Partially Linear Single Index Models for Repeated Measurements [1]

Shujie Ma, Hua Liang, and Chih-Ling Tsai

Abstract

In this article, we study the estimations of partially linear single-index models (PLSiM) with repeated measurements. Specifically, we approximate the nonparametric function by the polynomial spline, and then employ the quadratic inference function (QIF) together with profile principle to derive the QIF-based estimators for the linear coefficients. The asymptotic normality of the resulting linear coefficient estimators and the optimal convergence rate of the nonparametric function estimate are established. In addition, we employ a penalized procedure to simultaneously select significant variables and estimate unknown parameters. The resulting penalized QIF estimators are shown to have the oracle property, and Monte Carlo studies support this finding. An empirical example is also presented to illustrate the usefulness of penalized estimators.

KEY WORDS: Consistency, model selection, oracle estimator, polynomial spline, profile principle, quadratic inference function, SCAD

Short Title: PLSiM for repeated measurements

[1]Corresponding Author: Shujie Ma (E-mail: shujie.ma@ucr.edu). Shujie Ma (E-mail: shujie.ma@ucr.edu) is Assistant Professor, Department of Statistics, University of California, Riverside, CA 92521. Chih-Ling Tsai (E-mail: cltsai@ucdavis.edu) is Professor, Graduate School of Management, University of California, Davis, CA 95616. Hua Liang (E-mail:hliang@gwu.edu) is Professor, Department of Statistics, The George Washington University, 801 22nd St. NW, Washington, D.C. 20052. Ma's research was partially supported by NSF grant DMS 1306972 and UCR Regents Faculty Fellowship. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444.

1 Introduction

In regression analysis with repeated measures over time (i.e., longitudinal/panel data or cluster data), semiparametric models have been used to take into account both linear and nonlinear effects of covariates. The pioneering work in this context can be traced back to Severini and Staniswalis (1994), followed by a series of efforts, such as Chen et al. (2012); Chen and Jin (2006); Fan and Li (2004); He et al. (2005); Lin and Carroll (2001, 2006); and Su and Ullah (2006). All of these works focus on the case in which the models either contain one nonparametric component or a summand of nonparametric functions. The typical approach for estimating parameters in these models is kernel-based backfitting and local scoring, proposed by Buja et al. (1989). Although those models are very flexible, they have some limitations. For example, the model with one multivariate nonparametric function suffers from the "curse of dimensionality" when the number of covariates is moderate or large, while the model with a summand of nonparametric functions does not take into account the interaction effects among covariates. In addition, the backfitting estimation algorithm can become computationally expensive when the number of covariates is large. This is because it requires extensive iterations to update the estimator of each nonparametric component (Härdle et al.; 2004; Yu et al.; 2008). This motivates us to adapt partially linear single-index models (PLSiM) to analyze repeatedly measured data. Semiparametric single-index models have been widely used as an appealing and effective statistical tool to model the relationship between the response variable and multivariate covariates, since it achieves dimension reduction and relaxes the restrictive parametric assumptions. See Horowitz (1998) for detailed discussion and illustration of the usefulness of this model. The PLSiM as a natural extension allows discrete explanatory variables to be modeled in the linear part. See Carroll et al. (1997); Chen et al. (2013); Liang and Wang (2005); Xia and Li (1999) for studies and applications of PLSiM.

To estimate parameters in PLSiM, various methods have been proposed, including the backfitting algorithm (Carroll et al.; 1997), the penalized spline (Yu and Ruppert; 2002), the average derivative estimation method (ADE, Powell et al.; 1989), the minimum average variance estimation (MAVE, Xia and Härdle; 2006; Xia et al.; 1999), the profile least squares approach (PrLS, Ichimura; 1993; Liang et al.; 2010), and the estimating function method (EFM, Cui et al.; 2011). It is noteworthy that the backfitting algorithm can be unstable and penalized spline estimation may not be efficient. Although the ADE, MAVE and profile methods overcome these limitations, ADE may suffer from the curse of dimensionality as mentioned in Xia (2006), MAVE may encounter the sparseness problem as noted by Cui et al. (2011), and the PrLS estimator is not easy to obtain due to minimizing a high-dimensional nonlinear objective function. In addition, the above methods mainly focus on cross-sectional data rather than repeatedly measured data. Moreover, the true correlation structure within each cluster is often unknown, and ignoring such a correlation could yield biased estimators (Wang; 2003), inefficient estimators, and low power in hypothesis testing. Hence, we need to use a sophisticated method to parsimoniously separate within-subject and between-subject variation, and should not simply treat repeated measurements as cross-sectional observations. Consequently, developing an effective estimation procedure and then establishing its theoretical justifications for PLSiM with repeatedly measured data become an important and challenging task.

To alleviate the impact of correlation misspecification and to pursue estimation efficiency, we employ the quadratic inference function (QIF) proposed by Qu and Lindsay (2003); Qu et al. (2000); used for estimation in parametric models. This approach enables us to take into account the within-cluster correlation without specifying the covariance function. Furthermore, it is more efficient than the generalized estimating equation (GEE) approach when the working correlation is misspecified, as demonstrated in Qu et al. (2000). In this paper, we propose a QIF estimation procedure by incorporating the profile principle (Severini and Wong; 1992) for PLSiM with repeated measurements. Specifically, it consists of two steps: (i) For the given parametric components, employ the QIF approach to obtain an estimate of the nonparametric component by polynomial splines– It is noteworthy that the resulting nonparametric estimator is a function of the given parametric components; (ii) Based on the nonparametric estimator, construct the profiled QIF objective function for the parametric components and then obtain their estimators.

The QIF approach has been recently applied to estimation in single-index models (Bai et al.

(2009)) and PLSiM (Lai et al. (2013)) with the nonparametric functions estimated by penalized-splines and local linear smoothing, respectively. Spline estimation approach is known as computationally faster and more efficient than kernel smoothing in semiparametric models with correlated data (Lin et al. (2004)). It is noteworthy that Bai et al. (2009) studied the asymptotic normality for the index parameters by assuming that the true nonparametric link function is known. Comparing to Bai et al. (2009), we derive root-n consistency and a sandwich formula for the covariance matrix of the parametric estimators by estimating the link function with polynomial splines. Therefore, to obtain the asymptotic properties, we face significant theoretical challenges since the parametric QIF estimators are involved in the nonparametric functional estimator with divergent parameters. Thus, the classical asymptotic theory cannot be directly applied. Accordingly, we explore a new approach to establish the asymptotic normality of the parametric estimators in the PLSiM. Another contribution in our paper is that we introduce the penalized QIF (PQIF) to reduce the model complexity, which shrinks irrelevant coefficients of the linear and single-index components to zero. The resulting estimators of the nonzero coefficients are shown to be asymptotically normal and have the oracle property. Furthermore, Xue et al. (2010) applied the QIF to additive models and studied the optimal convergence rate of the spline estimators for the additive nonparametric functions.

The paper is organized as follows. Section 2 introduces the models and then applies QIF to obtain parametric and nonparametric estimations. The theoretical properties of parametric estimators are established. Section 3 proposes a penalized quadratic inference function approach for PLSiM to simultaneously estimate parameters and select variables, and the resulting estimators possess the oracle property. The practical implementations are developed in Section 4, and simulation studies and an empirical example are presented in Section 5. The last section concludes the article with a brief discussion, and technical proofs are given in the Appendix.

2 Models and Estimation Methods

2.1 Models

Suppose that the data consist of n independent subjects and the i-th ($i = 1, \ldots, n$) cluster has m_i repeated measures. Let Y_{ij} be the response variable, and $X_{ij} = (X_{ij,1}, \ldots, X_{ij,d_1})^{\mathrm{T}}$ and

$Z_{ij} = (Z_{ij,1}, \ldots, Z_{ij,d_2})^{\mathrm{T}}$ be $d_1 \times 1$ and $d_2 \times 1$ covariate vectors, respectively, for the j-th observation in the i-th cluster. Denote $Y_i = (Y_{i1}, \ldots, Y_{im_i})^{\mathrm{T}}$, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{im_i})^{\mathrm{T}}_{m_i \times d_1}$, and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{im_i})^{\mathrm{T}}_{m_i \times d_2}$. Consider the partially linear single-index model (PLSiM),

$$E\left(Y_{ij} \,|X_{ij}, Z_{ij}\right) = \mu_{ij} = g\left(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}\right) + \boldsymbol{\alpha}^{\mathrm{T}} Z_{ij}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is an unknown index vector which belongs to the parameter space $\{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_1})^{\mathrm{T}} : \|\boldsymbol{\beta}\|_2 = 1, \beta_1 > 0, \boldsymbol{\beta} \in R^{d_1}\}$, $\|\boldsymbol{\beta}\|_2 = \left(\beta_1^2 + \cdots + \beta_{d_1}^2\right)^{1/2}$ is the Euclidean norm of $\boldsymbol{\beta}$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_2})^{\mathrm{T}}$. In addition, $g(\cdot)$ is an unknown differentiable function of $U_{ij}(\boldsymbol{\beta}) = \boldsymbol{\beta}^{\mathrm{T}} X_{ij}$. Finally, assume that $e_i = Y_i - \mu_i$, where $\mu_i = (\mu_{i1}, \ldots, \mu_{im_i})^{\mathrm{T}}$, $e_i \sim (0, \Sigma_i)$, and Σ_i is a positive definite matrix. It is noteworthy that (2.1) reduces to Fan and Li's (2004) mean function of their semiparametric model when $d_1 = 1$.

2.2 Estimation method

In the estimation procedure, we approximate the smooth function g in (2.1) by polynomial splines. To this end, let $N = N_n$ be the number of interior knots. For theoretical reasons, we assume that, for any $\boldsymbol{\beta}$ in the neighborhood of its true parameter value, $U_{ij}(\boldsymbol{\beta})$ is distributed on a compact interval $[a, b]$ for $1 \leq j \leq m_i$ and $1 \leq i \leq n$, so that the range of the B-splines can be well defined. See Sun et al. (2008) for the same assumption. Divide $[a, b]$ into $(N + 1)$ subintervals: $I_J = [\xi_J, \xi_{J+1})$ for $J = 0, \ldots, N - 1$ and $I_N = [\xi_N, 1]$, where $(\xi_J)_{J=1}^N$ is a sequence of interior knots that is given as

$$\xi_{-(p-1)} = \cdots = a = \xi_0 < \xi_1 < \cdots < \xi_N < b = \xi_{N+1} = \cdots = \xi_{N+p}.$$

The resulting distance between neighboring knots is $h_J = \xi_{J+1} - \xi_J$ for $0 \leq J \leq N$, and let $h = \max_{0 \leq J \leq N} h_J$. Furthermore, define the p-th order B-spline basis as $B_p(u) = \{B_{J,p}(u) : 1 - p \leq J \leq N\}^{\mathrm{T}}$ (de Boor; 2001; Shen et al.; 1998), and let $G = G^{(p-2)}$ be the space spanned by $B_p(u)$. Then, the unknown function g in (2.1) can be approximated by a linear combination of the B-spline functions such that $g(u) \approx \sum_{J=1-p}^N \gamma_J B_{J,p}(u)$ with a set of coefficients $\boldsymbol{\gamma} = (\gamma_{1-p}, \ldots, \gamma_N)^{\mathrm{T}}$. Accordingly,

$$\mu_{ij} \approx \sum\nolimits_{J=1-p}^{N} \gamma_J B_{J,p}\left(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}\right) + \boldsymbol{\alpha}^{\mathrm{T}} Z_{ij} = \sum\nolimits_{J=1-p}^{N} \gamma_J B_{J,p}\left(U_{ij}(\boldsymbol{\beta})\right) + \boldsymbol{\alpha}^{\mathrm{T}} Z_{ij}. \tag{2.2}$$

To estimate $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and g, we next apply QIF to efficiently incorporate the within-cluster correlation structure. For the sake of simplicity, we assume that the number of repeated measurements are equal, i.e., $m_i = m < \infty$. Let $\boldsymbol{R}$ be a common working correlation matrix. Following the QIF approach, the inverse of $\boldsymbol{R}$ can be approximated by a linear combination of k basis matrices, i.e.,

$$\boldsymbol{R}^{-1} \approx a_1 \boldsymbol{M}_1 + \cdots + a_k \boldsymbol{M}_k, \tag{2.3}$$

where $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_k$ are known symmetric basis matrices and $a_1, \ldots, a_k$ are unknown constants. As stated in Qu et al. (2000), this is a sufficiently rich class that accommodates, or at least approximates, the correlation structures most commonly used. For example, if $\boldsymbol{R}$ is of compound symmetric structure with correlation ρ, $\boldsymbol{R}^{-1}$ can be represented by $a_1\boldsymbol{M}_1 + a_2\boldsymbol{M}_2$, where $\boldsymbol{M}_1 = \boldsymbol{I}$ (the identity matrix) and $\boldsymbol{M}_2$ is a matrix with diagonal entries 0 and off-diagonal entries 1, $a_1 = -\left\{(m-2)\rho + 1\right\}/k_1$, $a_2 = \rho/k_1$, $k_1 = (m-1)\rho^2 - (m-2)\rho - 1$, and m is the dimension of $\boldsymbol{R}$. See Qu et al. (2000) for more examples. Subsequently, we estimate the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ and the nonparametric function $g(\cdot)$ by a profile QIF approach in two steps below.

