 DOI: 10.1111/j.1541-0420.2005.00490.x

Quadratic Inference Functions for Varying-Coefficient Models with Longitudinal Data

Annie Qu

Department of Statistics, Oregon State University, Corvallis, Oregon 97331, U.S.A.
email: qu@stat.orst.edu

and

Runze Li

Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

Summary. Nonparametric smoothing methods are used to model longitudinal data, but the challenge remains to incorporate correlation into nonparametric estimation procedures. In this article, we propose an efficient estimation procedure for varying-coefficient models for longitudinal data. The proposed procedure can easily take into account correlation within subjects and deal directly with both continuous and discrete response longitudinal data under the framework of generalized linear models. The proposed approach yields a more efficient estimator than the generalized estimation equation approach when the working correlation is misspecified. For varying-coefficient models, it is often of interest to test whether coefficient functions are time varying or time invariant. We propose a unified and efficient nonparametric hypothesis testing procedure, and further demonstrate that the resulting test statistics have an asymptotic chi-squared distribution. In addition, the goodness-of-fit test is applied to test whether the model assumption is satisfied. The corresponding test is also useful for choosing basis functions and the number of knots for regression spline models in conjunction with the model selection criterion. We evaluate the finite sample performance of the proposed procedures with Monte Carlo simulation studies. The proposed methodology is illustrated by the analysis of an acquired immune deficiency syndrome (AIDS) data set.

Key words: Generalized method of moments; Goodness of fit; Model selection; Penalized spline; Quadratic inference function; Smoothing spline; Varying-coefficient model.

1. Introduction

Longitudinal data often occur in biomedical research where data are collected at irregular and possibly subject-specific time points. Due to their unbalanced nature, it is difficult to directly apply traditional multivariate regression techniques. To explore possible time-dependent effects, time-varying coefficient models and their extensions have been proposed for longitudinal data analysis. See, for example, Hoover et al. (1998), Wu, Chiang, and Hoover (1998), Fan and Zhang (2000), Martinussen and Scheike (2001), Chiang, Rice, and Wu (2001), Huang, Wu, and Zhou (2002), and references therein. These authors propose various estimation procedures for varying-coefficient models under longitudinal data settings, but they have not discussed how to incorporate information on the correlation structure within subjects into their estimation procedures. Furthermore, the aforementioned works only discuss continuous responses under a linear model framework. In this article, we are interested in developing a general approach for both continuous and discrete responses under a generalized linear model framework.

There are many nonparametric approaches for correlated data (Wang, 1998a, 1998b; Opsomer, Wang, and Yang, 2001, and references therein), but most of the nonparametric literature focuses on consistent and efficient estimation, including recent kernel and spline approaches by Lin and Carroll (2000), Wang (2003), Lin et al. (2004), and Wang, Carroll, and Lin (2005). Hart (1997) and references therein provide nonparametric goodness-of-fit tests, however most of these approaches treat response variables as normal outcomes. More recently, Zhang (2004) proposed generalized linear mixed models for hypothesis testing for varying-coefficient models, where the response variables could be nonnormal such as binary or Poisson; however, there is a strong parametric assumption for random effects, and typically the random effects are assumed to be normal. Since the high-dimensional likelihood for correlated discrete data usually does not have a closed form, numerical approximations such as the Laplace method might be required. In addition, Zhang's approach is to transform testing the varying-coefficient into a variance component testing problem, and the variance component has a nonnegative

constraint; therefore, Zhang's test is a mixture of chi-squared asymptotically. However, the large sample approximation of mixture chi-squared often performs poorly in simulation studies (Crainiceanu and Ruppert, 2004). Testing for more than one variance component could be even more complicated.

Our research is motivated by an analysis of a subset of longitudinal data from the Multi-Center AIDS Cohort study. The data set contains the HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. The number and time of measurements vary between individuals, with at least 1 to a maximum of 14, measurements. Huang et al. (2002) analyzed this data set by taking CD4 cell counts as the response variable, and demonstrated that the effects on baseline intercept, smoking status, age, and pre-CD4 counts might be time varying. However, their analysis ignored within-subject correlation. It is important to incorporate correlation structures in nonparametric local modeling for longitudinal data, since as Wang (2003) and Lin et al. (2004) demonstrated, kernel and smoothing spline estimators using the true covariance are more efficient than the independent structure.

Empirically, it is also difficult to estimate an unstructured covariance matrix for the following reasons: (1) the covariance matrix has to be positive definite; however, the estimator of the unstructured covariance matrix is often nonpositive definite for unbalanced longitudinal data (Lipsitz et al., 2000); (2) high-dimensional nuisance parameters could be involved if the data are measured over a long period of time; (3) the inverse of the covariance is needed and essential, therefore the smallest eigenvalues of the covariance matrix are the most important; but these are poorly estimated when the dimension of the covariance matrix is large (Qu and Lindsay, 2003). Even if the correlation matrix is assumed to possess a simplified working structure as in generalized estimating equations (GEEs) (Liang and Zeger, 1986), the estimation of the correlation matrix could still be nonpositive definite (Crowder, 1995).

Our goal in this article is to develop a unified approach that enables us to handle high-dimensional problems without losing efficiency. We propose an estimation procedure for varying-coefficient models using the penalized spline (Ruppert and Carroll, 2000) and quadratic inference function approaches (Qu, Lindsay, and Li, 2000). The proposed method allows us to directly incorporate correlations into model building, but does not require us to estimate the nuisance parameters associated with correlations. Under certain regularity conditions we establish the asymptotic normality of the resulting estimator, and show that our estimator is asymptotically efficient within the class where the moment conditions are satisfied.

Another goal of interest is to examine whether coefficient functions are time varying or are invariant. In general this is still challenging; as we mentioned earlier, the likelihood functions are intractable when data are correlated and discrete; in addition, a specific parametric alternative is not desirable in nonparametric models, and therefore a typical likelihood ratio test might not be applicable. Huang et al. (2002) constructed test statistics based on the difference of the residual sum of squares under the null and alternative. Their approach does not require likelihood functions; however, the asymptotic properties of their test have not been developed and they proposed bootstrap sampling strategies to determine a critical value.

This leads us to consider nonparametric goodness-of-fit tests for large samples. We propose a simple and efficient statistical inference procedure that does not require likelihood functions. The proposed test statistics have a chi-squared limiting distribution under the null hypothesis. In addition, we are able to perform a goodness-of-fit test for the model assumption. This provides an objective criterion for choosing basis functions in regression spline models and determining the number of knots in penalized spline approaches. A goodness-of-fit test has not been developed in the nonparametric literature for cases where the likelihood function is not available. In addition, we also apply Andrews's (1999) generalized method of moments Bayesian information criterion (BIC), which allows us to select between models when the goodness-of-fit tests fail to reject.

This article is organized as follows: In Section 2, we propose an estimation procedure under the varying-coefficient model, and further establish the strong consistency and asymptotic normality of the proposed estimator. In Section 3, we discuss some practical issues to implement the proposed estimation procedures. In Section 4, a nonparametric goodness-of-fit test and a model selection procedure are illustrated. We assess the finite sample performance of the proposed procedure with Monte Carlo simulation and illustrate the proposed methodology by the analysis of an AIDS data set in Section 5. Discussion is given in Section 6. Technical conditions and proofs are provided in the Appendix.

2. A New Estimation Procedure

In this section, we will illustrate how to estimate coefficient functions using the penalized spline, and how to incorporate correlation structures using quadratic inference functions. We start with a brief introduction to quadratic inference functions.

2.1 *Quadratic Inference Functions*

For longitudinal data, let $y_i(t)$ be a response variable and $x_i(t)$ be a $p \times 1$ vector of covariates, measured at time $t = t_1, \ldots, t_{n_i}$ for subjects $i = 1, \ldots, N$. We assume that the model satisfies the first moment model assumption

$$\mu_{it} = E\{y_i(t)\} = \mu\{x_i(t)'\beta\},$$

where $\mu(\cdot)$ is a known inverse link function and β is a p-dimensional parameter vector. The quasi-likelihood equation (Wedderburn, 1974) for longitudinal data is

$$\sum_{i=1}^{N} \dot{\mu}_i' V_i^{-1}(y_i - \mu_i) = 0,$$

where $V_i = \mathrm{var}(y_i), y_i = (y_i(t_1), \ldots, y_i(t_{n_i}))', \mu_i = (\mu_{it_1}, \ldots, \mu_{it_{n_i}})'$, and $\dot{\mu}_i = \partial\mu_i/\partial\beta$. In practice, V_i is often unknown, and the empirical estimator of V_i based on sample variance could be unreliable, especially when there is a small number of replications relative to a large number of variance components. Liang and Zeger (1986) introduced generalized estimating equations to simplify V_i by assuming $V_i = A_i^{1/2} R A_i^{1/2}$, where A_i is a diagonal marginal variance matrix and R is a common working correlation, which involves

a small number of nuisance parameters. If the working correlation R is misspecified, the estimator of the regression parameter is still consistent, but is not efficient within the same class of estimating functions.

