdots,w_{ij(j-1)},0,\cdots,0)$$

is a $q\times n_i$ matrix and a_{jk} is the (j,k) element of lower triangle matrix L^{-1}, it then can be obtained

$$\sum_{i=1}^{m}G_{0i} = \sum_{i=1}^{m}\sum_{j=2}^{n_i}\frac{1}{\sigma_{0ij}^2}[W_{ij}+\sum_{k=1}^{j-1}a_{0jk}W_{ik}]\cdot$$
$$D_{0i}[W_{ij}+\sum_{k=1}^{j-1}a_{0jk}W_{ik}]^{\mathrm{T}} \qquad (A6)$$

with a_{0jk} being the (j,k) element of lower triangle matrix L^{-1} evaluated at γ_0.

Proof of Theorem 2.2

Eqs. (14) and (16) are exactly same as that in Ref. [8]. We only give a proof of Eq. (15). For this purpose, it is sufficient to prove it when all D_i are known as D_{0i}. By Taylor expansion and the second estimating equation in Eq. (4), we have

$$\sqrt{n}(\hat{\gamma}-\gamma_0)=\left(-\frac{1}{m}\frac{\partial S_2}{\partial \gamma'}\Big|_{\gamma=\tilde{\gamma}}\right)^{-1}\frac{1}{\sqrt{m}}S_2(\gamma_0),$$

where $\tilde{\gamma}=w\gamma_0+(1-w)\hat{\gamma}$ with some $0<w<1$.

According to the computation of the Hessian matrix and its expectation, we have

$$\frac{\partial S_2(\gamma)}{\partial \gamma'}=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\sigma_{0ij}^2}\frac{\partial \varepsilon_{ij}}{\partial \gamma}\left(\frac{\partial \varepsilon_{ij}}{\partial \gamma}\right)'+\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{1}{\sigma_{0ij}^2}\frac{\partial^2 \varepsilon_{ij}}{\partial \gamma\partial \gamma'}\varepsilon_{ij}.$$

Let

$$T_i=\sum_{j=1}^{n_i}\frac{1}{\sigma_{0ij}^2}\left[\frac{\partial \varepsilon_{ij}}{\partial \gamma}\left(\frac{\partial \varepsilon_{ij}}{\partial \gamma}\right)'+\frac{\partial^2 \varepsilon_{ij}}{\partial \gamma\partial \gamma'}\varepsilon_{ij}\right],$$

where

$$\frac{\partial^2 \varepsilon_{ij}}{\partial \gamma\partial \gamma'}=-\sum_{k=1}^{j-1}\left[w_{ijk}\frac{\partial \varepsilon_{ik}}{\partial \gamma'}+\frac{\partial \varepsilon_{ik}}{\partial \gamma}w'_{ijk}+l_{ijk}\frac{\partial^2 \varepsilon_{ik}}{\partial \gamma\partial \gamma'}\right],$$

then $\dfrac{\partial S_2(\gamma)}{\partial \gamma}=\sum_{i=1}^{m}T_i$. It has been shown that

$$E(T_i\mid_{(\beta_0,\gamma_0,\lambda_0)})=G_{0i}<\infty.$$

Further, noting that $\dfrac{\partial^2 \varepsilon_{ij}}{\partial \gamma\partial \gamma'}$ can be re-expressed as

$$\frac{\partial^2 \varepsilon_{ij}}{\partial \gamma\partial \gamma'}=-\sum_{k=1}^{j}\left[a_{ijk}W_{ik}\frac{\partial \varepsilon_i}{\partial \gamma'}+a_{ijk}\left(\frac{\partial \varepsilon_i}{\partial \gamma'}\right)'W'_{ik}\right]=$$

$$-\sum_{k=1}^{j}a_{ijk}\sum_{l=1}^{k-1}(w'_{ikl}(\sum_{r=1}^{l}a_{ilr}\varepsilon'_iW'_{ir})+(\sum_{s=1}^{l}a_{ils}W_{is}\varepsilon_i)w'_{ikl}).$$

Therefore, it can be obtained that

$$T_i=\sum_{j=1}^{n_i}\frac{1}{\sigma_{0ij}^2}\{\sum_{k=1}^{j}a_{ijk}W_{ik}\varepsilon_i(\sum_{k=1}^{j}a_{ijk}\varepsilon'_iW_{ik'})-U_i\varepsilon_{ij}\},$$

where

$$U_i=[\sum_{k=1}^{j}a_{ijk}\sum_{l=1}^{k-1}(w_{ikl}(\sum_{r=1}^{l}a_{ilr}\varepsilon'_iW'_{ir})+(\sum_{s=1}^{l}a_{ils}W_{is}\varepsilon_i)w'_{ikl})].$$

Hence $\mathrm{Var}(T_i\mid_{(\beta_0,\gamma_0,\lambda_0)})$ is bounded for each $i=1,2,\cdots,n$, it is verifiable that $\sum_{i=1}^{\infty}\mathrm{Var}(T_i)/i^2<\infty$. By the Kolmogorov's strong law of large numbers we have that

$$\frac{1}{m}\frac{\partial S'_2(\gamma)}{\partial \gamma}-\frac{1}{m}E\left\{\frac{\partial S'_2(\gamma)}{\partial \gamma}\right\}_{(\beta_0,\gamma_0,\lambda_0)}=\frac{1}{m}\frac{\partial S'_2(\gamma)}{\partial \gamma}-\frac{1}{m}\sum_{i=1}^{m}G_{0i}\to 0 \qquad (A7)$$

almost surely as $m\to\infty$.