Step 1. For a fixed $\boldsymbol{\theta}$, let $\boldsymbol{B}_p(U_i(\boldsymbol{\beta})) = \left\{B_p(U_{i1}(\boldsymbol{\beta})), \ldots, B_p(U_{im}(\boldsymbol{\beta}))\right\}_{J_n \times m}$ and $J_n = N + p$, where $U_i(\boldsymbol{\beta}) = (U_{i1}(\boldsymbol{\beta}), \ldots, U_{im_i}(\boldsymbol{\beta}))^{\mathrm{T}}$. Then, from equations (2.1), (2.2), and (2.3), we obtain the estimating function of γ that is

$$\phi_i(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{B}_p(U_i(\boldsymbol{\beta}))\boldsymbol{A}^{-1/2}\boldsymbol{M}_1\boldsymbol{A}^{-1/2}\left\{\boldsymbol{B}_p(U_i(\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} - Y_i\right\} \\ \vdots \\ \boldsymbol{B}_p(U_i(\boldsymbol{\beta}))\boldsymbol{A}^{-1/2}\boldsymbol{M}_k\boldsymbol{A}^{-1/2}\left\{\boldsymbol{B}_p(U_i(\boldsymbol{\beta}))^{\mathrm{T}}\boldsymbol{\gamma} + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha} - Y_i\right\} \end{bmatrix}, \tag{2.4}$$

where $\boldsymbol{A}$ is an $m \times m$ diagonal matrix with its diagonal entries being the marginal variances of Y_{ij} for $j = 1, \cdots, m$. Let

$$\Phi_n(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(\boldsymbol{\gamma}, \boldsymbol{\theta}) \text{ and } \Psi_n(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \frac{1}{n^2}\sum_{i=1}^{n}\phi_i(\boldsymbol{\gamma}, \boldsymbol{\theta})\phi_i(\boldsymbol{\gamma}, \boldsymbol{\theta})^{\mathrm{T}}.$$

Since the dimension of (2.4) is kJ_n, which is larger than the number of unknown parameters, we follow Qu et al.'s (2000) approach to estimate $\boldsymbol{\gamma}$ by minimizing the following equation,

$$Q_n(\boldsymbol{\gamma}, \boldsymbol{\theta}) = \Phi_n(\boldsymbol{\gamma}, \boldsymbol{\theta})^{\mathrm{T}}\Psi_n(\boldsymbol{\gamma}, \boldsymbol{\theta})^{-1}\Phi_n(\boldsymbol{\gamma}, \boldsymbol{\theta}), \tag{2.5}$$

and the resulting estimator is $\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}}(\boldsymbol{\theta}) = \{\hat{\gamma}_J^{\mathrm{QIF}}(\boldsymbol{\theta})\}_{J=1-p}^{N} = \arg\min_{\boldsymbol{\gamma}} Q_n(\boldsymbol{\gamma}, \boldsymbol{\theta})$. Accordingly, the nonparametric estimator of g in (2.1) is

$$\tilde{g}\left(u,\boldsymbol{\theta}\right)=\sum\nolimits_{J=1-p}^{N}\hat{\gamma}_{J}^{\text{QIF}}\left(\boldsymbol{\theta}\right)B_{J,p}(u)=B(u)^{\mathrm{T}}\hat{\boldsymbol{\gamma}}^{\text{QIF}}\left(\boldsymbol{\theta}\right). \tag{2.6}$$

For estimating regression parameters, we need to estimate not only g but also its first derivative g'. To this end, we follow de Boor's (2001) approach to approximate g' by the spline functions with one order less than that of g. As a result,

$$\tilde{g}'\left(u,\boldsymbol{\theta}\right)=\sum\nolimits_{J=2-p}^{N}\hat{\gamma}_{J}^{\text{QIF},(1)}\left(\boldsymbol{\theta}\right)B_{J,p-1}(u),$$

where $\hat{\gamma}_{J}^{\text{QIF},(1)}\left(\boldsymbol{\theta}\right)=(p-1)\left\{\hat{\gamma}_{J}^{\text{QIF}}\left(\boldsymbol{\theta}\right)-\hat{\gamma}_{J-1}^{\text{QIF}}\left(\boldsymbol{\theta}\right)\right\}/\left(\xi_{J+p-1}-\xi_{J}\right)$ for $2-p\leq J\leq N$. In addition, it can be re-expressed as $\tilde{g}'\left(u,\boldsymbol{\theta}\right)=B_{p-1}(u)^{\mathrm{T}}D_{1}\hat{\boldsymbol{\gamma}}^{\text{QIF}}\left(\boldsymbol{\theta}\right)$, where $B_{p-1}(u)=\left\{B_{J,p-1}(u):2-p\leq J\leq N\right\}^{\mathrm{T}}$, and

$$D_{1}=(p-1)\begin{pmatrix}\frac{-1}{\xi_{1}-\xi_{2-p}} & \frac{1}{\xi_{1}-\xi_{2-p}} & 0 & \cdots & 0\\ 0 & \frac{-1}{\xi_{2}-\xi_{3-p}} & \frac{1}{\xi_{2}-\xi_{3-p}} & \cdots & 0\\ \vdots & \vdots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & \frac{-1}{\xi_{N+p-1}-\xi_{N}} & \frac{1}{\xi_{N+p-1}-\xi_{N}}\end{pmatrix}_{(J_{n}-1)\times J_{n}}.$$

Step 2. Before estimating $\boldsymbol{\theta}=(\boldsymbol{\beta}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$, we eliminate β_1 and reform the space of $\boldsymbol{\beta}$ to ensure identifiability (see Cui et al.; 2011), and the resulting space of $\boldsymbol{\beta}$ is $\{((1-\sum_{l=2}^{d_1}\beta_l^2)^{1/2},\beta_2,\ldots,\beta_{d_1})^{\mathrm{T}}:\sum_{l=2}^{d_1}\beta_l^2<1\}$.

Let $\boldsymbol{\beta}^{(1)}=(\beta_2,\ldots,\beta_{d_1})^{\mathrm{T}}$ and $\boldsymbol{J}=\partial\boldsymbol{\beta}/\partial\boldsymbol{\beta}^{(1)}$ be the Jacobian matrix of size $d_1\times(d_1-1)$, which is

$$\boldsymbol{J}=\begin{pmatrix}-\boldsymbol{\beta}^{(1)^{\mathrm{T}}}/\sqrt{1-\left\|\boldsymbol{\beta}^{(1)}\right\|^{2}}\\ \boldsymbol{I}_{d_1-1}\end{pmatrix}.$$

Then, we apply the same techniques used in the estimation of g to estimate $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\alpha}$. Let

$$\tilde{g}\left(U_{i}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)=\left\{\tilde{g}\left(U_{i1}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right),\ldots,\tilde{g}\left(U_{im}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)\right\}^{\mathrm{T}}$$

and its gradients with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ be

$$\nabla_{\beta}\tilde{g}\left(U_{i}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)=\left\{\frac{\partial\tilde{g}\left(U_{i1}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)}{\partial\boldsymbol{\beta}^{(1)}},\ldots,\frac{\partial\tilde{g}\left(U_{im}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)}{\partial\boldsymbol{\beta}^{(1)}}\right\}_{m\times(d_1-1)}^{\mathrm{T}}$$

and

$$\nabla_{\alpha}\tilde{g}\left(U_{i}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)=\left\{\frac{\partial\tilde{g}\left(U_{i1}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)}{\partial\boldsymbol{\alpha}},\ldots,\frac{\partial\tilde{g}\left(U_{im}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)}{\partial\boldsymbol{\alpha}}\right\}_{m\times d_2}^{\mathrm{T}},$$

respectively. Denote $\hat{D}_{i}\left(\boldsymbol{\theta}\right)=\left[\nabla_{\beta}\tilde{g}\left(U_{i}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right),\left\{\nabla_{\alpha}\tilde{g}\left(U_{i}\left(\boldsymbol{\beta}\right),\boldsymbol{\theta}\right)+\boldsymbol{Z}_{i}\right\}\right]_{(d_1-1+d_2)\times m}^{\mathrm{T}}$.

Based on equations (2.1), (2.2), and (2.3), we obtain the estimating function of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ that is

$$\omega_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{bmatrix} \hat{D}_i(\boldsymbol{\theta}) \boldsymbol{A}^{-1/2} \boldsymbol{M}_1 \boldsymbol{A}^{-1/2} \left\{ \tilde{g}\left(U_i(\boldsymbol{\beta}), \boldsymbol{\theta}\right) + \boldsymbol{Z}_i \boldsymbol{\alpha} - Y_i \right\} \\ \vdots \\ \hat{D}_i(\boldsymbol{\theta}) \boldsymbol{A}^{-1/2} \boldsymbol{M}_k \boldsymbol{A}^{-1/2} \left\{ \tilde{g}\left(U_i(\boldsymbol{\beta}), \boldsymbol{\theta}\right) + \boldsymbol{Z}_i \boldsymbol{\alpha} - Y_i \right\} \end{bmatrix}. \tag{2.7}$$

Since the dimension of (2.7) is $k(d_1 - 1 + d_2)$ which is larger than the number of unknown parameters, we adopt Qu et al.'s (2000) approach to estimate $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\alpha}$ by minimizing the following function,

$$Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \Omega_n(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}} \Xi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \Omega_n(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{2.8}$$

where

$$\Omega_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^{n} \omega_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \text{ and } \Xi_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{n^2} \sum_{i=1}^{n} \omega_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \omega_i(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}. \tag{2.9}$$

As a result, we obtain the estimators $\hat{\boldsymbol{\beta}}^{(1)\text{QIF},}$ and $\hat{\boldsymbol{\alpha}}^{\text{QIF}}$ of $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\alpha}$. Then, using the fact that $\beta_1 = \sqrt{1 - \left\|\boldsymbol{\beta}^{(1)}\right\|^2}$, we have $\hat{\beta}_1^{\text{QIF}}$. Let $\boldsymbol{\beta}^0 = \left\{\beta_1^0, \left(\boldsymbol{\beta}^{(1)0}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0$ be the true parameters in model (2.1). For the sake of simplicity, define $\tilde{\xi}_{ij} = \xi_{ij} - E\left(\xi_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$ and $\hat{\xi}_{ij} = \xi_{ij} - \hat{E}\left(\xi_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$, where $\hat{E}\left(\xi_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$ is an estimator of $E\left(\xi_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$ obtained by using polynomial splines with order p for the random variable (or vector) ξ_{ij}. For example, $\tilde{X}_{ij} = X_{ij} - E\left(X_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$ and $\hat{X}_{ij} = X_{ij} - \hat{E}\left(X_{ij} \left| U_{ij}\left(\boldsymbol{\beta}^0\right)\right.\right)$. After simplification, it can be shown that for all $1 \le j \le m$ and $1 \le i \le n$,

$$\frac{\partial \tilde{g}\left(U_{ij}\left(\boldsymbol{\beta}^0\right), \boldsymbol{\theta}^0\right)}{\partial \boldsymbol{\beta}^{(1)}} = g'\left(U_{ij}\left(\boldsymbol{\beta}^0\right), \boldsymbol{\theta}^0\right) \boldsymbol{J}^{\mathrm{T}} \hat{X}_{ij} \left\{1 + o_p(1)\right\},$$
$$\frac{\partial \tilde{g}\left(U_{ij}\left(\boldsymbol{\beta}^0\right), \boldsymbol{\theta}^0\right)}{\partial \boldsymbol{\alpha}} = -\hat{E}\left(Z_{ij} \left| U_{ij}(\boldsymbol{\beta}^0)\right.\right) \left\{1 + o_p(1)\right\}.$$

Accordingly, $\hat{D}_i(\boldsymbol{\theta}^0) = [\{g'(U_{i1}(\boldsymbol{\beta}^0), \boldsymbol{\theta}^0)\hat{X}_{i1}, \ldots, g'(U_{im}(\boldsymbol{\beta}^0), \boldsymbol{\theta}^0)\hat{X}_{im}\}^{\mathrm{T}} \boldsymbol{J}, (\hat{Z}_{i1}, \ldots, \hat{Z}_{im})^{\mathrm{T}}]^{\mathrm{T}}\{1 + o_p(1)\}$.

Let $\hat{\boldsymbol{\theta}}^{(1)\text{QIF}} = \left\{\left(\hat{\boldsymbol{\beta}}^{(1)\text{QIF}}\right)^{\mathrm{T}}, \left(\hat{\boldsymbol{\alpha}}^{\text{QIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\theta}^{(1)0} = \left\{\left(\boldsymbol{\beta}^{(1)0}\right)^{\mathrm{T}}, \left(\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$. Moreover, define

$$D_i\left(\boldsymbol{\beta}^0\right) = \left[\left\{g'\left(\boldsymbol{\beta}^{0\mathrm{T}} X_{i1}\right) \tilde{X}_{i1}, \ldots, g'\left(\boldsymbol{\beta}^{0\mathrm{T}} X_{im}\right) \tilde{X}_{im}\right\}^{\mathrm{T}} \boldsymbol{J}, \left(\tilde{Z}_{i1}, \ldots, \tilde{Z}_{im}\right)^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{2.10}$$

$$\dot{\Omega}\left(\boldsymbol{\beta}^0\right) = E \left\{ \begin{matrix} D_i\left(\boldsymbol{\beta}^0\right) \Lambda_1 D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \\ \vdots \\ D_i\left(\boldsymbol{\beta}^0\right) \Lambda_k D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \end{matrix} \right\}, \tag{2.11}$$

and

$$\Xi\left(\boldsymbol{\beta}^0\right)=E\left\{\begin{array}{ccc} D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{1,1}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} & \cdots & D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{1,k}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{k,1}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} & \cdots & D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{k,k}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \end{array}\right\}, \tag{2.12}$$

where $\Gamma_{r,r'}=\Lambda_r\Sigma\Lambda_{r'}$ with $\Lambda_r=\boldsymbol{A}^{-1/2}\boldsymbol{M}_r\boldsymbol{A}^{-1/2}$ for $r,r'=1,\ldots,k$. Then, the asymptotic properties of parametric estimators are given below.

Theorem 1. *Under Conditions (C1)-(C6) given in Appendix A.1 and the assumption that* $nN^{-(2p+2)}\to 0$ *and* $nN^{-4}\to\infty$ *as* $n\to\infty$*, we have that,* $\left\|\hat{\boldsymbol{\theta}}^{(1)\mathrm{QIF}}-\boldsymbol{\theta}^{(1)0}\right\|=O_p\left(n^{-1/2}\right)$*, and as* $n\to\infty$*,* $\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(1)\mathrm{QIF}}-\boldsymbol{\theta}^{(1)0}\right)\to N\left(0,\Sigma^{-1}_{\boldsymbol{\theta}^{(1)0}}\right)$*, where* $\Sigma_{\boldsymbol{\theta}^{(1)0}}=\dot{\Omega}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Xi\left(\boldsymbol{\beta}^0\right)^{-1}\dot{\Omega}\left(\boldsymbol{\beta}^0\right)$.