Qu et al. (2000) introduced the quadratic inference function by assuming that the inverse of the working correlation can be approximated by a linear combination of several basis matrices, that is,

$$R^{-1} \approx a_0 I + a_1 M_1 + \cdots a_m M_m, \tag{1}$$

where I is the identity matrix and M_i are symmetric matrices. The advantage of this approach is that it does not require estimation of linear coefficients a_i's that can be viewed as nuisance parameters, since the generalized estimating equation is an approximate linear combination of elements of the estimating function

$$\begin{aligned}\bar{g}_N(\beta) &= \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) \\ &= \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N}(\dot{\mu}_i)' A_i^{-1}(y_i - \mu_i) \\ \sum_{i=1}^{N}(\dot{\mu}_i)' A_i^{-1/2} M_1 A_i^{-1/2}(y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N}(\dot{\mu}_i)' A_i^{-1/2} M_m A_i^{-1/2}(y_i - \mu_i) \end{pmatrix}. \end{aligned} \tag{2}$$

Because the dimension of the above estimating equation is greater than the number of unknown parameters, we cannot set each component in (2) to be zero to solve for β. Instead we estimate β by setting $\bar{g}_N$ as close to zero as possible, in the sense of minimizing the quadratic function

$$\hat{\beta} = \arg\min_{\beta} \bar{g}_N' \Omega^{-1} \bar{g}_N, \tag{3}$$

where $\Omega = \mathrm{var}(g_i)$, and we assume that N subjects are independent and identically distributed. The estimator in (3) is also called a generalized method of moments estimator in the econometrics literature (Hansen, 1982). The covariance Ω is invertible if estimating equations in g_i are not linearly dependent. The covariance Ω in (3) is often unknown, but it can be estimated consistently by $\bar{C}_N = N^{-1}\sum_{i=1}^{N} g_i g_i'$. Note that the additional requirement for $\bar{C}_N$ being invertible is $N \geq \dim(g_i)$. The quadratic function,

$$Q_N(\beta) = N \bar{g}_N' \bar{C}_N^{-1} \bar{g}_N, \tag{4}$$

is called the quadratic inference function (Qu et al., 2000), because it provides an inference function for testing of β. This approach also provides an optimal linear combination of given estimating functions such that the asymptotic variance of the estimator attains the minimum in the sense of Löwner ordering (e.g., Pukelsheim, 1993, p. 12).

2.2 *An Effective Estimation Procedure via Penalized Quadratic Inference Functions*

Under generalized linear model settings, varying-coefficient models assume the following mean structure:

$$E\{y_i(t_{ij}) \,|\, x(t_{ij})\} = h\{x_i'(t_{ij})\beta(t_{ij})\} = \mu_{ij}, \tag{5}$$

where $h(\cdot)$ is a known inverse link function. Varying-coefficient models were systematically introduced by Hastie and Tibshirani (1993). Here, we are interested in the longitudinal data setting where data are correlated within the same subject. Suppose $B_{uv}(t)$ is a set of basis functions of the functional space to which $\beta_u(\cdot)$ belongs, and $\beta_u(t)$ can be approximated by a linear combination of the basis functions. Specifically,

$$\beta_u(t) \approx \sum_{v=0}^{V_u} \gamma_{uv} B_{uv}(t), \quad \text{for} \quad u = 1, \ldots, p,$$

where γ_{uv}'s are constants, and V_u is associated with the number of basis functions for the uth coefficient. Substituting the approximation of $\beta_u(t)$ into (5), the mean function in (5) can be approximated by

$$h\{X_i'(t_{ij})\beta(t_{ij})\} \approx h\left[\sum_{u=1}^{p}\sum_{v=0}^{V_u}\{x_{iu}(t_{ij})B_{uv}(t_{ij})\}\gamma_{uv}\right].$$

The basis functions can be selected as polynomials, Fourier basis functions, or splines. Our approach is not restricted to one specific choice of basis functions, but here we consider only the q-degree truncated power spline basis with knots $\kappa_1, \ldots, \kappa_{K_u}$, that is

$$1, t, \ldots, t^q, (t-\kappa_1)_+^q, \ldots, (t-\kappa_{K_u})_+^q,$$

where $(z)_+^q = z^q I(z \geq 0)$. In Section 4, we will discuss how to choose q and the number of knots using a goodness-of-fit test, and the generalized method of moments model selection criteria. With the truncated power spline basis, the coefficient function can be modeled by

$$\beta_u(t) = \gamma_{u0} + \gamma_{u1} t + \cdots + \gamma_{uq} t^q + \sum_{k=1}^{K_u} \gamma_{u(q+k)}(t-\kappa_k)_+^q. \tag{6}$$

To incorporate correlation into the model, we apply the idea of the quadratic inference function in Section 2.1 and construct estimating functions as follows. We create $\bar{g}_N$ as in (2) with the mean of response μ_i in (5) approximated by using the basis in (6). We can further derive a quadratic inference function $Q_N(\gamma)$ in (5), which is a function of parameters $\gamma = \{\gamma_{uv},\, u = 1, \ldots, p;\, v = 0, \ldots, V_u\}$. Minimizing $Q_N(\gamma)$ yields an estimator for γ. Plugging the estimator of γ_{uv} into the basis expansion (6), we obtain an estimator for $\beta_u(t)$. However, it is well known that the model in (6) usually over-parameterizes the coefficient function, and therefore the resulting estimator is undersmoothed and has a large variance. To overcome this drawback, we borrow the idea of the penalized spline (Ruppert and Carroll, 2000) and propose a penalized quadratic inference function

$$N^{-1} Q_N(\gamma) + \lambda \gamma' D \gamma, \tag{7}$$

where D is a diagonal matrix with 1 if γ_{uv} is the coefficient of the truncated power function associated with the knots in (6), and 0 otherwise, and λ is a smoothing parameter, which can be chosen by data-driven methods such as cross-validation and generalized cross-validation. If λ is large, it has more shrinkage toward a polynomial fit, less weights on selected knots, and it is oversmoothed. On the other hand, if λ_N is small, it is undersmoothed. See Wand (1999), Ruppert (2002), Yu and Ruppert (2002, 2004), Kim, Cohen, and Carroll (2003), and Jarrow, Ruppert, and Yu (2004) on the penalized spline approach.

2.3 *Asymptotic Properties of Estimators*

In this section, we will study the asymptotic properties of the penalized quadratic inference function estimator. Asymptotic properties of penalized spline regression for independent and identically distributed observations have been studied by several authors using different formulations (Wand, 1999; Yu and Ruppert, 2002; Hall and Opsomer, 2005). Here, we focus only on fixed-knot asymptotics, since fixed-knot spline regression might be more useful for developing a practical statistical methodology, as argued by Yu and Ruppert (2002). We first establish the asymptotic properties of the penalized quadratic inference function estimator when the smoothing parameter λ_N goes to 0 as sample size N goes to infinity. Theorem 1 shows the strong consistency of the resulting estimator of the penalized quadratic inference function, and Theorem 2 establishes the root N consistency and asymptotic normality of the resulting estimator.

THEOREM 1: *Under Conditions A–D in the Appendix, if the smoothing parameter $\lambda_N = o(1)$, then the spline regression parameter estimator $\hat{\gamma}$ by minimizing* (7) *exists and converges to γ_0 almost surely.*

THEOREM 2: *Under Conditions A–E in the Appendix, if the smoothing parameter $\lambda_N = o(N^{-1/2})$, then the spline regression parameter estimator $\hat{\gamma}$ by minimizing* (7) *is asymptotically normal and efficient (i.e., the asymptotic variance of the estimator reaches lower bound). That is*

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N_k\big(0, \big(J_0' C_0^{-1} J_0\big)^{-1}\big),$$

where γ_0 is the parameter satisfying $E_{\gamma_0} g_i = 0, k = \sum_{u=1}^{p} V_u + p$, and $\frac{\partial \bar{g}_N}{\partial \gamma}$ and $\bar{C}_N = N^{-1} \sum_{i=1}^{N} g_i g_i'$ converge to J_0 and C_0 in probability.

It is important to point out that if the inverse of the true correlation belongs to the class in (1), then the quadratic inference function approach is as efficient as the GEE approach under the true correlation; if not, for example, the working correlation is misspecified, then the quadratic inference function approach is still optimal within the family where the inverse of the misspecified working correlation has an approximate linear representation of basis matrices (Qu et al., 2000). In other words, the quadratic inference function estimator is more widely efficient than the GEE estimator.