Clearly,

$$S_2(\gamma_0)=\sum_{i=1}^{m}\left[\left(\frac{\partial \varepsilon'_i}{\partial \gamma}\right)D_{0i}^{-1}\varepsilon_i\right]_{\gamma=\gamma_0}=$$

$$\sum_{i=1}^{m}((L_{0i}^{-1}r_{0i})'\otimes I_q+(L_{0i}^{-1}\nabla\mu_i)'\otimes I_q)\cdot$$

$$\Omega'_i\Sigma_{0i}^{-1}(r_{0i}+\nabla\mu_i)=$$

$$S_{20}+\sum_{i=1}^{m}((L_{0i}^{-1}\nabla\mu_i)'\otimes I_q)\Omega'_i\Sigma_{0i}^{-1}\nabla\mu_i+$$

$$\sum_{i=1}^{m}((L_{0i}^{-1}\nabla\mu_i)'\otimes I_q)\Omega'_i\Sigma_{0i}^{-1}r_{0i}=$$

$$S_{20}+J_{11}+J_{12} \qquad (A8)$$

where $\Sigma_{0i}=L_{0i}D_{0i}L'_{0i}$, $r_{0i}=y_i-\mu_{0i}$, $\nabla\mu_i=\mu_{0i}-\mu_i$, I_q is the $q\times q$ identity matrix, Ω_i is $n_i\times n_iq$ matrix with the jth row

$-W_{ij}=(-w'_{ij1},-w'_{ij2},\cdots,-w'_{ij(j-1)},0,\cdots,0)$.

Let $a\in\mathbb{R}^q$ satisfy $a'a=1$, then

$$|Ea'J_{11}|\leqslant\sum_{i=1}^{m}|a'((L_{0i}^{-1}\nabla\mu_i)'\otimes I_q)\Omega'_i\Sigma_{0i}^{-1}\nabla\mu_i|\leqslant$$

$$C\sum_{i=1}^{m}\|\nabla\mu_i\|^2=O_p(n^{\frac{1}{2s+1}})=o_p(\sqrt{m}).$$

Similarly, we have $J_{12}=o_p(\sqrt{m})$, thus $S_2(\gamma_0)=S_{20}+o_p(\sqrt{m})$. The proof is then completed by an application of the Slutsky theorem. The asymptotic normality can then be obtained by following the Liapounov form of the multivariate central limit theorem.

References

[1] Diggle P J, Heagerty P J, Liang, K Y, et al. Analysis of Longitudinal Data [M]. Oxford, UK: Oxford University Press, 2002.

[2] Liang K Y, Zeger S L. Longitudinal data analysis using generalized linear models[J]. Biometrika, 1986, 73: 13-22.

[3] Wang Y G, Carey V. Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance[J]. Biometrika, 2003, 90: 29-41.

[4] Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parametrization[J]. Biometrika, 1999, 86: 677-690.
[5] Pourahmadi M. Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix[J]. Biometrika, 2000, 87: 425-435.
[6] Pan J, Mackenzie G. Model selection for joint mean-covariance structures in longitudinal studies [J]. Biometrika, 2003, 90: 239-244.
[7] Ye H, Pan J. Modelling of covariates structures in generalized estimating equations for longitudinal data [J]. Biometrika, 2006, 93: 927-941.
[8] Leng C, Zhang W, Pan J. Semiparametric mean-covariance regression analysis for longitudinal data[J]. Journal of the American Statistical Association, 2010, 105: 181-193.
[9] Rothman A J, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions[J]. Biometrika, 2010, 97(3): 539-550.
[10] Zhang W, Leng C. A moving average Cholesky factor model in covariance modeling for longitudinal data[J]. Biometrika, 2012, 99: 141-150.
[11] Zeger S L, Diggle P J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters[J]. Biometrics, 1994, 50: 689-699.
[12] He X, Shi P. Bivariate tensor-product B-splines in a partly linear model [J]. Journal of Multivariate Analysis, 1996, 58: 162-181.
[13] Carroll R J, Fan J, Gijbels I, et al. Generalized partially linear single-index models[J]. Journal of the American Statistical Association, 1997, 92(438): 477-489.
[14] Zhang D, Lin X, Raz J, et al. Semiparametric stochastic mixed models for longitudinal data [J]. Journal of the American Statistical Association, 1998, 93(442): 710-719.
[15] Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis [J]. Journal of American Statistical Association, 2004, 99(467): 710-723.
[16] Chen K, Jin Z. Partial linear regression models for clustered data[J]. Journal of the American Statistical Association, 2006, 101(473):195-204.
[17] Wu W, Pourahmadi M. Nonparametric estimation of large covariance matrices of longitudinal data [J]. Biometrika, 2003, 90: 831-844.
[18] Fan J, Huang T, Li R. Analysis of longitudinal data with semiparametric estimation of covariance function [J]. Journal of the American Statistical Association, 2007, 35: 632-641.
[19] Fan J, Wu Y. Semiparametric estimation of covariance matrices for longitudinal data [J]. Journal of the American Statistical Association, 2008, 103: 1 520-1 533.
[20] He X, Zhu Z Y, Fung W K. Estimating in a semiparametric model for longitudinal data with unspecified dependence structure [J]. Biometrika, 2002, 89: 579-590.
[21] Heckman N E. Spline smoothing in a partly linear model[J]. Journal of the Royal Statistical Society Ser B, 1986, 48: 244-248.
[22] He X, Fung W K, Zhu Z Y. Robust estimation in generalized partial linear models for clustered data[J]. Journal of the American Statistical Association, 2005, 472: 1 176-1 184.
[23] Stone C. Additive regression and other nonparametric models [J]. The Annals of Statistics, 1985, 13: 689-705.
[24] Schumaker L L. Spline Functions[M]. New York: Wiley, 1981.