In addition, partition $\Sigma^{-1}_{\boldsymbol{\theta}^{(1)0}}$ into $\left\{\begin{array}{cc}\left(\Sigma^{11}_{\boldsymbol{\theta}^{(1)0}}\right)_{(d_1-1)\times(d_1-1)} & \left(\Sigma^{12}_{\boldsymbol{\theta}^{(1)0}}\right)_{(d_1-1)\times d_2} \\ \left(\Sigma^{21}_{\boldsymbol{\theta}^{(1)0}}\right)_{d_2\times(d_1-1)} & \left(\Sigma^{22}_{\boldsymbol{\theta}^{(1)0}}\right)_{d_2\times d_2}\end{array}\right\}$, and then we obtain the following results.

Theorem 2. *Under Conditions (C1)-(C6) in Appendix A.1 and the assumption of that* $nN^{-(2p+2)}\to 0$ *and* $nN^{-4}\to\infty$ *as* $n\to\infty$*, we have that, as* $n\to\infty$*,*
$\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(1)\mathrm{QIF}}-\boldsymbol{\beta}^{(1)0}\right)\to N\left(0,\Sigma^{11}_{\boldsymbol{\theta}^{(1)0}}\right)$ *and* $\sqrt{n}\left(\hat{\boldsymbol{\alpha}}^{\mathrm{QIF}}-\boldsymbol{\alpha}^0\right)\to N\left(0,\Sigma^{22}_{\boldsymbol{\theta}^{(1)0}}\right)$.

Theorem 1, together with the multivariate delta-method, leads to $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{\mathrm{QIF}}-\boldsymbol{\beta}^0\right)\to N\left(0,\Sigma_{\boldsymbol{\beta}^0}\right)$, as $n\to\infty$, where $\Sigma_{\boldsymbol{\beta}^0}=\boldsymbol{J}\Sigma^{11}_{\boldsymbol{\theta}^{(1)0}}\boldsymbol{J}^{\mathrm{T}}$.

After obtaining the estimator $\hat{\boldsymbol{\theta}}^{\mathrm{QIF}}=\{(\hat{\boldsymbol{\beta}}^{\mathrm{QIF}})^{\mathrm{T}},(\hat{\boldsymbol{\alpha}}^{\mathrm{QIF}})^{\mathrm{T}}\}$, we replace $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}^{\mathrm{QIF}}$ in $\tilde{g}\left(u,\boldsymbol{\theta}\right)$ defined in (2.6) and obtain $\hat{g}(u)$. To study the properties of $\hat{g}(u)$, we define

$$\tilde{\dot{\Phi}}_n\left(\boldsymbol{\gamma}\right)=\tilde{\dot{\Phi}}_n=\frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Lambda_1\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Lambda_k\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}}\end{array}\right\} \tag{2.13}$$

and

$$\tilde{\Psi}_n=n^{-2}\sum_{i=1}^{n}\left\{\begin{array}{ccc}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Gamma_{1,1}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}} & \cdots & \boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Gamma_{1,k}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Gamma_{k,1}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}} & \cdots & \boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Gamma_{k,k}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}}\end{array}\right\}_{kJ_n\times kJ_n}. \tag{2.14}$$

Furthermore, let $B^*_p\left(\cdot\right)$ be the p-th Bernoulli polynomial that is inductively defined as follows:

$$B^*_0(u)=1, B^*_i(u)=\int_0^u iB^*_{i-1}(u)d(u)+b_i,$$

where $b_i = -i\int_0^1\int_0^x B_{i-1}^*(u)dudx$ is the i-th Bernoulli number. Subsequently, for $u \in I_J$ and $0 \le J \le N$, define

$$b^*(u) = -\frac{g^{(p)}(u)}{p!}B_p^*\left(\frac{u-\xi_J}{h}\right) \text{ and } \vartheta_n = n^{-1}\sum_{i=1}^{n}\left\{\begin{array}{c}\boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Lambda_1 b^*\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\\ \vdots \\ \boldsymbol{B}_p\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\Lambda_k b^*\left(U_i\left(\boldsymbol{\beta}^0\right)\right)\end{array}\right\}. \quad (2.15)$$

Moreover, denote $d^*(u) = B_p(u)^{\mathrm{T}}(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n)^{-1}\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\vartheta_n$. Then, we obtain the asymptotic results of $\hat{g}(u)$ given below.

Theorem 3. *Under Conditions (C1)-(C6) given in Appendix A.1 and the assumption that $nN^{-(2p+2)} \to 0$ and $nN^{-4} \to \infty$ as $n \to \infty$, we have that, as $n \to \infty$,*

(i) $\sigma_n^{-1}(u)\{\hat{g}(u) - g(u) - b^*(u)h^p + d^*(u)h^p\} \to N(0,1)$ *for any $u \in I_J$ and $0 \le J \le N$, where $\sigma_n^2(u) = \boldsymbol{B}_p(u)^{\mathrm{T}}(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n)^{-1}\boldsymbol{B}_p(u)$, $\widetilde{\dot{\Phi}}, \tilde{\Psi}$, and $b^*(u)$ and ϑ_n are defined in (2.13), (2.14), and (2.15), respectively;*

(ii) $|\hat{g}(u) - g(u)| = O_p\left\{(nh)^{-1/2} + h^p\right\}$ *uniformly in $u \in [0,1]$.*

Remark 1: By letting $N \asymp n^{1/(2p+1)}$ which satisfies the assumption on N in the above theorem, the spline estimator $\hat{g}(u)$ has the optimal convergence rate, which is $O_p\left\{n^{-p/(2p+1)}\right\}$. It is noteworthy that Shen et al. (1998) established the same convergence rate for spline estimation in univariate nonparametric regression.

3 Penalized QIF for PLSiM

In practice, the true model is often unknown. This motivates us to penalize the QIF in (2.8) and then simultaneously select significant variables and estimate resulting parameters $(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}})^{\mathrm{T}}$of the parametric components $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$. By the assumption that $\beta_1 > 0$ in model (2.1), we assume that the first component X_1 of $X = (X_1, \ldots, X_{d_1})^{\mathrm{T}}$ is a significant covariate. Accordingly, the penalized quadratic inference function (PQIF) is as follows:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) + n\sum_{l=2}^{d_1}p_{\lambda_{1l}}(|\beta_l|) + n\sum_{s=1}^{d_2}p_{\lambda_{2s}}(|\alpha_s|), \quad (3.1)$$

where $p_\lambda(\cdot)$ is a penalty function with a regularization parameter λ. We allow $(\beta_l)_{l=2}^{d_1}$ and $(\alpha_s)_{s=1}^{d_2}$ to have different penalty functions with different regularization parameters.

There are various penalty functions available in literature such as the L_1 and L_2 penalties, that yield the LASSO-type and ridge-type estimators, respectively. Here, we consider the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001), whose first derivative is defined as

$$p'_\lambda(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\},$$

where $p_\lambda(0) = 0$, $a = 3.7$, and $(t)_+ = tI(t > 0)$. For the given tuning parameters, the penalized estimators $\hat{\boldsymbol{\beta}}^{\text{PQIF},(1)}$ and $\hat{\boldsymbol{\alpha}}^{\text{PQIF}}$ are obtained by minimizing $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\alpha}$. As a result, β_1 is estimated by $\hat{\beta}_1^{\text{PQIF}} = \sqrt{1 - \left\|\hat{\boldsymbol{\beta}}^{\text{PQIF},(1)}\right\|_2^2}$.

We next study the theoretical properties of estimators $\hat{\boldsymbol{\beta}}^{\text{PQIF}} = \{\hat{\beta}_1^{\text{PQIF}}, (\hat{\boldsymbol{\beta}}^{\text{PQIF},(1)})^{\mathrm{T}}\}^{\mathrm{T}}$ and $\hat{\boldsymbol{\alpha}}^{\text{PQIF}}$ with the SCAD penalty. Without loss of generality, we assume that the correct model in (2.1) has regression coefficients $\boldsymbol{\beta}^0 = \left\{\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}}, \left(\boldsymbol{\beta}_2^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0 = \left\{\left(\boldsymbol{\alpha}_1^0\right)^{\mathrm{T}}, \left(\boldsymbol{\alpha}_2^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$, where $\boldsymbol{\beta}_1^0 = \{\beta_1^0, (\boldsymbol{\beta}_1^{(1)0})^{\mathrm{T}}_{(d_{10}-1)\times 1}\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1^0$ are $d_{10} \times 1$ and $d_{20} \times 1$ nonzero components of $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$, respectively, and $\boldsymbol{\beta}_2^0$ and $\boldsymbol{\alpha}_2^0$ are $(d_1 - d_{10}) \times 1$ and $(d_2 - d_{20}) \times 1$ vectors with zeros. In addition, let $X_{ij}^1 = \{(X_{ij,1}, \ldots, X_{ij,d_{10}})^{\mathrm{T}}\}_{d_{10}\times 1}$ and $Z_{ij}^1 = \{(Z_{ij,1}, \ldots, Z_{ij,d_{20}})^{\mathrm{T}}\}_{d_{20}\times 1}$. Then, define

$$\boldsymbol{J}_1 = \begin{pmatrix} -\boldsymbol{\beta}_1^{(1)0^{\mathrm{T}}} \Big/ \sqrt{1 - \left\|\boldsymbol{\beta}_1^{(1)0}\right\|^2} \\ \boldsymbol{I}_{d_{10}-1} \end{pmatrix}_{d_{10}\times(d_{10}-1)}, \tag{3.2}$$

$$D_{1i}(\boldsymbol{\beta}_1) = \left[\left\{g'\left(\boldsymbol{\beta}_1^{\mathrm{T}} X_{i1}^1\right)\tilde{X}_{i1}^1, \ldots, g'\left(\boldsymbol{\beta}_1^{\mathrm{T}} X_{im}\right)\tilde{X}_{im}^1\right\}^{\mathrm{T}} \boldsymbol{J}_1, \left(\tilde{Z}_{i1}^1, \ldots, \tilde{Z}_{im}^1\right)^{\mathrm{T}}\right]^{\mathrm{T}},$$

$$\dot{\Omega}_1\left(\boldsymbol{\beta}_1^0\right) = E\left\{\begin{array}{c} D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Lambda_1 D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} \\ \vdots \\ D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Lambda_k D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} \end{array}\right\},$$

and

$$\Xi_1\left(\boldsymbol{\beta}_1^0\right) = E\left\{\begin{array}{ccc} D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Gamma_{1,1} D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} & \cdots & D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Gamma_{1,k} D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Gamma_{k,1} D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} & \cdots & D_{1i}\left(\boldsymbol{\beta}_1^0\right)\Gamma_{k,k} D_{1i}\left(\boldsymbol{\beta}_1^0\right)^{\mathrm{T}} \end{array}\right\}$$

where $\tilde{X}_{ij}^1 = X_{ij}^1 - E\left(X_{ij}^1 \,\middle|\, \boldsymbol{\beta}_1^{\mathrm{T}} X_{ij}^1\right)$ and $\tilde{Z}_{ij}^1 = Z_{ij}^1 - E\left(Z_{ij}^1 \,\middle|\, \boldsymbol{\beta}_1^{\mathrm{T}} X_{ij}^1\right)$. Moreover, denote $\hat{\boldsymbol{\beta}}^{\text{PQIF},(1)} = \left\{\left(\hat{\boldsymbol{\beta}}_1^{\text{PQIF},(1)}\right)^{\mathrm{T}}, \left(\hat{\boldsymbol{\beta}}_2^{\text{PQIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\hat{\boldsymbol{\alpha}}^{\text{PQIF}} = \{(\hat{\boldsymbol{\alpha}}_1^{\text{PQIF}})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}_2^{\text{PQIF}})^{\mathrm{T}}\}^{\mathrm{T}}$. Subsequently, we obtain the oracle properties of the penalized estimators given below.

Theorem 4. *Under Conditions (C1)-(C6) in the Appendix A.1, if* $nN^{-(2p+2)} \to 0$, $nN^{-4} \to \infty$, $\lambda_{1l} \to 0$, $n^{1/2}\lambda_{1l} \to \infty$, $\lambda_{2s} \to 0$, *and* $n^{1/2}\lambda_{2s} \to \infty$ *for all* $1 \leq l \leq d_1$ *and* $1 \leq s \leq d_2$, *then we have that, as* $n \to \infty$, *the penalized estimators satisfy: i)* $P(\hat{\boldsymbol{\beta}}_2^{\text{PQIF}} = 0, \hat{\boldsymbol{\alpha}}_2^{\text{PQIF}} = 0) \to 1$*; ii)* $\sqrt{n}\left\{\begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(1)\text{PQIF}} - \boldsymbol{\beta}_1^{(1)0} \\ \hat{\boldsymbol{\alpha}}_1^{\text{PQIF}} - \boldsymbol{\alpha}_1^0 \end{pmatrix}\right\} \to N\left\{\mathbf{0}, \Sigma^{-1}_{\theta_1^{(1)0}}\right\}$, *where* $\Sigma_{\theta_1^{(1)0}} = \dot{\Omega}_1\left(\boldsymbol{\beta}_1^0\right)^{\text{T}} \Xi_1\left(\beta_1^0\right)^{-1} \dot{\Omega}_1\left(\boldsymbol{\beta}_1^0\right)$.

Under the conditions of Theorem 4, it can be easily shown that $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\text{PQIF}} - \boldsymbol{\beta}_1^0) \to N(0, \boldsymbol{J}\Sigma^{11}_{\theta_1^{(1)0}}\boldsymbol{J}^{\text{T}})$ and $\sqrt{n}(\hat{\boldsymbol{\alpha}}_1^{\text{PQIF}} - \boldsymbol{\alpha}_1^0) \to N(0, \Sigma^{22}_{\theta_1^{(1)0}})$, where $\Sigma^{11}_{\theta_1^{(1)0}}$ and $\Sigma^{22}_{\theta_1^{(1)0}}$ are the upper and lower block-diagonal matrices of $\Sigma^{-1}_{\theta_1^{(1)0}}$ with sizes $d_{10} - 1$ and d_{20}, respectively. As a result, the PQIF method yields consistent and asymptotically unbiased estimators, even though the correlation structure is misspecified. As demonstrated in Qu et al. (2000), the QIF approach is more efficient than the generalized estimating equation (GEE) approach when the working correlation is misspecified. Although the true correlation structure is unknown in practice, the closer the working correlation is to the true correlation, the more efficient the regression coefficient estimators can be, which leads to a better model. Hence, we propose employing the coefficient of determination to select an appropriate correlation structure.