Next, we establish the asymptotic distribution when λ is treated as fixed. Notice that $\hat{\gamma}$ by minimizing (7) is asymptotically equivalent to solving

$$J_0'(\gamma) C_0^{-1}(\gamma) \bar{g}_N(\gamma) + \lambda D \gamma = 0.$$

Let $s_i = J_0' C_0^{-1} g_i + \lambda D \gamma$, then $\hat{\gamma}(\lambda)$ solves $\sum_{i=1}^{N} s_i \{\gamma(\lambda), \lambda\} = 0$. If we assume that $E[s_1\{\gamma(\lambda), \lambda\}] = 0$ for a fixed λ, then we are solving unbiased estimating equations. The following asymptotic distribution of $\hat{\gamma}(\lambda)$ can be derived based on estimating function theory (e.g., Heyde, 1997, Chapter 2):

$$\sqrt{N}\{\hat{\gamma}(\lambda) - \gamma(\lambda)\} \xrightarrow{d} N_k(0, H^{-1} G H^{-1}),$$

where $H = E(\partial s_1 / \partial \gamma) = J_0' C_0^{-1} J_0 + \lambda D$ and $G = E(s_1 s_1')$.

3. Practical Implementation Issues

In practical implementation, one has to choose the basis for the inverse of the correlation matrix, determine the magnitude of λ, and calculate the standard error and confidence interval of the resulting estimator. In this section, we address these practical issues.

3.1 *Choice of the Basis for the Inverse of the Correlation Matrix*

We discuss the choice of basis matrices M_i in (1) in this section. If the working correlation is exchangeable, we can choose a basis matrix M_1 with 0 on the diagonal and 1 off-diagonal. If the working correlation is AR(1), then M_1^* can be 1 on the subdiagonal and 0 elsewhere, and M_2^* can be 1 on (1, 1) and (N, N) components and 0 elsewhere. However, M_2^* can often be dropped out of the model, as removing M_2^* does not affect the efficiency of the estimator too much, but could simplify the estimation procedure. If we use both M_1 and M_1^*, then they are effective for modeling either the exchangeable or AR(1) working correlation. This is useful when there is uncertainty as to which working correlation structure is appropriate. Our simulation also confirms this finding. If there is no prior information on working correlation, Qu and Lindsay (2003) provide an adaptive estimation equation approach to approximate the true correlation empirically, and their approach does not require the inversion of a large dimensional unstructured correlation matrix.

3.2 *Choice of Smoothing Parameter*

Selection of the smoothing parameter is crucial in model fitting. It is desirable to have an automatic and data-driven method to select the smoothing parameter. Here we extend generalized cross-validation to the penalized quadratic inference function. Following the conventional technique of penalized least squares (e.g., Ruppert, 2002), we define the effective degrees of freedom as

$$\mathrm{df} = \mathrm{trace}\big\{(\ddot{Q}_N + N\lambda D)^{-1} \ddot{Q}_N\big\},$$

where $\ddot{Q}_N$ is the second derivative of Q_N with respect to γ. Thus, a generalized cross-validation statistic is defined as

$$\mathrm{GCV}(\lambda) = \frac{N^{-1} Q_N}{(1 - N^{-1}\mathrm{df})^2}, \tag{8}$$

and further $\hat{\lambda} = \mathrm{argmin}_{\lambda} \mathrm{GCV}(\lambda)$. In practice, the above minimization can be carried out by searching over a grid of λ values.

3.3 *Standard Error Formula*

We can derive the standard error formula for the resulting estimator using the sandwich formula,

$$\widehat{\mathrm{cov}}(\hat{\gamma}) = \{\ddot{Q}_N(\gamma) + N\lambda D\}^{-1} \left(\sum s_i s_i'\right) \{\ddot{Q}_N(\gamma) + N\lambda D\}^{-1},$$

which can be shown to be a consistent estimator of $\mathrm{cov}(\hat{\gamma})$ as given in Theorem 2. Denote $B_u(t) = [B_{u1}(t), \ldots, B_{uV_u}(t)]'$ and $\hat{\gamma}_u = [\hat{\gamma}_{u1}, \ldots, \hat{\gamma}_{uV_u}]'$. Then an estimator for $\beta_u(t)$ is $\hat{\beta}_u(t) = B_u'(t)\hat{\gamma}_u$ and its covariance can be estimated by $\widehat{\mathrm{cov}}\{\hat{\beta}_u(t)\} = B_u'(t)\widehat{\mathrm{cov}}(\hat{\gamma}_u)B_u(t)$. Ignoring the approximation error in (6), a $100(1-\alpha)\%$ pointwise confidence interval of $\beta_u(t)$ is given by

$$\hat{\beta}_u(t) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{cov}}\{\hat{\beta}_u(t)\}}, \tag{9}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution, since the resulting estimator follows an asymptotic normal distribution by Theorem 2.

4. Nonparametric Goodness-of-Fit Tests and Model Selection

In practice, parsimonious models are always desirable to enhance model predictability. It is of interest to test whether parsimonious parametric models can be used to approximate coefficient functions from the shape of their estimators. For varying-coefficient models, it is of particular interest to test whether some coefficients are time varying or time invariant. In other words, we are interested in testing

$$H_0 : \beta_u(\cdot) \equiv \beta_{u0} \quad \text{versus} \quad H_1 : \beta_u(\cdot) \neq \beta_{u0}, \tag{10}$$

for some u, where β_{u0} is an unknown constant. This problem can be handled using truncated power splines regression as in (6). Crainiceanu et al. (2005) consider a similar hypothesis testing problem for partial linear models with independent and identical observations, and propose a likelihood ratio test using a linear mixed-model representation of the penalized spline. We next propose a test statistic for (10) using the quadratic inference function. Our test statistic has an asymptotic chi-square distribution that is different from that of Crainiceanu et al. (2005).

Based on (6), we can test the following null hypothesis:

$$H_0 : \gamma_{uv} = 0, \quad v = 1, \ldots, V_u. \tag{11}$$

Let $\tilde{\gamma}$ denote the estimator under H_0 and $\hat{\gamma}$ be the estimator under H_1. Since the quadratic inference function plays a similar role to the least-square function, $Q(\tilde{\gamma})$ and $Q(\hat{\gamma})$ measure how well the model fits the data under H_0 and H_1, respectively. Intuitively, under H_0 the difference between $Q(\tilde{\gamma})$ and $Q(\hat{\gamma})$ should be very small. However, under $H_1, Q(\tilde{\gamma})$ should be systematically larger than $Q(\hat{\gamma})$. Thus, an appropriate test statistic to test H_0 against H_1 would be

$$T = Q(\tilde{\gamma}) - Q(\hat{\gamma}). \tag{12}$$

Alternatively, if we apply the penalized quadratic inference function, we may also consider

$$T_a = Q(\tilde{\gamma}) + N\lambda_N\tilde{\gamma}'D\tilde{\gamma} - Q(\hat{\gamma}) - N\lambda_N\hat{\gamma}'D\hat{\gamma}.$$

We next demonstrate that under certain regularity conditions and under H_0, T and T_a have the same limiting distribution, and both of them have a chi-squared distribution.

THEOREM 3: *Under Conditions A–E in the Appendix, if the smoothing parameter $\lambda_N = o(N^{-1/2})$, then T and T_a asymptotically follow chi-squared with degrees of freedom equal to V_u under the null hypothesis in* (11).

An important issue arises here as to how we decide whether the varying-coefficient function in (6) is adequately modeled. To assess whether there is a sufficient number of basis functions in (6) such that the model assumption $E(g) = 0$ is satisfied, where g is the estimating function in (2) for a single observation, we apply the goodness-of-fit test (Hansen, 1982). Namely,

$$Q(\hat{\gamma}) \xrightarrow{d} \chi^2_{r-k},$$

where $\hat{\gamma}$ is the estimator by minimizing the quadratic inference function when given basis functions are applied in the model, r is the dimension of $\bar{g}_N$ in (2), and k is the dimension of γ. This test can also be useful to determine the number of knots to be selected in (6), as too many knots in the model might overfit the data and degrade the performance of spline estimators (Ruppert, 2002). Note that the above goodness-of-fit test is only applicable where there are more estimating functions than unknown parameters, so it works for our situation as the dimension of estimating functions in (2) is greater than the dimension of parameters.

It is also possible that the goodness-of-fit tests fail to reject several different models. How can we assess which model is better? Note that most of these models are not nested; since different models likely have different knots, here we let knots be equally spaced. Andrews (1999) proposed a model selection or moment selection criterion in the generalized method of moments framework, which can also be applied here. The main idea is to penalize the objective function $Q(\hat{\gamma})$ for the difference of the numbers of estimating equations and parameters. For example, Andrews's selection criterion for a model with r estimating equations and k parameters is

$$Q(\hat{\gamma}) - (r-k)c_N, \tag{13}$$

where c_N is $\ln N$ for BIC and 2 for Akaike's information criterion (AIC), which are commonly used in traditional BIC and AIC model selection criteria. A model with a smaller value in (13) is better. The BIC in general is better than the AIC, as the latter is not consistent for different sample sizes. Intuitively, the penalty term associated with $r-k$ can be explained in that $Q(\hat{\gamma})$ follows χ^2_{r-k} asymptotically, and the mean of the chi-squared is its degrees of freedom. In our setting, if we choose $m+1$ basis matrices as in (1), then $r-k = (m+1)k - k = mk$.

For the multivariate model with several varying-coefficients, we could set the upper limit of q as 5 and knots as 20, where knots are evenly distributed in the range of time t (Ruppert, 2002). In practice, these numbers could be reduced to 10 for the number of knots and 3-degree polynomial basis functions. This is because, in principle, the choice of basis functions does not affect the fit very much, although some basis functions are more numerically stable with simpler computation (Ruppert, Wand, and Carroll, 2003, p. 69). After we determine these, we could choose the optimal combination of knots and basis functions where the BIC is the minimum. The selected model with the minimum BIC provides the best fit for the data within a class where the upper limit number of knots and the degree of polynomial basis functions are determined. The model selection procedure could be computationally intensive (Ruppert et al., 2003, p. 64), as it is possible that the

total number of combinations is very high. Wand (2000) provides a review and comparison for some recent model selection approaches.

It is important to point out that model selection should be done at the beginning stage, since hypothesis testing for whether the coefficient is time varying or not depends on how we choose the full model. If the full models under the alternative are different, then the test statistics and corresponding p-values could be different, although it might not affect the statistical significance of our tests dramatically.

5. Simulation and Application

In this section, we assess the finite sample performance of the proposed procedures in Sections 2 and 3 with Monte Carlo simulation studies. We also demonstrate the proposed method with an analysis of an AIDS data set. The programming codes for the simulations and the real data example are available upon request.

5.1 *Simulation Studies*

Example 1 (binary response). We generate 200 subjects for each simulation. Each subject is supposed to have 31 repeated measurements at centered scheduled time points $\{-15, -14, \ldots, 15\}$, but in reality each subject has a 60% chance of missing the scheduled time except at the beginning time. The true time also varies around the unskipped schedule time following the uniform $(-0.5, 0.5)$ distribution. The response variable y_{ij} has the marginal distribution

$$P(y_{ij} = 1 \,|\, t_{ij}) = \exp\{\beta(t_{ij})\}/[1 + \exp\{\beta(t_{ij})\}],$$

where $i = 1, \ldots, 200$ and $j = 1, \ldots, n_i$. We simplify this simulation by assuming that covariates are constant 1. We model four kinds of varying-coefficients $\beta(t)$ as follows:

$$\beta_0(t) = \sin\left(\frac{(t+15)\pi}{60}\right) - 0.5, \quad \beta_1(t) = \cos\left\{\frac{(t-10)\pi}{15}\right\}$$

$$\beta_2(t) = -0.1(t-15), \quad \beta_3(t) = \frac{(5-t)^3}{4000} - 1.$$

To create correlated responses, we apply the algorithm following Park, Park, and Shin (1996) under exchangeable correlation structure with the correlation parameter as 0.5.

We first perform the goodness-of-fit test in Section 4 to select the degree of truncated power polynomial splines and the number of knots that are evenly distributed over the ranges of t_{ij}. The results are summarized as follows. For the basis function in (6), the goodness-of-fit test based on the first few simulations yields $q = 3$ for $\beta_0(t)$, $\beta_1(t)$, and $\beta_3(t)$, and $q = 2$ for $\beta_2(t)$; for number of knots, the test also chooses 5 knots for $\beta_1(t)$, and no knots for $\beta_0(t)$, $\beta_2(t)$, and $\beta_3(t)$. With the selected degrees and number of knots, we calculate the quadratic inference function estimators by minimizing (7). Here, we assume exchangeable working correlation for estimating equations in (2), that is, the basis matrix M_1 is 0 on the diagonal and 1 off-diagonal. The result by assuming AR(1) working correlation is similar to that under exchangeable correlation, and is not presented here. We define the mean absolute deviation of errors by minimizing (7)

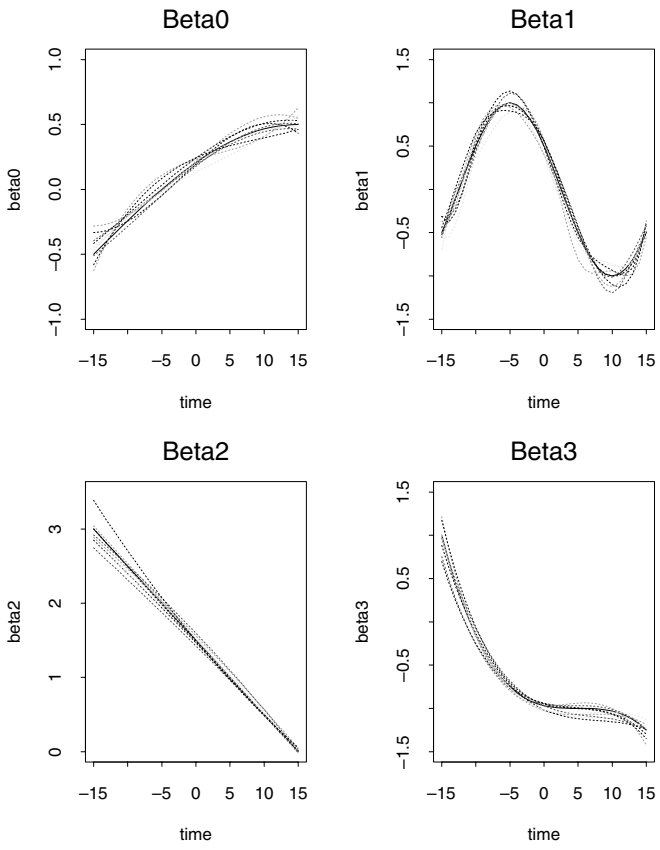

Figure 1. For binary responses, fitted varying-coefficient curves corresponding to nine deciles of mean absolute deviation of errors, from 1000 simulations. The solid lines are true coefficient curves.

$$\text{MADE} = \sum_{j=0}^{30} 31^{-1} |\hat{\beta}(t_j) - \beta(t_j)| / \text{range}(\beta),$$

where $t_j = (-15, \ldots, 15)$. Figure 1 provides fitted varying-coefficient curves corresponding to nine deciles of mean absolute deviation of errors from 1000 simulations. Figure 1 demonstrates that the quadratic inference function approach applying the penalized spline works well in settings where coefficients have linear, cubic, sine, and cosine relationships of time.

We apply the test result in (12) to illustrate how the quadratic inference function performs for testing whether coefficients vary over time in finite samples. We simulate data such that $\beta_1(t) = 0.5$. The null hypothesis $H_0 : \beta_1$ is constant over time. We let the basis functions for $\beta_1(t)$ under H_0 be 1 and basis functions for $\beta_1(t)$ under the alternative be 1, t, t^2, t^3, $(t+10)_+^3$, $(t)_+^3$, $(t-10)_+^3$. We calculate $\tilde{\gamma}_1$ and $\hat{\gamma}_1$ by minimizing (4) under H_0 and H_1, where $\bar{g}_N$ is constructed by assuming either exchangeable or AR(1) working correlation structures. Since the difference of the numbers for the basis functions under H_1 and H_0 is 6, the test statistic $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ asymptotically follows χ^2_6. Figure 2 provides quantile–quantile plots under both exchangeable and AR(1) working correlations and illustrates that under H_0 the

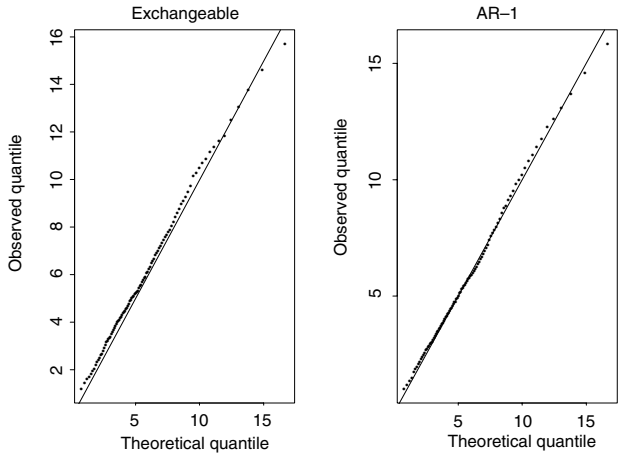

Figure 2. For binary responses, quantile–quantile plots for test statistics $Q(\tilde{\beta}) - Q(\hat{\beta})$ versus χ^2_6 under $H_0 : \beta_1$ is constant over time, from 1000 simulations: assume exchangeable working correlation; assume AR(1) working correlation.

empirical quantiles of $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ follow the theoretical chi-squared quantile rather well.