4 Estimation Algorithm

We propose an algorithm to find penalized estimators by minimizing equations (2.5) and (3.1). For given $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, the estimating equation of (2.5) for $\boldsymbol{\gamma}$, is

$$\dot{Q}_n(\boldsymbol{\gamma}) = 2\dot{\Phi}_n(\boldsymbol{\gamma})^{\text{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}) - \Phi_n(\boldsymbol{\gamma})^{\text{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \dot{\Psi}_n(\boldsymbol{\gamma}) \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}) = 0, \quad (4.1)$$

where $\dot{\Phi}_n(\boldsymbol{\gamma})$ is a $kJ_n \times J_n$ matrix $\{\partial\Phi_n(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}\}$, $\dot{\Psi}_n(\boldsymbol{\gamma})$ is a three dimensional array $\{\partial\Psi_n(\boldsymbol{\gamma})/\partial\gamma_1, \ldots, \partial\Psi_n(\boldsymbol{\gamma})/\partial\gamma_{J_n}\}$, $\Phi_n(\boldsymbol{\gamma})^{\text{T}}\Psi_n(\boldsymbol{\gamma})^{-1}\dot{\Psi}_n(\boldsymbol{\gamma})\Psi_n(\boldsymbol{\gamma})^{-1}\Phi_n(\boldsymbol{\gamma})$ is a $J_n \times 1$ vector

$$\left[\Phi_n(\boldsymbol{\gamma})^{\text{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \{\partial\Psi_n(\boldsymbol{\gamma})/\partial\gamma_J\} \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma})\right]_{J=1-p}^{N}.$$

And the second derivative of $Q_n(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ is $\ddot{Q}_n(\boldsymbol{\gamma}) = 2\dot{\Phi}_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \dot{\Phi}_n(\boldsymbol{\gamma}) + R_n$, where

$$
\begin{aligned}
R_n = {} & 2\ddot{\Phi}_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}) - 4\dot{\Phi}_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \dot{\Psi}_n(\boldsymbol{\gamma}) \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}) \\
& + 2\dot{\Phi}_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \Psi_n(\boldsymbol{\gamma})^{-1} \dot{\Psi}_n(\boldsymbol{\gamma}) \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}) \\
& - \Phi_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \ddot{\Psi}_n(\boldsymbol{\gamma}) \Psi_n(\boldsymbol{\gamma})^{-1} \Phi_n(\boldsymbol{\gamma}).
\end{aligned}
$$

Since $R_n = o_p(1)$, $\ddot{Q}_n(\boldsymbol{\gamma})$ can be approximated by $2\dot{\Phi}_n(\boldsymbol{\gamma})^{\mathrm{T}} \Psi_n(\boldsymbol{\gamma})^{-1} \dot{\Phi}_n(\boldsymbol{\gamma})$. By the definition of $Q_n(\boldsymbol{\gamma})$, its first and second derivatives implicitly depend on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For the sake of clarification, we denote $\dot{Q}_n(\boldsymbol{\gamma}) = \dot{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\ddot{Q}_n(\boldsymbol{\gamma}) = \ddot{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$. Then, applying Newton-Raphson method, we obtain the iterative estimation procedure,

$$
\hat{\boldsymbol{\gamma}}^{i+1} = \hat{\boldsymbol{\gamma}}^{i} - \ddot{Q}_n^{-1}\left(\hat{\boldsymbol{\gamma}}^{i}, \hat{\boldsymbol{\beta}}^{i}, \hat{\boldsymbol{\alpha}}^{i}\right) \dot{Q}_n\left(\hat{\boldsymbol{\gamma}}^{i}, \hat{\boldsymbol{\beta}}^{i}, \hat{\boldsymbol{\alpha}}^{i}\right), \tag{4.2}
$$

where $\left(\hat{\boldsymbol{\gamma}}^{i}, \hat{\boldsymbol{\beta}}^{i}, \hat{\boldsymbol{\alpha}}^{i}\right)$ is the estimated values of $(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ at the i-th step for $i \geq 0$. Accordingly, g and g' can be estimated by $\hat{g}^{i+1}(u) = B_p(u)^{\mathrm{T}} \hat{\boldsymbol{\gamma}}^{i+1}$ and $(\hat{g}^{i+1}(u))' = B_{p-1}(u)^{\mathrm{T}} D_1 \hat{\boldsymbol{\gamma}}^{i+1}$. Let $\boldsymbol{\theta}^{(1)} = \left\{\left(\boldsymbol{\beta}^{(1)}\right)^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right\}^{\mathrm{T}}$.

To estimate $\boldsymbol{\theta}^{(1)} = \left\{\left(\boldsymbol{\beta}^{(1)}\right)^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$, we next consider the following quantities:

$$
\begin{aligned}
\dot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \dot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \hat{g}, \hat{g}') = \left\{\partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \boldsymbol{\theta}^{(1)}\right\}_{(d_1-1+d_2)\times 1}, \\
\ddot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \ddot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \hat{g}, \hat{g}') = \left\{\partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \boldsymbol{\theta}^{(1)} \left(\boldsymbol{\theta}^{(1)}\right)^{\mathrm{T}}\right\}_{(d_1-1+d_2)\times(d_1-1+d_2)},
\end{aligned}
$$

$$
\begin{aligned}
\Sigma_{\lambda_1}^*(\boldsymbol{\beta}) &= \operatorname{diag}\left\{p'_{\lambda_{12}}(|\beta_2|) / |\beta_2|, \ldots, p'_{\lambda d_1}(|\beta_{d_1}|) / |\beta_{d_1}|\right\}, \\
\Sigma_{\lambda_2}^*(\boldsymbol{\alpha}) &= \operatorname{diag}\left\{p'_{\lambda_{21}}(|\alpha_1|) / |\alpha_1|, \ldots, p'_{\lambda d_2}(|\alpha_{d_2}|) / |\alpha_{d_2}|\right\},
\end{aligned}
$$

and $\Sigma_{\lambda}^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{pmatrix} \Sigma_{\lambda_1}^*(\boldsymbol{\beta}) & \mathbf{0} \\ \mathbf{0} & \Sigma_{\lambda_2}^*(\boldsymbol{\alpha}) \end{pmatrix}$. Then, set the initial values $\hat{\beta}_l^0$ of $\hat{\beta}_l$ to be the standardized values obtained by fitting the single index model using the np package in R, while the initial values $\hat{\alpha}_s^0$ of $\hat{\alpha}_s$ and $\hat{\boldsymbol{\gamma}}^0$ of $\hat{\boldsymbol{\gamma}}$ are calculated from the least squares approach. If the i-th iterative estimate $\hat{\beta}_l^i$ and $\hat{\alpha}_s^i$ are close to zero (i.e., $\left|\hat{\beta}_l^i\right| < \delta_1$ and $|\hat{\alpha}_s^i| < \delta_1$ for some small threshold value δ_1), we set $\hat{\beta}_l^{i+1} = \hat{\alpha}_s^{i+1} = 0$. When $\hat{\beta}_l^{i+1} \neq 0$ and $\hat{\alpha}_s^{i+1} \neq 0$, we adopt the approach of Fan and Li

(2001) and obtain the locally quadratic approximation of $\mathcal{L}\left(\hat{\boldsymbol{\beta}}^{i+1}, \hat{\boldsymbol{\alpha}}^{i+1}\right)$ in (3.1) as follows:

$$Q_n^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i\right) + \dot{Q}_n^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i, \hat{g}^{i+1}, \left(\hat{g}^{i+1}\right)'\right)^{\mathrm{T}}\left(\hat{\boldsymbol{\theta}}^{(1),i+1} - \hat{\boldsymbol{\theta}}^{(1),i}\right)$$
$$+\frac{1}{2}\left(\hat{\boldsymbol{\theta}}^{(1),i+1} - \hat{\boldsymbol{\theta}}^{(1),i}\right)^{\mathrm{T}} \ddot{Q}_n^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i, \hat{g}^{i+1}, \left(\hat{g}^{i+1}\right)'\right)\left(\hat{\boldsymbol{\theta}}^{(1),i+1} - \hat{\boldsymbol{\theta}}^{(1),i}\right) + n\sum_{l=2}^{d_1} p_{\lambda_{1l}}\left(\left|\hat{\beta}_l^i\right|\right)$$
$$+n\sum_{s=1}^{d_2} p_{\lambda_{2s}}\left(\left|\hat{\alpha}_s^i\right|\right) + \frac{1}{2}n\left(\hat{\boldsymbol{\theta}}^{(1),i+1}\right)^{\mathrm{T}}\Sigma_\lambda^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i\right)\left(\hat{\boldsymbol{\theta}}^{(1),i+1}\right) - \frac{1}{2}n\left(\hat{\boldsymbol{\theta}}^{(1),i}\right)^{\mathrm{T}}\Sigma_\lambda^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i\right)\hat{\boldsymbol{\theta}}^{(1),i}.$$

Minimizing the above function with respect to $\hat{\boldsymbol{\theta}}^{(1),i+1}$, we obtain that

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{(1),i+1} \\ \hat{\boldsymbol{\alpha}}^{i+1} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(1),i} \\ \hat{\boldsymbol{\alpha}}^{i} \end{pmatrix} - \left\{\ddot{Q}_n^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i, \hat{g}^{i+1}, \left(\hat{g}^{i+1}\right)'\right) + n\Sigma_\lambda^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i\right)\right\}^{-1}$$
$$\left\{\dot{Q}_n^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i, \hat{g}^{i+1}, \left(\hat{g}^{i+1}\right)'\right) + n\Sigma_\lambda^*\left(\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\alpha}}^i\right)\begin{pmatrix} \hat{\boldsymbol{\beta}}^{(1),i} \\ \hat{\boldsymbol{\alpha}}^{i} \end{pmatrix}\right\}, \tag{4.3}$$

and $\hat{\beta}_1^{i+1}$ is calculated from $\sqrt{1 - \left\|\hat{\boldsymbol{\beta}}^{(1),i+1}\right\|_2^2}$. Repeat (4.2) and (4.3) until $\left\|\hat{\boldsymbol{\theta}}^{i+1} - \hat{\boldsymbol{\theta}}^{i}\right\|_2 < \delta_2$ and $\left\|\hat{\gamma}^{i+1} - \hat{\gamma}^i\right\|_2 < \delta_2$ for some small threshold value δ_2. Using similar techniques given above, we are able to obtain the spline coefficient estimator $\hat{\boldsymbol{\gamma}}^{\text{QIF}}$ and unpenalized estimators $\left(\hat{\boldsymbol{\beta}}^{\text{QIF},(1)}, \hat{\boldsymbol{\alpha}}^{\text{QIF}}\right)$ by minimizing (2.5) and (2.8) iteratively.

It is noteworthy that the tuning parameters are unknown in the computation of penalized estimators. Hence, we need to select the tuning parameters in the iterative process to complete the whole computation algorithm. To this end, we employ the BIC criterion modified by Wang et al. (2007) to choose the tuning parameters p_{λ_1} and p_{λ_2} in the penalized-QIF procedure, which is

$$\text{BIC}\left(\lambda_1, \lambda_2\right) = \mathcal{L}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) + \log\left(n\right) \times \left(d_1 + d_2 - 1\right)/n.$$

The optimal λ_1 and λ_2 are selected by minimizing BIC, i.e., $(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\min_{(\lambda_1,\lambda_2)}\text{BIC}(\lambda_1, \lambda_2)$.

5 Numerical Studies

5.1 Simulation results

We present Monte Carlo studies to evaluate the finite sample performance of the proposed penalized QIF estimators. To this end, let $\mathcal{S}$ represent any candidate model and $\mathcal{S}_0$ be the true model. Then, define the model $\mathcal{S}$ to be overfitted, correct, and underfitted if $\mathcal{S} \supset \mathcal{S}_0$ and $\mathcal{S} \neq \mathcal{S}_0$, $\mathcal{S} = \mathcal{S}_0$, and $\mathcal{S} \not\supset \mathcal{S}_0$, respectively. In 500 realizations, we calculate the proportions of models correctly fitted (C), overfitted (O), and underfitted (U). In addition, we report the average number of truly nonzero coefficients that are correctly set to nonzero (NC), and the average number of truly nonzero coefficients that are incorrectly set to zero (NI). To assess the estimation accuracy, we compare the SCAD-penalized QIF (PQIF) estimate with the ORACLE estimate obtained by assuming that the true model is known *a priori*. The assessment measure is the squared root value of the estimated mean squared error (MSE) defined as $\sum_{k=1}^{500} \left\| \hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}^0 \right\|^2 /500$ and $\sum_{k=1}^{500} \left\| \hat{\boldsymbol{\alpha}}_{(k)} - \boldsymbol{\alpha}^0 \right\|^2 /500$, where $\hat{\boldsymbol{\beta}}_{(k)}$ and $\hat{\boldsymbol{\alpha}}_{(k)}$ are the parameter estimates calculated in the k-th realization.

In this study, we consider a "sine-bump" model (Carroll et al.; 1997)

$$Y_{ij} = \sin\left\{ \frac{\pi\left(\boldsymbol{\beta}^{0\mathrm{T}} X_{ij} - A\right)}{C - A} \right\} + \boldsymbol{\alpha}^{0\mathrm{T}} Z_{ij} + \varepsilon_{ij},$$

where $A = \sqrt{3}/2 - 1.645/\sqrt{12}$, $C = \sqrt{3}/2 + 1.645/\sqrt{12}$, $j = 1, ..., 5$, $i = 1, \cdots, n$, and $n = 100, 200$, and 500. Furthermore, covariates $X_{ij} = (X_{ij,1}, \ldots, X_{ij,7})^{\mathrm{T}}$ are independently generated from uniform $[0, 1]$, $Z_{ij,1}$ are simulated from Bernoulli(0.5), and $(Z_{ij,2}, Z_{ij,3}, Z_{ij,4})^{\mathrm{T}}$ follows a multivariate normal distribution with mean 0 and covariance (i.e., cov$(Z_{ij,k_1}, Z_{ij,k_2}) = 0.5^{|k_1-k_2|}$ for $2 \leq k_1, k_2 \leq 4$). Their corresponding regression parameter vectors are $\boldsymbol{\beta}^0 = 1/\sqrt{14}(3, 2, 0, 0, 1, 0, 0)^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0 = (1, 0, 0, -0.5)^{\mathrm{T}}$. Moreover, the random errors $(\varepsilon_{i1}, ..., \varepsilon_{i5})^{\mathrm{T}}$ are generated from a multivariate normal distribution with mean 0, variance σ^2, and an associated exchangeable correlation parameter $\rho = 0.6$. The two standard errors, $\sigma = 0.2, 0.5$, are used for examining the performance of the proposed estimators via the signal-to-noise ratio. To assess the robustness of covariance setting, we finally consider three different working correlation structures: independent (IND), exchangeable (EXCH), and AR(1), although the data are simulated from the exchangeable setting. In this study, we use a cubic spline to estimate the nonparametric function, and assume that the family of candidate knots contains the operating knots (i.e., the nearest representation of the true knots). In the literature, selection criteria for the number of interior knots are usually classified into two categories: consistent (e.g., the Bayesian information criterion BIC) and efficient (e.g., the Akaike information criterion AIC and the cross-validation CV). For the sake of consistency, we follow the approach of Qu and Li (2006) and Xue et al. (2010), and employ BIC to select the number of interior knots N. However, this does not exclude the possibility of using AIC or CV (e.g., see Huang et al. (2002)) to choose the number of knots when the user focuses on efficiency.

Based on the order assumptions on the number of interior knots N in Theorem 1 and Remark 1, we choose N from a given interval $\left[\left[n^{1/(2p+1)}\right], 5\left[n^{1/(2p+1)}\right]\right]$ by minimizing $\mathrm{BIC}(N) = Q_n(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}) + \log(n) \times (N+p)/n$, where $[a]$ denotes the closest integer to a. As a result, $N = 2$ on average among 500 replications for $n = 100, 200$, and 500.

Tables 1 and 2 report the variable selection and estimation results for $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$, respectively. Both tables correspondingly indicate the three and two truly nonzero coefficients of $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ that are correctly set to nonzero. On the other hand, the average numbers of truly nonzero coefficients that are incorrectly set to zero decrease as the sample size increases. In addition, the proportions of models correctly fitted increase with the sample size, while the difference between the PQIF and ORACLE estimators in terms of the squared root MSE measure is diminishing as the sample size increases. The above findings confirm theoretical results under different signal-to-noise ratios. It is noteworthy that although the correct working correlation structure (i.e., EXCH) performs the best, the other two working correlation structures (IND and AR(1)) also yield good performance. This finding corroborates Theorem 4 by showing that the PQIF estimators are consistent even though the working correlation is mis-specified.

Finally, we evaluate the nonparametric estimate of g. Figure 1 respectively depicts the spline estimated functions $\hat{g}\left(\hat{u}\right)$ (solid curve) together with the true nonlinear function $g(u)$ (dashed curve) under the exchangeable, AR(1), and independent error structures, when $n = 100$, $\sigma = 0.2$. It shows that all three estimated curves are fairly similar, and close to the true curve. In sum, our proposed PQIF approach performs well in estimating both parametric and nonparametric components.

5.2 Empirical example

To illustrate the practical usefulness of PQIF, we consider a data set from Frees (2004) that studies the debt maturity structure of a firm. The data contain 328 observations of unregulated firms over the period 1980-1989. The response variable (DEBTMAT) is the value-weighed average of the maturities of the firm's debt that is calculated based on the formula given in Stohs and Mauer (1996). The explanatory variables are defined as follows: LEVERAGE is the ratio of total debt (the sum of long-term debt, long-term debt due within 1 year, and short-term debt) to the market value of the firm; ASSETMAT is the (book) value-weighted average of the maturities of current assets and net property, plant, and equipment; MV/BV is the market value of the firm scaled by the book value of assets; SIZE is the natural logarithm of the estimate of firm value measured in 1982 dollars using the producer price index deflator; CHANGEEPS is the difference between next year's and this year's earnings per share scaled by this year's common stock price per share; The firm's effective tax rate (TAXRATE) is split into two variables GTAXRATE and BTAXRATE, where GTAXRATE=TAXRATE and BTAXRATE= 0 if TAXRATE is between zero and one, and GTAXRATE= 0 and BTAXRATE=TAXRATE otherwise; VAR is the ratio of the standard deviation of the first difference in earnings before interest, depreciation, and taxes to the average of assets over the ten year period (1980-89); TERM is the difference between the long-term and short-term yields on government bonds; LOWBOND is a dummy variable, in which LOWBOND equals one if the firm has a rating of CCC or is unrated, and zero otherwise; and HIGHBOND is also a dummy variable and it equals one if the firm is rated AA or higher, and zero otherwise.

We fit the data with PLSiM by using the standardized continuous variables as the single index components, including X_1 =LEVERAGE, X_2 =ASSETMAT, X_3 =MV/BV, X_4 =SIZE, X_5 =CHNGEEPS, X_6 =GTAXRATE, X_7 = BTAXRATE, X_8 =VAR, and X_9 =TERM. In addition, the two dummy variables, LOWBOND and HIGHBOND, are the linear components, i.e., Z_1 =LOWBOND and Z_2 =HIGHBOND. Since the distribution of DEBTMAT is highly right skewed, we take the natural logarithm transformation of DEBTMAT so that $Y = \log$ (DEBTMAT+ 1). As a result, we have that $E(Y_{ij}|X_{ij}, Z_{ij}) = g\,(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}) + \boldsymbol{\alpha}^{\mathrm{T}} Z_{ij}$, where $X_{ij} = (X_{ij,1}, \ldots, X_{ij,9})^{\mathrm{T}}$,

$Z_{ij} = (Z_{ij,1}, Z_{ij,2})^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ for $i = 1, \cdots, 328$ and $j = 1, \cdots, 10$. We use cubic splines to estimate $g(\cdot)$ and the number of interior knots N is selected by the BIC criterion described in Section 5.1.

Table 3 reports the variable selection and estimation results by using the proposed PQIF method with the exchangeable, AR(1), and independent working correlation structures. In the single index component, four of the six selected variables are common across the three working correlation structures. Among those four variables, LEVERAGE, ASSETMAT, and MV/BV have significant positive effects on DEBTMAT, while VAR has a significant negative effect on DEBTMAT. In the linear component, the signs of coefficient estimates of both dummy variables, LOWBOND and HIGHBOND, are negative. In addition, the p-values of all variables selected by the penalized QIF method are less than 0.05. It is noteworthy that the coefficient signs for LEVERAGE, ASSETMAT, LOWBOND and HIGHBOND are consistent with the agency cost hypothesis given in Stohs and Mauer (1996), while MV/BV shows a different sign from the agency cost hypothesis. To further understand this sign difference, we calculate the correlation coefficient between MV/BV and LEVERAGE. It is -0.458, which shows a moderately negative correlation. This inverse relationship between MV/BV and LEVERAGE may cause MV/BV to have a different sign from the agency cost hypothesis regarding LEVERAGE, by controlling for the LEVERAGE variable in the regression model. This finding is also consistent with Stohs and Mauer (1996).

In addition to the four common variables mentioned above, the variables SIZE and GTAXRATE have been selected in the AR(1) and IND settings and the EX and AR(1) settings, respectively. The coefficient estimates of SIZE are positive, which is consistent with the agency cost hypothesis. In contrast, the positive coefficients of GTAXRATE contradict the agency cost hypothesis (see Stohs and Mauer, 1996). In sum, although the estimated regression coefficients obtained from the three working correlation structures, EXCH, AR(1), and IND, are slightly different, they yield the same significant effect on DEBTMAT. To assess the model fitting, we next calculate the coefficients of determination under the EXCH, AR(1) and IND settings, which are 0.488, 0.495, and 0.475, respectively. Since these values are similar, the model fitting via the PQIF method is robust against the three correlation structures. Based on the coefficient of determination, one may favor AR(1). However, its resulting model yields a contradictory sign of GTAXRATE. Consequently, we slightly prefer IND to EXCH and AR(1).

To study the relationship between the response variable and the aggregate (i.e., linear combination) of continuous variables, Figure 2 depicts estimated nonparametric functions of $\hat{g}\left(\hat{u}\right)$ versus $\hat{u}$ (thick curves) together with their 95% confidence intervals (upper and lower thin curves) for three working correlation settings, where $\hat{u} = \hat{\boldsymbol{\beta}}^{\mathrm{T}} X$. The three functions exhibit a similar pattern and show an increasing trend in general. Specifically, they increase sharply in the beginning, and then become less steep as the index increases. There is a drop towards at higher index values and then they increase again. In conclusion, the relationship between the debt maturity and index is nonlinear rather than identity, after controlling for two dummy variables, and this provides an insightful finding on corporate debt maturity structure.

6 Discussion

In partially linear single-index models with repeated measurements, we employed the quadratic influence function together with the profile approach to obtain parameter estimators for both parametric and nonparametric components. The asymptotic properties of the resulting estimators are established. Furthermore, we proposed a penalized quadratic function approach to select variables and estimate parameters. The resulting PQIF estimator shares the same asymptotic distribution as the oracle. To broaden the usefulness of the proposed approach, we could consider a mixed effects model by assuming that part of $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ has random effects. In practice, incorporating the missing data and measurement errors in model structure is also worth further study. Finally, extending the model structure by including the quasi likelihood function and censor data would enhance the applicability in real data analysis.

References

Bai, Y., Fung, W. and Zhu, Z. (2009). Penalized quadratic inference functions for single-index models with longitudinal data, *Journal of Multivariate Analysis* **100**: 152–161.

Barrow, D. L. and Smith, P. W. (1978). Asymptotic properties of best $l_2\,[0,1]$ approximation by spline with variable knots, *Quarterly of Applied Mathematics* **36**: 293–304.

Bosq, D. (1998). *Nonparametric Statistic for Stochastic Process*, Vol. 10 of *Lecture notes in Statistics*, 2 edn, Springer, New York.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models, *The Annals of Statistics* **17**: 453–510.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models, *Journal of the American Statistical Association* **92**: 477–489.

Chen, J., Gao, J. and Li, D. (2012). Semiparametric trending regression in panel data models with cross–sectional dependence, *Journal of Econometrics* **171**: 71–85.

Chen, J., Gao, J. and Li, D. (2013). Estimation in partially linear single-index panel data models with fixed effects, *Journal of Business and Economic Statistics* **In press**.

Chen, K. and Jin, Z. (2006). Partial linear regression models for clustered data, *Journal of the American Statistical Association* **101**: 195–204.

Cui, X., Härdle, W. and Zhu, L. (2011). The EFM approach for single-index models, *The Annals of Statistics* **39**: 1658–1688.

de Boor, C. (2001). *A Practical Guide to Splines*, Vol. 27 of *Applied Mathematical Sciences*, revised edn, Springer-Verlag, New York.

Demko, S. (1986). Spectral bounds for $|a^{-1}|_\infty$, *Journal of Approximation Theory* **48**: 207–212.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, Vol. 303 of *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, Berlin.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**: 1348–1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *Journal of the American Statistical Association* **99**: 710–723.

Frees, E. W. (2004). *Longitudinal and Panel Data-Analysis and Applications in the Social Sciences*, Cambridge University Press, New York.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer-Verlag, New York.

He, X., Fung, W. and Zhu, Z. (2005). Robust estimation in generalized partial linear models for clustered data, *Journal of the American Statistical Association* **100**: 1176–1184.

Horowitz, J. (1998). *Lecture Notes in Statistics*, Springer, New York.

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.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models, *Journal of Econometrics* **58**: 71–120.

Lai, P., Li, G. and Lian, H. (2013). Quadratic inference functions for partially linear single-index models with longitudinal data, *Journal of Multivariate Analysis* **118**: 115–127.

Liang, H., Liu, X., Li, R. and Tsai, C. L. (2010). Estimation and testing for partially linear single-index models, *The Annals of Statistics* **38**: 3811–3836.

Liang, H. and Wang, N. S. (2005). Partially linear single-index measurement error models, *Statistica Sinica* **15**: 99–116.

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

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

Lin, X., Wang, N., Welsh, A. and Carroll, R. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data, *Biometrika* **91**: 177193.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients, *Econometrica* **57**: 1403–1430.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data, *Biometrics* **62**(2): 379–391.

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**(1): 127–142.

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

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models, *The Annals of Statistics* **20**(4): 1768–1802.

Shen, X., Wolfe, D. A. and Zhou, S. (1998). Local asymptotics for regression splines and confidence regions, *Annals of Statistics* **26**: 1760–1782.

Stohs, M. H. and Mauer, D. C. (1996). The determinants of corporate debt maturity structure, *The Journal of Business* **69**: 279–312.

Su, L. and Ullah, A. (2006). Profile likelihood estimation of partially linear panel data models with fixed effects, *Economics Letters* **92**: 75–81.

Sun, J., Kopciuk, K. A. and Lu, X. (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models, *Journal Computational Statistics & Data Analysis* **53**: 176–188.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method, *Biometrika* **94**: 553–568.

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

Xia, Y. (2006). Asymptotic distributions for two estimators of the singleindex model, *Econometric Theory* **22**: 1112–1137.

Xia, Y. C. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models, *Journal of Multivariate Analysis* **97**: 1162–1184.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models, *Journal of the American Statistical Association* **94**: 1275–1285.

Xia, Y., Tong, H. and Li, W. K. (1999). On extended partially linear single-index models, *Biometrika* **86**: 831–842.

Xue, L., Qu, A. and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data, *Journal of the American Statistical Association* **105**: 1518–1530.

Yu, K., Park, B. and Mammen, E. (2008). Smooth backfitting in generalized additive models, *The Annals of Statistics* **36**: 228–260.