We next examine the power of the quadratic inference function approach when $\beta_1(t)$ deviates from the constant. Let

$$\beta_1(t, \eta) = 0.5 - \eta * \cos\left\{\frac{(t-10)\pi}{15}\right\},$$

where $0 \le \eta \le 1$. We calculate test statistics $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ from 1000 simulations for various η, and find the percentage of test statistics greater to or equal to 12.59, the 95% quantile of χ^2_6. Figure 3 illustrates the power function curve. Note that when η is close to 0, the test size is approximately 0.05; and when η reaches 0.25, the probability of rejection reaches 1.

Example 2 (continuous response). In this example, we generate 200 subjects, and each subject is supposed to have repeated measurements at scheduled time points $\{0, 1, \ldots, 30\}$, but has a 60% chance of missing the scheduled time except at time 0. The true time also varies around the unskipped schedule time following the uniform $(-0.5, 0.5)$ distribution. The response variable y_{ij} is modeled as

$$y_{ij} = \beta_0(t_{ij}) + \sum_{k=1}^{3} X_i^{(k)}(t_{ij})\beta_k(t_{ij}) + \varepsilon_i(t_{ij}),$$

where $i = 1, \ldots, 200$ and $j = 1, \ldots, n_i$. The time-varying-coefficients satisfy

$$\beta_0(t) = 15 + 20\sin\left(\frac{t\pi}{60}\right), \quad \beta_1(t) = 2 - 3\cos\left\{\frac{(t-25)\pi}{15}\right\},$$

$$\beta_2(t) = 6 - 0.2t, \quad \beta_3(t) = -4 + \frac{(20-t)^3}{1000}.$$

The covariates are generated as follows: $X_i^{(1)}$ has a uniform $(t/10,\ 2 + t/10)$ distribution; conditioning on $X_i^{(1)}$, $X_i^{(2)}$ has a normal distribution with mean 0 and variance $(1 + X_i^{(1)})/(2 + X_i^{(1)})$; and $X_i^{(3)}$ has a Bernoulli(0.6) distribution. The error ε_i follows a multivariate normal distribution with mean 0 and the marginal variance matrix $2I$, I is an identity matrix, and the correlation is exchangeable with correlation 0.8.

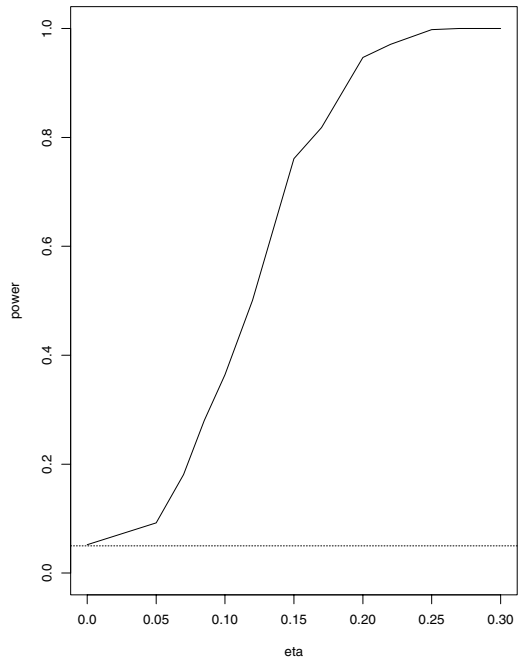

Figure 3. For binary responses, power of quadratic inference function against η for testing $H_0 : \beta_1$ is constant over time, from 1000 simulations.

We select the degrees q and the number of knots using the goodness-of-fit test based on the first few simulations. The test provides $q = 3$ for all coefficients, 5 knots for $\beta_1(t)$, and no knots for $\beta_0(t)$, $\beta_2(t)$, and $\beta_3(t)$. With the selected degrees and the number of knots, we calculate the quadratic inference function estimators by minimizing (7) with exchangeable working correlation structure.

As in Example 1, Figure 4 provides fitted varying-coefficient curves corresponding to nine deciles of mean absolute deviation of errors from 1000 simulations. Figure 4 clearly shows that the quadratic inference function approach applying the penalized spline works well in settings where coefficients have linear, cubic, sine, and cosine relationships of time.

We also use this simulation to illustrate how different working correlations could affect our estimations. The true correlation is exchangeable here. We calculate the quadratic inference function estimators by minimizing (7) using AR(1) working correlation and independent structures in addition to the above exchangeable correlation structure and we calculate the median of the absolute deviation of errors between the fitted and true values of β_0, β_1, β_2, and β_3, and plot them against time $t = 0, 1, \ldots, 30$. Figure 5 shows that the estimation based on independent structure has the largest median of the absolute deviation of errors for every time point. Estimations using AR(1) and exchangeable have similar errors, though using the true exchangeable correlation structure

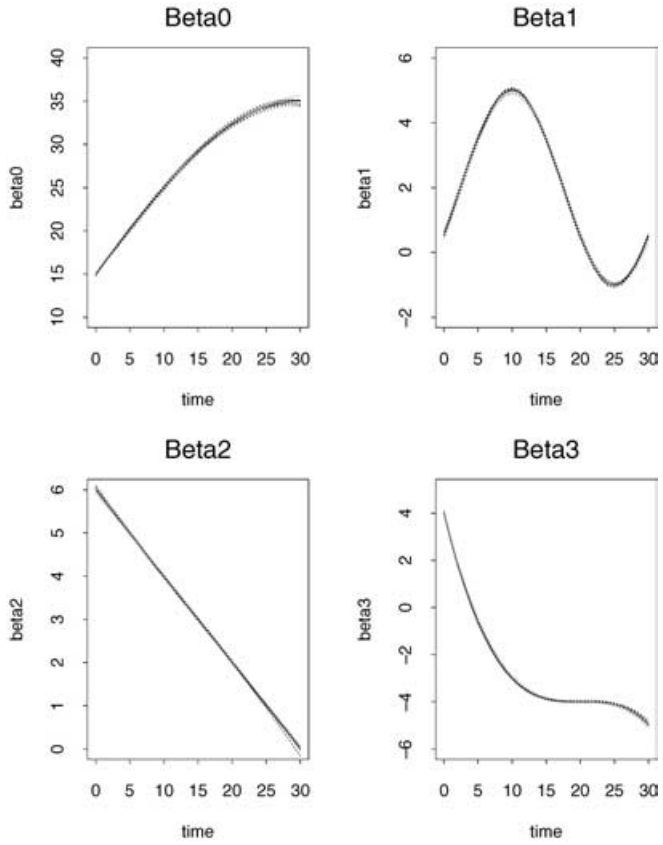

Figure 4. For continuous responses, fitted varying-coefficient curves corresponding to nine deciles of mean absolute deviation of errors from 1000 simulations. The solid lines are true coefficient curves.

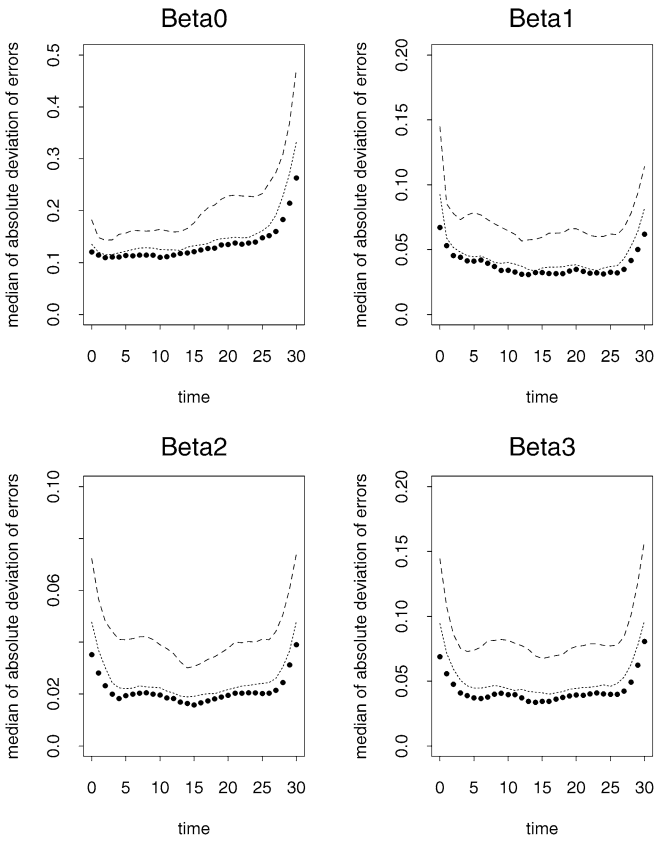

Figure 5. For continuous responses, the median of absolute deviation of errors between fitted and true values from 1000 simulations for three different working correlation structures. The bold dotted line is from exchangeable correlation, the dotted line is from AR(1) correlation, and the dashed line is from independent structure.

produces the smallest median of the absolute deviation of errors. If we take the weighted average of the median of the absolute deviation of errors for four coefficients, where weights are $1/\text{range}(\beta_i)$, $i = 0, 1, 2, 3$, then the ratio of the weighted average of errors between the exchangeable (true) and AR(1) is 0.87, and 0.55 between the exchangeable and independent structures. This illustrates that using misspecified AR(1) will lose some efficiency, but is still better than when assuming independence. However, this would be a different case for GEE if the true exchangeable correlation were misspecified as the AR(1), since GEE requires one to estimate the correlation ρ for misspecified AR(1), and the estimator of ρ may not be valid (Crowder, 1995).