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

Appendix

We begin this appendix by presenting some notation that will be used in the proofs of theorems. For any positive numbers a_n and b_n, $a_n \asymp b_n$ means that $\lim_{n\to\infty} a_n/b_n = c$, where c is a positive constant. In addition, denote the space of the p-th order smooth functions ϕ as $C^{(p)}([0,1]) = \left\{\phi \,\middle|\, \phi^{(p)} \in [0,1]\right\}$. For any two functions ϕ and φ, let $\phi(U_i)$ and $\varphi(U_i)$ be $m \times 1$ vectors. Then, define the empirical inner product and the empirical norm as $\langle \phi, \varphi \rangle_n = n^{-1} \sum_{i=1}^{n} \phi(U_i)^{\mathrm{T}} \varphi(U_i)$ and $\|\phi\|_n^2 = \langle \phi, \phi \rangle_n$, respectively. If functions ϕ and φ are L^2-integrable, we further define the theoretical inner product and theoretical L^2 norm as $\langle \phi, \varphi \rangle = E\left\{\phi(U_i)^{\mathrm{T}} \varphi(U_i)\right\}$ and $\|\phi\|^2 = E\left\{\phi(U_i)^{\mathrm{T}} \phi(U_i)\right\}$, respectively. For $1 \leq r \leq \infty$, denote $\|\zeta\|_r = (|\zeta_1|^r + \cdots + |\zeta_s|^r)^{1/r}$, for any vector $\zeta = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$, and $\|\zeta\|_\infty = \max(|\zeta_1|, \cdots, |\zeta_s|)$. For any symmetric matrix $\boldsymbol{A}$, denote its L_r norm as $\|\boldsymbol{A}\|_r = \max_{\zeta \in R^s, \zeta \neq 0} \|\boldsymbol{A}\zeta\|_r \|\zeta\|_r^{-1}$. For any matrix $\boldsymbol{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\boldsymbol{A}\|_\infty = \max_{1 \leq i \leq s} \sum_{j=1}^{t} |A_{ij}|$. To develop the theoretical results of the proposed estimators, we next present the following technical conditions.

A.1 Six Conditions

(C1) The density function $f_{\boldsymbol{\beta}^{\mathrm{T}} X_{ij}}(\cdot)$ of the random variable $\boldsymbol{\beta}^{\mathrm{T}} X_{ij}$ is bounded away from 0 on $[a, b]$ for β in a neighborhood of β_0 and it satisfies the Lipschitz condition of order 1.

(C2) $g(\cdot) \in C^{(p)}([0,1])$ and $p \geq 3$.

(C3) There exists $0 < c_\xi < \infty$, such that the distances between neighboring knots satisfy

$$\max_{0 \leq J \leq N-1} |h_{J+1} - h_J| = o\left(N^{-1}\right) \text{ and } h / \min_{0 \leq J \leq N} h_J < c_\xi.$$

(C4) The eigenvalues of $\boldsymbol{M}_r$, $1 \leq r \leq k$ are bounded away from 0 and infinity. Let $\Gamma = (\Gamma_{r,r'})_{r,r'=1}^{k} = (\Gamma_{j,j'r,r'})_{j,j'=1,r,r'=1}^{m,k}$. For any $1 \leq j \leq m$, and any given vector $a = (a_r)_{r=1}^{k} \in R^k$, there exist constant $0 < c_\Gamma < C_\Gamma < \infty$, such that $c_\Gamma \sum_{r=1}^{k} a_r^2 \leq \sum_{r,r'=1}^{k} a_r a_{r'} \Gamma_{j,j,r,r'} \leq C_\Gamma \sum_{r=1}^{k} a_r^2$.

(C5) The eigenvalues of $\dot{\Omega}\left(\boldsymbol{\theta}^0\right)$ and $\Xi\left(\boldsymbol{\theta}^0\right)$ are bounded away from 0 and infinity.

(C6) $E\left(X_{ij} \,|\boldsymbol{\beta}^{\mathrm{T}} X_{ij} = u_{ij}\right) \in C^{(1)}([0,1])$ and $E\left(Z_{ij} \,|\boldsymbol{\beta}^{\mathrm{T}} X_{ij} = u_{ij}\right) \in C^{(1)}([0,1])$ for $u_{ij} \in S_\beta$.

It is noteworthy that Condition (C1) is the same as Condition (d) in Cui et al. (2011). Condition (C2) is given in Theorem 2.1 of Shen et al. (1998), and Condition (C6) is weaker than the last one of Condition (a) in Cui et al. (2011). In addition, Condition (C3) is given in equation (3) and Remark 4 of Shen et al. (1998). Finally, Condition (C4) is required for the existence of the asymptotic variances of the estimators, while Condition (C5) is needed for the convergence rates of the parametric and nonparametric estimators. Before proving the proposition and theorems, we provide eight lemmas below.

A.2 Eight Lemmas

For notational simplicity, we denote $U_{ij} = U_{ij}\left(\boldsymbol{\beta}^0\right)$ and $U_i = (U_{i1}, \ldots, U_{im})^{\mathrm{T}}$.

Lemma A.1. *Under Conditions (C1) and (C3), we have that, as $n \to \infty$,*

$$\max_{1-p \le J, J' \le N} \left| \langle B_{J,p}, B_{J',p} \rangle_n - \langle B_{J,p}, B_{J',p} \rangle \right| = O_{a.s.} \left\{ \sqrt{(\log n)\, h/n} \right\}.$$

Proof. Using Bernstein's inequality from Bosq (1998), we can directly prove the result. □

Lemma A.2. *Under Conditions (C1) and (C3), for any $\boldsymbol{a} \in R^{J_n}$, there exist constants $0 < c_1 < C_1 < \infty$ such that*

$$c_1 h \left\| \boldsymbol{a} \right\|_2^2 \le \left\| \boldsymbol{B}_p\left(U_i\right)^{\mathrm{T}} \boldsymbol{a} \right\| \le C_1 h \left\| \boldsymbol{a} \right\|_2^2.$$

Proof. Lemma A.2 follows from Theorem 5.4.2 of DeVore and Lorentz (1993). □

The next lemma can be found in Demko (1986), which plays an important role in the proof of Lemma A.4.

Lemma A.3. *Suppose that $\boldsymbol{A}$ is a positive definite Hermitian matrix $\boldsymbol{A}$ such that $\boldsymbol{A}$ has no more than k nonzero entries in each row. Then, we have that $\left\| \boldsymbol{A}^{-1} \right\|_\infty \le 33\sqrt{k} \left\| \boldsymbol{A}^{-1} \right\|_2^{5/4} \left\| \boldsymbol{A} \right\|_2^{1/4}$.*

Lemma A.4. *Assume Conditions (C1), (C3), and (C4) hold. Then, as $n \to \infty$, there exist constants $0 < C_\Psi < \infty$, $0 < c_2 < C_2 < \infty$, $0 < C_3 < \infty$, and $0 < c_4 < C_4 < \infty$, such that, with probability 1, (i) $c_2 h^{-1} \boldsymbol{I}_{kJ_n} \le n^{-1} \tilde{\Psi}_n^{-1} \le C_2 h^{-1} \boldsymbol{I}_{kJ_n}$, (ii) $\left\| n^{-1} \tilde{\Psi}_n^{-1} \right\|_\infty \le C_\Psi h^{-1}$, (iii) $\left\| \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n \right)^{-1} \right\|_\infty \le C_3 (nh)^{-1}$, and (iv) $c_4 (nh)^{-1} \boldsymbol{I}_{kJ_n} \le \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n \right)^{-1} \le C_4 (nh)^{-1} \boldsymbol{I}_{kJ_n}$, where $\widetilde{\dot{\Phi}}_n$ and $\tilde{\Psi}_n$ are defined in (2.13) and (2.14), respectively.*

Proof. Define $\dot{\Phi} = E\left\{\begin{array}{c} \boldsymbol{B}_p(U_i)\Lambda_1\boldsymbol{B}_p(U_i)^{\mathrm{T}} \\ \vdots \\ \boldsymbol{B}_p(U_i)\Lambda_k\boldsymbol{B}_p(U_i)^{\mathrm{T}} \end{array}\right\}_{kJ_n\times J_n}$ and

$$\Psi = E\left\{\begin{array}{ccc} \boldsymbol{B}_p(U_i)\Gamma_{1,1}\boldsymbol{B}_p(U_i)^{\mathrm{T}} & \cdots & \boldsymbol{B}_p(U_i)\Gamma_{1,k}\boldsymbol{B}_p(U_i)^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{B}_p(U_i)\Gamma_{k,1}\boldsymbol{B}_p(U_i)^{\mathrm{T}} & \cdots & \boldsymbol{B}_p(U_i)\Gamma_{k,k}\boldsymbol{B}_p(U_i)^{\mathrm{T}} \end{array}\right\}_{kJ_n\times kJ_n}.$$

Let $\Gamma_{r,r'} = (\Gamma_{j,j'r,r'})_{j,j'=1}^{m}$. Then, for $1 \le r, r' \le k$, we have that

$$\begin{aligned} E\left\{\boldsymbol{B}_p(U_i)\Gamma_{r,r'}\boldsymbol{B}_p(U_i)^{\mathrm{T}}\right\} &= \left[\sum_{j,j'=1}^{m} E\left\{B_{J,p}(U_{ij})\Gamma_{j,j'r,r'}B_{J',p}(U_{ij'})\right\}\right]_{J,J'=1-p}^{N} \\ &= \Theta_{r,r'}^{(1)} + \Theta_{r,r'}^{(2)}, \end{aligned}$$

where

$$\begin{aligned} \Theta_{r,r'}^{(1)} &= \left[\sum_{j=1}^{m} E\left\{B_{J,p}(U_{ij})\Gamma_{j,jr,r'}B_{J',p}(U_{ij})\right\}\right]_{J,J'=1-p}^{N} \quad \text{and} \\ \Theta_{r,r'}^{(2)} &= \Big[\sum_{j\neq j'} E\big\{B_{J,p}(U_{ij})\Gamma_{j,j'r,r'}B_{J',p}(U_{ij'})\big\}\Big]_{J,J'=1-p}^{N}. \end{aligned}$$

By B-spline properties, $\Theta_{r,r'}^{(1)}$ has no more than $2p-1$ nonzero entries in each row, and $\Theta_{r,r'}^{(2)} = O(h^2)$. Next define $\Psi_1 = \left\{\begin{array}{ccc} \Theta_{1,1}^{(1)} & \cdots & \Theta_{1,k}^{(1)} \\ \vdots & \ddots & \vdots \\ \Theta_{k,1}^{(1)} & \cdots & \Theta_{k,k}^{(1)} \end{array}\right\}_{kJ_n\times kJ_n}$ and $\Psi_2 = \left\{\begin{array}{ccc} \Theta_{1,1}^{(2)} & \cdots & \Theta_{1,k}^{(2)} \\ \vdots & \ddots & \vdots \\ \Theta_{k,1}^{(2)} & \cdots & \Theta_{k,k}^{(2)} \end{array}\right\}_{kJ_n\times kJ_n}$. Accordingly, $\Psi = \Psi_1 + \Psi_2 = \Psi_1 + O(h^2)$, and Ψ_1 has no more than $(2p-1)k$ nonzero entries in each row. By Lemma A.2 and Condition (C4), for any vector $\boldsymbol{a} = (a_{r,J})_{r=1,J=1-p}^{k,N} \in R^{kJ_n}$, we have that

$$\begin{aligned} \boldsymbol{a}^{\mathrm{T}}\Psi_1\boldsymbol{a} &\le C_\Gamma \sum_{j=1}^{m}\sum_{r=1}^{k}\sum_{J,J'=1-p}^{N} a_{r,J}a_{r,J'}E\left\{B_{J,p}(U_{ij})B_{J',p}(U_{ij})\right\} \\ &\le C_\Gamma C_1 h \sum_{j=1}^{m}\sum_{r=1}^{k}\sum_{J=1-p}^{N} a_{r,J}^2 = kC_\Gamma C_1 h\|\boldsymbol{a}\|_2^2 = C'h\|\boldsymbol{a}\|_2^2, \end{aligned}$$

where $C' = kC_\Gamma C_1$. Analogously, we can show that $\boldsymbol{a}^{\mathrm{T}}\Psi_1\boldsymbol{a} \ge c'h\|\boldsymbol{a}\|_2^2$. Therefore, as $n \to \infty$, $c'h\|\boldsymbol{a}\|_2^2 \le \boldsymbol{a}^{\mathrm{T}}\Psi\boldsymbol{a} \le C'h\|\boldsymbol{a}\|_2^2$. Moreover, Lemma A.1 implies that $\left\|\left(n\tilde{\Psi}_n - \Psi\right)\right\|_\infty = O_{a.s.}\left\{\sqrt{(\log n)h/n}\right\}$. Consequently, with probability 1, $c'h\|\boldsymbol{a}\|_2^2 \le \boldsymbol{a}^{\mathrm{T}}n\tilde{\Psi}_n\boldsymbol{a} \le C'h\|\boldsymbol{a}\|_2^2$ when $n \to \infty$. This leads to $c_2h^{-1}\boldsymbol{I}_{kJ_n} \le n^{-1}\tilde{\Psi}_n^{-1} \le C_2h^{-1}\boldsymbol{I}_{kJ_n}$ where $c_2 = C'^{-1}$ and $C_2 = c'^{-1}$. This completes the proof of part (i).

To demonstrate part (ii), we have that

$$\begin{aligned}\|\Psi_1\|_2 &= \sup_{\boldsymbol{a}} \left\{ (\Psi_1 \boldsymbol{a})^{\mathrm{T}} (\Psi_1 \boldsymbol{a}) / \|\boldsymbol{a}\|_2^2 \right\}^{1/2} \\ &\le \sup_{\boldsymbol{a}} \left\{ C' h (\Psi_1 \boldsymbol{a})^{\mathrm{T}} \Psi_1^{-1} (\Psi_1 \boldsymbol{a}) / \|\boldsymbol{a}\|_2^2 \right\}^{1/2} \\ &= C'^{1/2} h^{1/2} \sup_{\boldsymbol{a}} \left\{ \boldsymbol{a}^{\mathrm{T}} \Psi_1 \boldsymbol{a} / \|\boldsymbol{a}\|_2^2 \right\}^{1/2} \le C' h.\end{aligned}$$

Analogously, we can show that $\left\|\Psi_1^{-1}\right\|_2 \le c'^{-1} h^{-1}$. It is also noteworthy that Ψ_1 is a positive definite Hermitian matrix. The above results, together with Lemma A.3, yield

$$\left\|\Psi_1^{-1}\right\|_\infty \le 33\sqrt{(2p-1)} \left\|\Psi_1^{-1}\right\|_2^{5/4} \|\Psi_1\|_2^{1/4} \le C'_\Psi h^{-1},$$

where $C'_\Psi = 33\sqrt{(2p-1)} c'^{-5/4} C'^{1/4}$. Let $\xi = \Psi\eta$, where η is any given $kJ_n \times 1$ vector. Then,

$$\left\|\Psi^{-1}\xi\right\|_\infty \le \left\|\Psi^{-1}\right\|_\infty \|\xi\|_\infty \le C'_\Psi h^{-1} \|\xi\|_\infty .$$

Subsequently, $\|\Psi\eta\|_\infty \ge C'^{-1}_\Psi h \|\eta\|_\infty$. By Lemma A.1, we further have that

$$\left\|\left(n\tilde{\Psi}_n - \Psi\right)\eta\right\|_\infty \le \left\|n\tilde{\Psi}_n - \Psi\right\|_\infty \|\eta\|_\infty = O_{a.s.}\left\{\sqrt{(\log n)\, h/n}\right\} \|\eta\|_\infty .$$

Accordingly, with probability 1, $\left\|n\tilde{\Psi}_n\eta\right\|_\infty \ge (1/2)\, C'^{-1}_\Psi h \|\eta\|_\infty$, as $n \to \infty$. Moreover, let $\xi_1 = n\tilde{\Psi}_n\eta$. Then, with probability 1, $\left\|n^{-1}\tilde{\Psi}_n^{-1}\xi_1\right\|_\infty \le 2C'_\Psi h^{-1} \|\xi_1\|_\infty$ as $n \to \infty$, where $C_\Psi = 2C'_\Psi$. This completes the proof of part (ii).