To illustrate how the quadratic inference function performs for testing whether coefficients vary over time in finite samples, we simulate the same data as above, except let $\beta_1(t) = 2$ such that under the null hypothesis $H_0 : \beta_1$ is constant over time. We let the basis functions for β_1 under H_0 be 1, and the basis functions for $\beta_1(t)$ under the alternative be $1, t, t^2, t^3, (t-5)_+^3, (t-10)_+^3, (t-15)_+^3, (t-20)_+^3, (t-25)_+^3$. We calculate $\tilde{\gamma}_1 = \tilde{\gamma}_{10}$ and $\hat{\gamma}_1 = (\hat{\gamma}_{10}, \hat{\gamma}_{11}, \ldots, \hat{\gamma}_{18})$ by minimizing (4), where $\bar{g}_N$ is constructed by assuming either exchangeable or AR(1) working correlation structures. Since the difference of the numbers for the basis functions under H_1 and H_0 is 8, the test statistic $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ asymptotically follows χ_8^2. Figure 6 provides quantile–quantile plots under the exchangeable and AR(1) working correlations and illustrates that under H_0 the empirical quantiles of $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ follow the theoretical chi-squared quantile rather well.

We also examine the power of the quadratic inference function approach when $\beta_1(t)$ deviates from the constant. Let

$$\beta_1(t, \eta) = 2 - \eta * 3 \cos\left\{\frac{(t-25)\pi}{15}\right\},$$

where $0 \le \eta \le 1$. We calculate test statistics $Q(\tilde{\gamma}) - Q(\hat{\gamma})$ from 1000 simulations for various η using exchangeable, AR(1) correlation, and independent structures, and find the percentage of test statistics greater than or equal to 15.51, the 95% quantile of χ_8^2. Figure 7 illustrates the power function curves for three correlation structures. Note that when η is close to 0, the test sizes are all approximately 0.05; and when η reaches 0.06, the probabilities of rejection reach 1. In addition, the power function using true exchangeable correlation structure is much higher than the power using AR(1) and independent structures when $\eta < 0.06$. The test power using AR(1) and independent structures are close; however, the test power

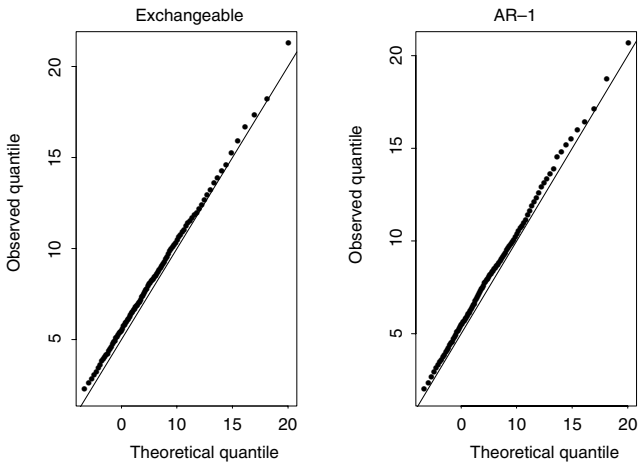

Figure 6. For continuous responses, quantile–quantile plots for test statistics $Q(\tilde{\beta}) - Q(\hat{\beta})$ versus χ^2_8 under $H_0 : \beta_1$ is constant over time from 1000 simulations: assume exchangeable working correlation; assume AR(1) working correlation.

assuming independent structure is the worst. This might be explained in that when the true correlation in the exchangeable structure is very high (0.8), the performance of tests under the misspecified AR(1) (or independent) and true exchangeable correlation structures may be very different.

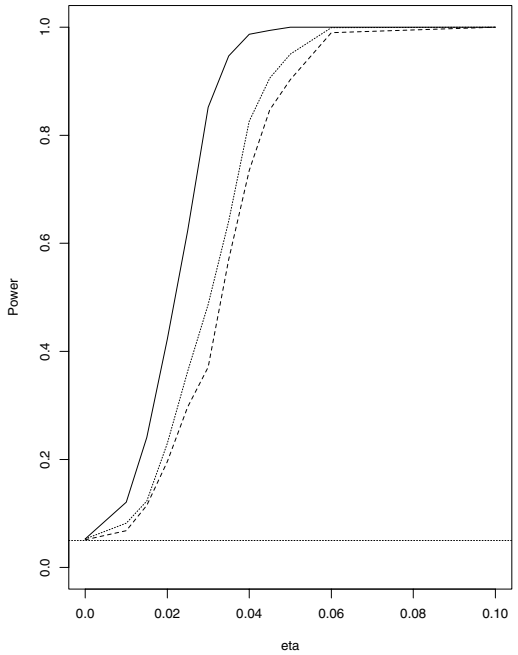

Figure 7. For continuous responses, power of quadratic inference function against η for testing $H_0 : \beta_1$ is constant over time from 1000 simulations. The solid line is the power using exchangeable correlation, the dotted line is the power using AR(1) correlation, and the dashed line is the power using independent structure. The true correlation structure is exchangeable.

5.2 *Application to AIDS Data*

We apply AIDS data (Huang et al., 2002) to illustrate how the penalized quadratic inference function approach works for real data. This data set consists of 283 homosexual males who were HIV positive between 1984 and 1991. Each patient was supposed to have measurements taken every 6 months, but it often happened that patients missed or rescheduled their appointments. Therefore, each patient had a different number of repeated measurements and the true observation times were not equally spaced. Here time t is defined as the time (in years) since subjects had their visits after their HIV infection. We observe that each patient has minimum 1 and maximum 14 measurements for these data. It is known that HIV destroys CD4 cells, so by measuring CD4 cell counts and percentages in the blood, doctors are able to monitor progression of the disease. The response variable is the CD4 percentage over time. Three covariates were also collected: patient's age, smoking status with 1 as smoker and 0 as nonsmoker, and the CD4 cell percentage before their infection. We model it as

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{Smoke} + \beta_2(t_{ij})\text{Age} + \beta_3(t_{ij})\text{Pre-CD4} + \varepsilon_{ij}.$$

We perform the goodness-of-fit test and choose $q = 3$ in the basis functions of (6) for all coefficients, and the number of knots to be 0, 5, 1, and 3 for $\beta_u(t)$, $u = 0, 1, 2, 3$, respectively. These choices are made to facilitate comparison with Huang et al.'s (2002) paper.

Figure 8 provides fitted curves (the solid line) for coefficients of intercept, smoking, age, and pre-infection CD4 effects. We also provide 95% pointwise confidence intervals (dotted lines) for four varying-coefficients. Figure 8 also implies that the intercept decreases over time, though the rate of decreasing drops; smoking and age effects are not significant; and the pre-infection CD4 percentage appears to have a positive relationship with the post-infection CD4 percentage. These findings are consistent with Wu and Chiang (2000), Fan and Zhang (2000), and Huang et al. (2002).

To quantify these findings, we apply the quadratic inference function test in (12) to test whether smoking and age effects are statistically significant. For intercept and pre-infection CD4 effects, we also test whether they are time invariant or not. We fit the model with a 3-degree polynomial function of t for all varying-coefficients, a total of 25 parameters with 4 (0 knots) for intercept, 9 (5 knots) for smoking, 5 (1 knot) for age, and 7 (3 knots) for pre-infection CD4. We assume an exchangeable working correlation, therefore the number of estimating equations $r = 25 \times 2 = 50$. The goodness-of-fit test statistic $Q(\tilde{\gamma}) = 23.1$. By Section 4, with $r - k = 25$ degrees of freedom, the asymptotic p-value from the chi-squared test is 0.57. This indicates that the model fits the data reasonably well based on the asymptotic results in Section 4. For testing whether smoking is significant, we set all 9 parameters for the smoking coefficient to be zero, and estimate the rest of the 16 parameters by minimizing the quadratic inference function in (7). The test statistic under H_0 is 36.1, the difference between the two test statistics under H_1 and H_0 is then 13.0, and the corresponding p-value is 0.163 based on chi-squared with $25 - 16 = 9$ degrees of freedom. Similarly, we obtain the p-value for age as 0.172. None of these are statistically significant.