Using the above results in the proof of part (ii), we have that $\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n \asymp nh^{-1} \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \widetilde{\dot{\Phi}}_n$. This, in conjunction with Lemma A.1, yields $\left\|\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \widetilde{\dot{\Phi}}_n - \dot{\Phi}^{\mathrm{T}}\dot{\Phi}\right\|_\infty = O_{a.s.}\left(h^{3/2} n^{-1/2} \sqrt{\log n}\right)$. Next, let $\Lambda_r = (\Lambda_{j,j'r})_{j,j'=1}^m$. Then, from the definition of $\dot{\Phi}$, we obtain that

$$\begin{aligned}\dot{\Phi}^{\mathrm{T}}\dot{\Phi} &= \sum_{r=1}^k \left[\left[\sum_{j,j'=1}^m E\left\{B_{J,p}(U_{ij}) \Lambda_{j,j'r} B_{J',p}(U_{ij'})\right\}\right]_{J,J'=1-p}^N\right]^2 \\ &= \sum_{r=1}^k \left[\left[\sum_{j=1}^m E\left\{B_{J,p}(U_{ij}) \Lambda_{j,j'r} B_{J',p}(U_{ij})\right\} + O\left(h^2\right)\right]_{J,J'=1-p}^N\right]^2 \\ &= \widetilde{\dot{\Phi}}^{\mathrm{T}} \widetilde{\dot{\Phi}} + O\left(h^3\right),\end{aligned}$$

where

$$\widetilde{\dot{\Phi}}^{\mathrm{T}} \widetilde{\dot{\Phi}} = \sum_{r=1}^k \left[\left[\sum_{j=1}^m E\left\{B_{J,p}(U_{ij}) \Lambda_{j,jr} B_{J',p}(U_{ij})\right\}\right]_{J,J'=1-p}^N\right]^2 .$$

By B-spline properties, $\widetilde{\dot{\Phi}}^{\mathrm{T}} \widetilde{\dot{\Phi}}$ has no more than $4p-3$ nonzero entries in each row. Then, applying the same techniques used in the proof of Part (ii), we are able to show that there exists a constant $0 < C_3' < \infty$ such that $\|(\widetilde{\dot{\Phi}}^{\mathrm{T}}\widetilde{\dot{\Phi}})^{-1}\|_\infty \le C_3' h^{-2}$. This, together with the above result, implies that, $\|(\dot{\Phi}^{\mathrm{T}}\dot{\Phi})^{-1}\|_\infty \le C_3' h^{-2}\{1+o(1)\}$, Accordingly, with probability 1, $\|(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\widetilde{\dot{\Phi}}_n)^{-1}\|_\infty \le C_3' h^{-2}\{1+o(1)\}$, as $n\to\infty$. Subsequently, using the result that $\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}, \widetilde{\dot{\Phi}}_n \asymp nh^{-1}\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\widetilde{\dot{\Phi}}_n$, there is a constant $0 < C_3 < \infty$, such that, with probability 1, $\|(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n)^{-1}\|_\infty \le C_3 n^{-1}h^{-1}$, when $n\to\infty$. This completes the proof of part (iii). Finally, we can employ the similar techniques used in part (i) to prove part (iv). □

To find the asymptotic approximation of the nonparametric function estimation $\hat{g}$, we next introduce a lemma obtained from equation (2.7) of Barrow and Smith (1978).

Lemma A.5. *Under Conditions (C1)-(C3), we have that, for any $u \in I_J$ and $0 \le J \le N$, there exists $\boldsymbol{\gamma}^0 \in R^{J_n}$ such that $g(u) - \left\{B_p(u)^{\mathrm{T}}\boldsymbol{\gamma}^0 - b^*(u)h^p\right\} = o(h^p)$, where $b^*(u)$ is given in (2.15).*

Before using the above lemma, we first have that $\left|\Psi_n(\boldsymbol{\gamma}^0) - \tilde{\Psi}_n\right| = o_p(n^{-1})$, where $\tilde{\Psi}_n$ is given in (2.14). Let $\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}} = \hat{\boldsymbol{\gamma}}^{\mathrm{QIF}}(\boldsymbol{\theta}^0)$, where $\boldsymbol{\theta}^0 = \left(\boldsymbol{\beta}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}}$. By similar arguments as given in Qu and Li (2006), we can prove that $\|\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}} - \gamma^0\|_\infty = o_{\mathrm{a.s.}}(1)$. Then, by the Taylor expansion,

$$\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}} - \boldsymbol{\gamma}^0 = \left\{\ddot{Q}_n(\boldsymbol{\gamma}^0)\right\}^{-1}\left\{\dot{Q}_n(\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}}) - \dot{Q}_n(\boldsymbol{\gamma}^0)\right\}\{1+o_p(1)\},$$

where $\ddot{Q}_n(\boldsymbol{\gamma}^0) = \left\{\partial Q_n(\boldsymbol{\gamma}^0)/\partial\boldsymbol{\gamma}^0\partial(\boldsymbol{\gamma}^0)^{\mathrm{T}}\right\}$ and it is asymptotically equivalent to $2\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n$. Since $\dot{Q}_n(\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}}) = 0$, we obtain that

$$\begin{aligned}\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}} - \boldsymbol{\gamma}^0 &= -\left(2\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n\right)^{-1}\dot{Q}_n(\boldsymbol{\gamma}^0)\{1+o_p(1)\}\\ &= -\left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n\right)^{-1}\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\Phi_n(\boldsymbol{\gamma}^0)\{1+o_p(1)\} = \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n\right)^{-1}\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\times\\ &\quad n^{-1}\sum_{i=1}^n\begin{bmatrix}\boldsymbol{B}_p(U_i)\Lambda_{i,1}\left\{Y_i - \boldsymbol{B}_p(U_i)^{\mathrm{T}}\boldsymbol{\gamma}^0 - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}\right\}\\ \vdots\\ \boldsymbol{B}_p(U_i)\Lambda_{i,k}\left\{Y_i - \boldsymbol{B}_p(U_i)^{\mathrm{T}}\boldsymbol{\gamma}^0 - \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}\right\}\end{bmatrix}\{1+o_p(1)\},\end{aligned} \tag{A.1}$$

where $g(U_i) = \{g(U_{i1}),\ldots,g(U_{im_i})\}^{\mathrm{T}}$. This, together with Lemma A.5, leads to $\hat{\boldsymbol{\gamma}}^{\mathrm{QIF}} = \left(\tilde{\boldsymbol{r}}_e^{\mathrm{QIF}} + \tilde{\boldsymbol{r}}_g^{\mathrm{QIF}}\right)\{1+o_p(1)\}$, where

$$\tilde{\boldsymbol{r}}_e^{\mathrm{QIF}} = \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}\widetilde{\dot{\Phi}}_n\right)^{-1}\widetilde{\dot{\Phi}}_n^{\mathrm{T}}\tilde{\Psi}_n^{-1}n^{-1}\sum_{i=1}^n\left\{\begin{array}{c}\boldsymbol{B}_p(U_i)\Lambda_{i,1}e_i\\ \vdots\\ \boldsymbol{B}_p(U_i)\Lambda_{i,k}e_i\end{array}\right\}, \tag{A.2}$$

$$\tilde{\boldsymbol{r}}_g^{\text{QIF}} = \boldsymbol{\gamma}^0 - \left(\widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} \vartheta_n h^p + \tau_n(u) o\left(h^p\right), \tag{A.3}$$

ϑ_n is defined in (2.15), and

$$\tau_n(u) = \left(\widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} n^{-1} \sum_{i=1}^{n} \left\{ \begin{array}{c} \boldsymbol{B}_p\left(U_i\right) \Lambda_{i,1} 1_m \\ \vdots \\ \boldsymbol{B}_p\left(U_i\right) \Lambda_{i,k} 1_m \end{array} \right\}.$$

Let $\boldsymbol{\theta}^0 = \left\{\left(\boldsymbol{\beta}^0\right)^{\text{T}}, \left(\boldsymbol{\alpha}^0\right)^{\text{T}}\right\}^{\text{T}}$ be the vector of the true parameters. Accordingly, $\tilde{g}\left(u, \boldsymbol{\theta}^0\right) = \tilde{g}(u) = \left\{\tilde{g}_e(u) + \tilde{g}_g(u)\right\}\left\{1 + o_p(1)\right\}$, where $\tilde{g}_e(u) = B_p\left(u\right)^{\text{T}} \tilde{\boldsymbol{r}}_e^{\text{QIF}}$ and $\tilde{g}_g(u) = B_p\left(u\right)^{\text{T}} \tilde{\boldsymbol{r}}_g^{\text{QIF}}$. The next three lemmas present some asymptotic properties of $\tilde{g}_e$ and $\tilde{g}_g$.

Lemma A.6. *Under Conditions (C1), (C3), and (C4),* $\left\{\text{Var}\left(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}\right)\right\}^{-1/2} \tilde{g}_e(u) \to N\left(0, \mathbf{1}\right)$ *as* $n \to \infty$*, where* $\text{Var}\left(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}\right) = \boldsymbol{B}_p\left(u\right)^{\text{T}} \left(\dot{\Phi}_n^{\text{T}} \tilde{\Psi}_n^{-1} \dot{\Phi}_n\right)^{-1} \boldsymbol{B}_p(u)$, $\mathbb{X}= (\boldsymbol{X}_1^{\text{T}}, \ldots, \boldsymbol{X}_n^{\text{T}})^{\text{T}}$, *and* $\mathbb{Z}= (\boldsymbol{Z}_1^{\text{T}}, \ldots, \boldsymbol{Z}_n^{\text{T}})^{\text{T}}$.

Proof. By the definition of $\tilde{g}_e(u)$ and (A.2), it can be written as $\tilde{g}_e(u) = \sum_{i=1}^n \mathbf{a}_i^{\text{T}} \epsilon_i$, in which

$$\mathbf{a}_i^{\text{T}} \mathbf{a}_i = n^{-2} B_p(u)^{\text{T}} \left(\dot{\Phi}_n^{\text{T}} \tilde{\Psi}_n^{-1} \dot{\Phi}_n\right)^{-1} \dot{\Phi}_n^{\text{T}} \tilde{\Psi}_n^{-1} \times$$

$$\left\{ \begin{array}{ccc} \mathbf{B}_p\left(U_i\right) \Gamma_{1,1} \mathbf{B}_p\left(U_i\right)^{\text{T}} & \cdots & \mathbf{B}_p\left(U_i\right) \Gamma_{1,k} \mathbf{B}_p\left(U_i\right)^{\text{T}} \\ \vdots & \ddots & \vdots \\ \mathbf{B}_p\left(U_i\right) \Gamma_{k,1} \mathbf{B}_p\left(U_i\right)^{\text{T}} & \cdots & \mathbf{B}_p\left(U_i\right) \Gamma_{k,k} \mathbf{B}_p\left(U_i\right)^{\text{T}} \end{array} \right\} \tilde{\Psi}_n^{-1} \dot{\Phi}_n \left(\dot{\Phi}_n^{\text{T}} \tilde{\Psi}_n^{-1} \dot{\Phi}_n\right)^{-1} B_p(u),$$

and conditional on $(\mathbb{X},\mathbb{T})$, ϵ_i are independent with mean 0 and covariance $\mathbf{I}_m$. Hence, the proof of this lemma follows if the Lindeberg-Feller conditions are satisfied. To this end, it suffices to verify that $\max_{1\le i\le n} \boldsymbol{a}_i^{\text{T}} \boldsymbol{a}_i = o_{a.s.}\left(\sum_{i=1}^n \boldsymbol{a}_i^{\text{T}} \boldsymbol{a}_i\right)$. Let 1_{J_n} be a $J_n \times 1$ vector with elements 1. By Lemma A.4, we have that, with probability 1,

$$\begin{aligned} \max_{1\le i\le n} \boldsymbol{a}_i^{\text{T}} \boldsymbol{a}_i \;\; \le \;\; & n^{-2} \left\|B_p(u)\right\|_\infty^2 \left\| \left(\widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \right\|_\infty^2 \left\|\widetilde{\dot{\Phi}}_n^{\text{T}}\right\|_\infty \left\|\tilde{\Psi}_n^{-1} 1_{J_n}\right\|_\infty^2 \left\|\widetilde{\dot{\Phi}}_n\right\|_\infty \\ & \times \max_{1\le i\le n} \max_{J,J',r,r'} \left| \sum_{j,j'=1}^{m} B_{J,p}\left(U_{ij}\right) \Gamma_{j,j'r,r'} B_{J,p}\left(U_{ij'}\right) \right| \\ \le \;\; & n^{-2} C C_3^2 n^{-2} h^{-2} h C_\Psi^2 n^2 h^{-2} = O\left(n^{-2} h^{-3}\right), \end{aligned}$$

as $n \to \infty$. After algebraic simplification, $\text{Var}\left(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}\right) = B_p\left(u\right)^{\text{T}} \left(\widetilde{\dot{\Phi}}_n^{\text{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} B_p(u)$. By (iv) in Lemma A.4, we have that, for any $u \in [0,1]$, $\text{Var}\left(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}\right) \le C_4\left(nh\right)^{-1} \left\|B_p(u)\right\|_2^2 \le$