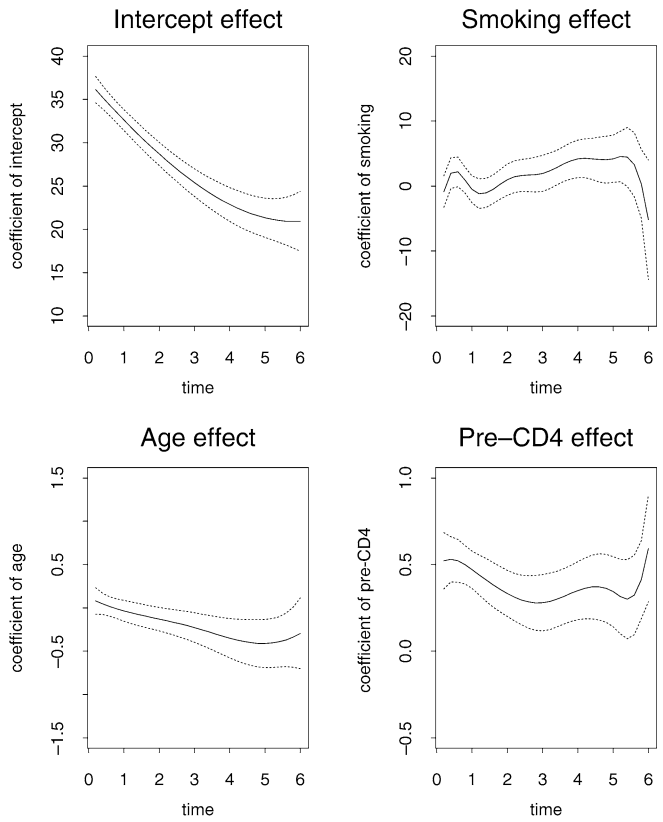

Figure 8. Fitted varying-coefficients for AIDS data, where solid lines are fitted curves and dotted lines are 95% pointwise confidence intervals.

Table 1
Hypothesis testing for AIDS data, comparisons between Huang et al.'s (2002) *bootstrap approach and the quadratic inference function approach*

	Bootstrap (Independent)	Quadratic inference function (Exchangeable)		
Null hypothesis	p-value	T	d.f.	p-value
Constant baseline	0.000	81.9	3	0.000
Smoking has no effect	0.176	13.0	9	0.163
Age has no effect	0.301	7.7	5	0.172
Constant pre-CD4	0.059	12.9	6	0.045*

* Statistically significant at a 0.05 nominal level.
T: test statistic defined in (12).
d.f.: degrees of freedom.

To test whether the intercept effect is time varying, we set the three parameters associated with time t to be zero (the intercept coefficient has no knots, so there is a total of four parameters). The test statistic under H_0 is 105.0, the difference between the test statistics under H_0 and H_1 is 81.9, and the corresponding p-value is close to zero based on the χ^2_3 test, where the degrees of freedom are calculated by $25 - 22 = 3$. Similarly, we find the p-value for constant pre-infection CD4 effects is 0.045, which is statistically significant. The final result for pre-infection CD4 effects differs from that in Huang et al.'s (2002) bootstrap approach, where their bootstrap approximate p-value is 0.059. Examination of Figure 8 on pre-infection CD4 effects favors our conclusion. Table 1 provides comparisons of p-values based on Huang et al.'s (2002) approach and our method. From Table 1, our p-values calculated from quadratic inference function test statistics are consistently smaller than Huang et al.'s results, and this might be explained in that the quadratic inference function test takes into account the correlation information and thus may be more powerful.

6. Discussion

We propose nonparametric modeling using the penalized quadratic inference function to incorporate correlation for longitudinal data. Our approach works for continuous and discrete cases as it only requires correct specification of the mean structure that is modeled nonparametrically, and we do not require any likelihood or approximation of the likelihood function for estimation and hypothesis testing such as in the generalized linear mixed model approach. The quadratic inference function approach is relatively simple and numerically feasible since it does not involve any nuisance parameters associated with the working correlation. This advantage becomes more important in nonparametric settings as there are many more parameters involved in nonparametric modeling than in parametric or semiparametric modeling. Existing smoothing spline methods are focused mainly on correlated continuous responses. For correlated discrete data, existing nonparametric approaches either ignore correlation and treat data as independent, or could be computationally intensive and complex. For example, the generalized linear mixed effects model typically has a parametric model assumption of random effect, and it requires numerical approximation to compute high-dimensional likelihood functions even under the assumption that the random effects are normal. Estimation and hypothesis testing for variance components in generalized linear mixed effects are even more complicated if the random effects are not normal.

Another advantage of the quadratic inference function approach is that the inference function has an explicit asymptotic form, which allows us to test whether coefficients are time varying or time invariant for varying-coefficient models. Further, it also enables us to do goodness-of-fit tests for checking model assumptions, and provides an objective criterion for choosing a sufficient number of basis functions and knots for varying-coefficients. An important issue that arises here is how to select the basis functions and the number of knots when candidate models are not nested to each other. This could be done using cross-validation and Mallows's (1973) C_p criterion if the response is continuous normal (e.g., Ruppert et al., 2003). In the setting where there are more estimating equations than parameters, such as in our case, we apply Andrews's (1999) generalized method of moments BIC criterion for selecting nonparametric basis functions and knots. This model selection criterion can be applied for both continuous and discrete responses.

Acknowledgements

The authors are grateful for many constructive suggestions from two referees, the associate editor, and the co-editor. Qu's research was supported in part by NSF grant DMS 0348764. Li's research was supported by National Institute on Drug Abuse (NIDA) grant 2-P50-DA10075.

References

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* **67,** 543–564.

Billingsley, P. (1995). *Probability and Measure*, 3rd edition. New York: John Wiley & Sons.

Chiang, C.-T., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying-coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association* **96,** 605–619.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B* **66,** 165–185.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92,** 91–103.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82,** 407–410.

Fan, J. and Zhang, J.-T. (2000). Functional linear models for longitudinal data. *Journal of the Royal Statistical Society, Series B* **62,** 303–322.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **31,** 1208–1212.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika* **92,** 105–118.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50,** 1029–1054.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757–776.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. New York: Springer-Verlag.

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

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89,** 111–128.

Jarrow, R., Ruppert, D., and Yu, Y. (2004). Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association* **99,** 57–66.

Kim, I., Cohen, N., and Carroll, R. J. (2003). Semiparametric regression splines in matched case-control studies. *Biometrics* **59,** 1158–1169.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73,** 12–22.

Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* **95,** 520–534.

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

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M., and Ibrahim, J. (2000). GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* **56,** 528–536.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15,** 661–675.

Martinussen, T. and Scheike, T. H. (2001). Sampling adjusted analysis of dynamic additive regression models for longitudinal data. *Scandinavian Journal of Statistics* **28,** 303–323.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16,** 134–153.

Park, C. G., Park, T., and Shin, D. W. (1996). A simple method for generating correlated binary variates. *American Statistician* **60,** 306–310.

Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: John Wiley & Sons.

Qu, A. and Lindsay, B. G. (2003). Building adaptive estimating equations when inverse-of-covariance estimation is difficult. *Journal of the Royal Statistical Society, Series B* **65,** 127–142.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87,** 823–836.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11,** 735–757.

Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42,** 205–223.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Wand, M. P. (1999). On the optimal amount of smoothing in penalized spline regression. *Biometrika* **86,** 936–940.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics* **15,** 443–462.

Wang, Y. (1998a). Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B* **60,** 159–174.

Wang, Y. (1998b). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association* **93,** 341–348.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90,** 43–52.

Wang, N., Carroll, R., and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/cluster data. *Journal of the American Statistical Association* **100,** 147–157.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* **61,** 439–447.

Wu, C. O. and Chiang, C.-T. (2000). Kernel smoothing on varying-coefficient models with longitudinal dependent variable. *Statistica Sinica* **10,** 433–456.

Wu, C. O., Chiang, C.-T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* **93,** 1388–1402.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97,** 1042–1054.

Yu, Y. and Ruppert, D. (2004). Root-n consistency of penalized spline estimator for partially linear single-index models under general Euclidean space. *Statistica Sinica* **14,** 449–456.

Zhang, D. (2004). Generalized linear mixed models with varying-coefficient for longitudinal data. *Biometrics* **60,** 8–15.

Received October 2004. *Revised July* 2005.
Accepted August 2005.

Appendix

Regularity Conditions

To establish the root n consistency and asymptotic normality for the penalized quadratic inference function estimator, we need the following regularity conditions. Similar conditions are also given in Hansen (1982).

A. The weighting matrix $\bar{C}_N = N^{-1}\sum g_i g_i'$ converges almost surely to a constant matrix C_0, where C_0 is invertible. This condition holds based on the weak law of large numbers when N goes to infinity, and the maximum cluster size is fixed.

B. The spline regression parameter γ is identified, that is, there is a unique $\gamma_0 \in S$ satisfying the mean zero model assumption

$$E\{g_1(\gamma_0)\} = 0, \tag{A.1}$$

where S is the parameter space.

C. The parameter space S is compact, and γ_0 is an interior point of S.

D. We require that $E\{g(\gamma)\}$ is continuous in γ.

E. The first derivative of $\bar{g}_N$ exists and is continuous, and $\frac{\partial \bar{g}_N}{\partial \gamma}(\hat{\gamma})$ converges in probability to $J_0 = E[\frac{\partial g}{\partial \gamma}(\gamma_0)]$ when $\hat{\gamma}$ converges in probability to γ_0.

Proof of Theorem 1. First, the estimator $\hat{\gamma}$ by minimizing (7) exists since (7) has 0 as a lower bound and the global minimum exists. Second, we will show that it is impossible that $\hat{\gamma}$ remains outside of U, where U is any neighborhood of γ_0. First, U^c is compact and $|C_0^{-1/2}E[g(\gamma)]|^2$ is continuous, where $|\cdot|^2$ is the inner product of a vector. Following Conditions C and D, $|C_0^{-1/2}E[g(\gamma)]|^2$ achieves a minimum in U^c, and let γ^* be its minimum in U^c. Second, using identification of γ from Condition B, we have $|C_0^{-1/2}E[g(\gamma^*)]|^2 > 0$. Hence, if we show $|C_0^{-1/2}E[g(\hat{\gamma})]|^2 \to_{a.s.} 0$, then eventually $\hat{\gamma} \in U$.

Since $\hat{\gamma}$ is the minimizer of (7), then

$$\left|\bar{C}_N^{-1/2}\frac{1}{N}\sum_{i=1}^N g_i(\hat{\gamma})\right|^2 + \lambda_N \hat{\gamma}' D\hat{\gamma} \leq \left|\bar{C}_N^{-1/2}\frac{1}{N}\sum_{i=1}^N g_i(\gamma_0)\right|^2 + \lambda_N \gamma_0' D\gamma_0. \tag{A.2}$$

The right-hand side of (A.2) converges to zero almost surely since $\lambda_N = o(1)$, and by the strong law of large numbers and Condition A, that is,

$$\left|\bar{C}_N^{-1/2}\frac{1}{N}\sum_{i=1}^N g_i(\gamma_0)\right|^2 \to_{a.s.} \left|C_0^{-1/2}E[g(\gamma_0)]\right|^2 = 0.$$

Since S is compact, we apply the uniform law of large numbers or Glivenko–Cantelli theorem (Billingsley, 1995, p. 269),

$$\sup_{b\in S}\left|\frac{1}{N}\sum_{i=1}^N g_i(b) - E[g(b)]\right| \to_{a.s.} 0.$$

Then by Condition A and the continuity mapping theorem,

$$\left|\bar{C}_N^{-1/2}\frac{1}{N}\sum_{i=1}^N g_i(\hat{\gamma}) - C_0^{-1/2}E[g(\hat{\gamma})]\right| \to_{a.s.} 0.$$

From (A.2), it follows that

$$\left|C_0^{-1/2}E[g(\hat{\gamma})]\right|^2 \to_{a.s.} 0,$$

as was to be shown.

Proof of Theorem 2. We denote $\dot{Q}$ and $\ddot{Q}$ to be the first and second derivatives of the quadratic inference function Q with respect to γ. Using the Central Limit Theorem,

$$\sqrt{N}\bar{g}_N(\beta_0) \xrightarrow{d} N_r(0, C_0). \tag{A.3}$$

Since $\hat{\gamma}$ is obtained by minimizing (7), $\hat{\gamma}$ satisfies

$$N^{-1}\dot{Q}(\hat{\gamma}) + 2\lambda_N D\hat{\gamma} = 0. \tag{A.4}$$

By Taylor's expansion,

$$\begin{aligned} 0 &= N^{-1}\dot{Q}(\hat{\gamma}) + 2\lambda_N D\hat{\gamma} \\ &= N^{-1}\dot{Q}(\gamma_0) + 2\lambda_N D\gamma_0 + \left\{N^{-1}\ddot{Q}(\tilde{\gamma}) + 2\lambda_N D\right\}(\hat{\gamma} - \gamma_0), \end{aligned}$$

where $\tilde{\gamma}$ is between γ_0 and $\hat{\gamma}$. Therefore,

$$\hat{\gamma} - \gamma_0 = -\left\{N^{-1}\ddot{Q}(\tilde{\gamma}) + 2\lambda_N D\right\}^{-1}\left\{N^{-1}\dot{Q}(\gamma_0) + 2\lambda_N D\gamma_0\right\}. \tag{A.5}$$

Note that $N^{-1}\ddot{Q}(\tilde{\gamma})$ converges in probability to $2J_0'(\gamma_0)\times C_0^{-1}(\gamma_0)J_0(\gamma_0) = O_p(1)$ since $\tilde{\gamma}$ is between γ_0 and $\hat{\gamma}$, where $\hat{\gamma}$ converges to γ_0 in probability by Theorem 1, and by Conditions A and E. In addition, using the fact that $2\lambda_N D = o(N^{-1/2})$, therefore $\{N^{-1}\ddot{Q}(\tilde{\gamma}) + 2\lambda_N D\}^{-1} = 2(J_0'C_0^{-1}J_0)^{-1} + o_p(N^{-1/2})$. Similarly, $N^{-1}\dot{Q}(\gamma_0) + 2\lambda_N D\gamma_0 = J_0'C_0^{-1}\bar{g}_N(\gamma_0) + o(N^{-1/2})$. Therefore equation (A.5) becomes

$$\sqrt{N}(\hat{\gamma} - \gamma_0) = -\sqrt{N}\big(J_0'C_0^{-1}J_0\big)^{-1}\big(J_0'C_0^{-1}\bar{g}_N\big) + o_p(1).$$

Then, using (A.3),

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{d} N_p\big(0, \big(J_0'C_0^{-1}J_0\big)^{-1}\big).$$

Next we will show that $(J_0'C_0^{-1}J_0)^{-1}$ reaches minimum in Löwner ordering. Suppose we minimize $\bar{g}_N'C^{-1}\bar{g}_N$, where C is any arbitrary symmetric invertible matrix. Again, using the above arguments, the asymptotic $\mathrm{var}(N^{1/2}\hat{\gamma}) = (J_0'C^{-1}J_0)^{-1}(J_0'C^{-1}C_0C^{-1}J_0)(J_0'C^{-1}J_0)^{-1}$.

Let $D = (J_0'C^{-1}J_0)^{-1}J_0'C^{-1}C_0^{1/2} - (J_0'C_0^{-1}J_0)^{-1}J_0'C_0^{-1/2}$,

$$\begin{aligned}
DD' = {} & \big(J_0'C^{-1}J_0\big)^{-1}J_0'C^{-1}C_0C^{-1}J_0\big(J_0'C^{-1}J_0\big)^{-1} \\
& - \big(J_0'C^{-1}J_0\big)^{-1}J_0'C^{-1}J_0\big(J_0'C_0^{-1}J_0\big)^{-1} \\
& - \big(J_0'C_0^{-1}J_0\big)^{-1}J_0'C^{-1}J_0\big(J_0'C^{-1}J_0\big)^{-1} \\
& + \big(J_0'C_0^{-1}J_0\big)^{-1}J_0'C_0^{-1}J_0\big(J_0'C_0^{-1}J_0\big)^{-1} \\
= {} & \big(J_0'C^{-1}J_0\big)^{-1}J_0'C^{-1}C_0C^{-1}J_0\big(J_0'C^{-1}J_0\big)^{-1} - \big(J_0'C_0^{-1}J_0\big)^{-1}.
\end{aligned}$$

Since DD' is a nonnegative definite matrix, $(J_0'C^{-1}J_0)^{-1}\times J_0'C^{-1}C_0C^{-1}J_0(J_0'C^{-1}J_0)^{-1} \geq (J_0'C_0^{-1}J_0)^{-1}$ (where $\geq$ stands for Löwner ordering), the equality holds if and only if $D = 0$, and this occurs when $C = C_0$, the true variance of g_i when $E(g_i) = 0$. The efficiency proof here is a standard result as in estimating function theory (Godambe, 1960) and the generalized method of moments (Hansen, 1982).

Proof of Theorem 3. This proof is similar to the proof of Theorem 1 in Qu et al. (2000).