$C_4'(nh)^{-1}$, and $\operatorname{Var}(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}) \geq c_4 (nh)^{-1} \|B_p(u)\|_2^2 \geq c_4'(nh)^{-1}$, where $C_4' = C_4 \|B_p(u)\|_2^2$ and $c_4' = c_4 \|B_p(u)\|_2^2$. As a result, for any $u \in [0,1]$,

$$\sum\nolimits_{i=1}^{n} \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{a}_i = \operatorname{Var}(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}) \asymp (nh)^{-1}. \tag{A.4}$$

Consequently, by Condition (C4), $\max_{1\leq i\leq n} \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_i = o_{a.s.}(\sum_{i=1}^{n} \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{a}_i)$. We complete the proof. □

Lemma A.7. *Under Conditions (C1), (C3), and (C4), we have that, with probability* 1*, there exist constants* $0 < c_\sigma < C_\sigma < \infty$*, such that*

$$c_\sigma (nh)^{-1} \leq \sup_{u\in[0,1]} \operatorname{Var}(\tilde{g}_e(u)\,|\mathbb{X},\mathbb{Z}) \leq C_\sigma (nh)^{-1},$$

as $n \to \infty$*. As a consequence,* $\tilde{g}_e(u) = O_p\left\{(nh)^{-1}\right\}$ *uniformly in* $u \in [0,1]$*.*

Proof. Lemma A.7 follows immediately from (A.4). □

Lemma A.8. *Under Conditions (C1)-(C4), we have that, for any* $u \in I_J$ *and* $0 \leq J \leq N$*,*

$$\tilde{g}_g(u) - g(u) = b^*(u)h^p - B_p(u)^{\mathrm{T}} \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \vartheta_n h^p + o_p(h^p),$$

where $\dot{\Phi}$*,* $\tilde{\Psi}_n$*,* $b^*(u)$*, and* ϑ_n *are given in (2.13), (2.14) and (2.15), respectively.*

Proof. According to the definition of $\tilde{g}_g(u)$, equation (A.3), and Lemma A.5, we have that, for any $u \in I_J$ and $0 \leq J \leq N$,

$$\begin{aligned}
\tilde{g}_g(u) &= B_p(u)^{\mathrm{T}} \gamma^0 - B_p(u)^{\mathrm{T}} \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \vartheta_n h^p + B_p(u)^{\mathrm{T}} \tau_n(u) o(h^p) \\
&= g(u) + b^*(u)h^p - B_p(u)^{\mathrm{T}} \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \vartheta_n h^p \\
&\quad + B_p(u)^{\mathrm{T}} \tau_n(u) o(h^p) + o(h^p).
\end{aligned}$$

In addition, by Lemma A.4, we obtain that, for any $u \in [0,1]$,

$$\left|B_p(u)^{\mathrm{T}} \tau_n(u)\right| \leq \|B_p(u)\|_\infty \left\| \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \right\|_\infty \left\| \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \right\|_\infty \left\| \tilde{\Psi}_n^{-1} \right\|_\infty O_p(h) = O_p(1).$$

As a result,

$$\tilde{g}_g(u) = g(u) + b^*(u)h^p - B_p(u)^{\mathrm{T}} \left(\widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \widetilde{\dot{\Phi}}_n\right)^{-1} \widetilde{\dot{\Phi}}_n^{\mathrm{T}} \tilde{\Psi}_n^{-1} \vartheta_n h^p + o_p(h^p),$$

which completes the proof. □

Next we establish the asymptotic convergence rate of $\tilde{g}\left(u, \boldsymbol{\theta}^0\right)$ and $\tilde{g}'\left(u, \boldsymbol{\theta}^0\right)$, which will be used in the proof of Theorem 1.

Proposition A.1. *Under Conditions (C1)-(C4) given in the Appendix A.1, and $N \to \infty$ and $nN^{-3} \to \infty$, as $n \to \infty$,*

(i) $\left|\tilde{g}\left(u, \boldsymbol{\theta}^0\right) - g(u)\right| = O_p\left\{(nh)^{-1/2} + h^p\right\}$ *uniformly in* $u \in [0,1]$*;*

(ii) $\left|\tilde{g}'\left(u, \boldsymbol{\theta}^0\right) - g'(u)\right| = O_p\left(n^{-1/2}h^{-3/2} + h^{p-1}\right)$ *uniformly in* $u \in [0,1]$.

Proof. The results of (i) in Proposition A.1 follow immediately from Lemmas A.6-A.8. Using the fact that $g'(u)$ is approximated by the spline functions with one order less than that of g, we have $\tilde{g}'(u) = B_{p-1}(u)^{\mathrm{T}} D_1 \hat{\boldsymbol{\gamma}}^{\mathrm{QIF}}$. Then, employing similar techniques to those used in the proofs of Lemma A.8 and (A.4), we obtain that, with probability approaching 1, the bias and variance of $\hat{g}'(u)$ are of orders $O\left(h^{p-1}\right)$ and $O\left\{\left(n^{-1/2}h^{-3/2}\right)^2\right\}$, respectively. Accordingly, part (ii) holds. □

A.3 Proof of Theorem 1

Let $\boldsymbol{\theta}^0 = \left[\beta_1^0, \left\{\theta^{(1)0}\right\}^{\mathrm{T}}\right]^{\mathrm{T}}$. The root-n consistency of $\hat{\boldsymbol{\theta}}^{(1)\mathrm{QIF}}$ can be proved by following similar reasoning as in Ichimura (1993). In the following, we will prove the asymptotic normality. Applying the Taylor expansion, we obtain that

$$\hat{\boldsymbol{\theta}}^{(1)\mathrm{QIF}} - \boldsymbol{\theta}^{(1)0} = -\left\{\ddot{Q}_n^*\left(\boldsymbol{\theta}^0\right)\right\}^{-1} \dot{Q}_n^*\left(\boldsymbol{\theta}^0\right)\left\{1 + o_p(1)\right\},$$

where $\ddot{Q}_n^*\left(\boldsymbol{\theta}^0\right) = \left\{\partial Q_n^*\left(\boldsymbol{\theta}^0\right)/\partial \boldsymbol{\theta}^{(1)0} \partial\left(\boldsymbol{\theta}^{(1)0}\right)^{\mathrm{T}}\right\}$ is a $(d_1 - 1 + d_2) \times (d_1 - 1 + d_2)$ matrix and $\dot{Q}_n^*\left(\boldsymbol{\theta}^0\right) = \left\{\partial Q_n^*\left(\boldsymbol{\theta}^0\right)/\partial \boldsymbol{\theta}^{(1)0}\right\}$ is a $(d_1 - 1 + d_2) \times 1$ vector. By the definition of $Q_n^*\left(\boldsymbol{\theta}^0\right)$ in (2.8), we further have that

$$\dot{Q}_n^*\left(\boldsymbol{\theta}^0\right) = 2\dot{\Omega}_n\left(\boldsymbol{\theta}^0\right)^{\mathrm{T}} \Xi_n\left(\boldsymbol{\theta}^0\right)^{-1} \Omega_n\left(\boldsymbol{\theta}^0\right) + O_p\left(n^{-1}\right), \tag{A.5}$$

where $\dot{\Omega}_n\left(\boldsymbol{\theta}^0\right) = \left\{\partial \Omega_n\left(\boldsymbol{\theta}^0\right)/\partial \boldsymbol{\theta}^{(1)0}\right\}$ is a $(d_1 - 1 + d_2) \times k(d_1 - 1 + d_2)$ matrix. Moreover, Condition (C6) yields $\left|\hat{X}_{ij} - \tilde{X}_{ij}\right| = O_p\left\{(nh)^{-1/2} + h\right\}$ and $\left|\hat{Z}_{ij} - \tilde{Z}_{ij}\right| = O_p\left\{(nh)^{-1/2} + h\right\}$.

These results together with Proposition A.1 imply that

$$\left|\hat{D}_i\left(\boldsymbol{\theta}^0\right)-D_i\left(\boldsymbol{\beta}^0\right)\right|=O_p\left\{(nh)^{-1/2}+h+\left(nh^3\right)^{-1/2}+h^{p-1}\right\}=O_p\left\{h+\left(nh^3\right)^{-1/2}\right\}. \tag{A.6}$$

Define $\widetilde{\Omega}_n\left(\boldsymbol{\beta}^0\right) = -\frac{1}{n}\sum_{i=1}^n\left\{\begin{array}{c} D_i\left(\boldsymbol{\beta}^0\right)\Lambda_1 e_i \\ \vdots \\ D_i\left(\boldsymbol{\beta}^0\right)\Lambda_k e_i\end{array}\right\}$. According to Condition (C3), we have that $h \asymp N^{-1}$. Then, the assumption on N in the theorem implies that $nh^{2p+2}\to 0$ and $nh^4\to\infty$ as $n\to\infty$. As a result, one has

$$\begin{aligned}
&\Omega_n\left(\boldsymbol{\theta}^0\right)-\widetilde{\Omega}_n\left(\boldsymbol{\beta}^0\right)\\
=\ &\frac{1}{n}\sum_{i=1}^n\left\{\begin{array}{c}\left\{D_i\left(\boldsymbol{\beta}^0\right)\right\}\Lambda_1\left\{\tilde{g}\left(U_i\right)-g\left(U_i\right)\right\}\\ \vdots \\ \left\{D_i\left(\boldsymbol{\beta}^0\right)\right\}\Lambda_k\left\{\tilde{g}\left(U_i\right)-g\left(U_i\right)\right\}\end{array}\right\}+O_p\left\{h+\left(nh^3\right)^{-1/2}\right\}O_p\left\{h^p+(nh)^{-1/2}\right\}\\
=\ &O_p\left(n^{-1/2}\right)O_p\left\{h^p+(nh)^{-1/2}\right\}+O_p\left\{h+\left(nh^3\right)^{-1/2}\right\}O_p\left\{h^p+(nh)^{-1/2}\right\}\\
=\ &O_p\left(h^{p+1}+n^{-1/2}h^{1/2}+n^{-1/2}h^{p-3/2}+n^{-1}h^{-2}\right)=o_p\left(n^{-1/2}\right).
\end{aligned}$$

By (A.6), one has

$$\begin{aligned}
\dot{\Omega}_n\left(\boldsymbol{\theta}^0\right) &= \widetilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right)+O_p\left\{\left(nh^3\right)^{-1/2}+h\right\},\\
\Xi_n\left(\boldsymbol{\theta}^0\right) &= \widetilde{\Xi}_n\left(\boldsymbol{\theta}^0\right)+O_p\left[n^{-1}\left\{\left(nh^3\right)^{-1/2}+h\right\}\right],
\end{aligned} \tag{A.7}$$

where

$$\widetilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right)=\frac{1}{n}\sum_{i=1.}^n\left\{\begin{array}{c} D_i\left(\boldsymbol{\beta}^0\right)\Lambda_1 D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\\ \vdots \\ D_i\left(\boldsymbol{\beta}^0\right)\Lambda_k D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\end{array}\right\},$$

$$\widetilde{\Xi}_n\left(\boldsymbol{\theta}^0\right)=n^{-2}\sum_{i=1}^n\left\{\begin{array}{ccc} D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{1,1}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} & \cdots & D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{1,k}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\\ \vdots & \ddots & \vdots \\ D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{k,1}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} & \cdots & D_i\left(\boldsymbol{\beta}^0\right)\Gamma_{k,k}D_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\end{array}\right\}.$$

Thus

$$\begin{aligned}
\ddot{Q}_n\left(\boldsymbol{\theta}^0\right) &= 2\dot{\Omega}_n\left(\boldsymbol{\theta}^0\right)^{\mathrm{T}}\Xi_n\left(\boldsymbol{\theta}^0\right)^{-1}\dot{\Omega}_n\left(\boldsymbol{\theta}^0\right)+o_p(1)\\
&= 2\widetilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right)^{\mathrm{T}}\widetilde{\Xi}_n\left(\boldsymbol{\theta}^0\right)^{-1}\widetilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right)+o_p(1).
\end{aligned}$$

Accordingly,

$$\hat{\boldsymbol{\theta}}^{(1)\text{QIF}} - \boldsymbol{\theta}^{(1)0} = \left\{ \widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right)^{\mathrm{T}} \widetilde{\Xi}_n \left(\boldsymbol{\theta}^0\right)^{-1} \widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right) \right\}^{-1} \times \\ \widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right)^{\mathrm{T}} \widetilde{\Xi}_n \left(\boldsymbol{\theta}^0\right)^{-1} \widetilde{\Omega}_n \left(\boldsymbol{\beta}^0\right) \{1 + o_p(1)\}.$$

Note that $\widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right) = \dot{\Omega}\left(\boldsymbol{\beta}^0\right) + O_p\left(n^{-1/2}\right)$ and $n\widetilde{\Xi}_n \left(\boldsymbol{\theta}^0\right) = \Xi\left(\boldsymbol{\beta}^0\right) + O_p\left(n^{-1/2}\right)$, where $\dot{\Omega}\left(\boldsymbol{\beta}^0\right)$ and $\Xi\left(\boldsymbol{\beta}^0\right)$ are defined in (2.11) and (2.12), respectively. Thus, $\widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right)^{\mathrm{T}} \widetilde{\Xi}_n \left(\boldsymbol{\theta}^0\right)^{-1} \widetilde{\dot{\Omega}}_n \left(\boldsymbol{\theta}^0\right) = n\Sigma_{\boldsymbol{\theta}^{(1)0}} + O_p\left(n^{-1/2}\right)$. By Lindeberg-Feller Central Limit Theorem and Condition (C5), we have that $\sqrt{n}(\hat{\boldsymbol{\theta}}^{(1)\text{QIF}} - \boldsymbol{\theta}^{(1)0}) \rightarrow N(0, \Sigma^{-1}_{\boldsymbol{\theta}^{(1)0}})$, as $n \rightarrow \infty$. This completes the proof.

A.4 Proof of Theorem 3

Theorem 3 follows from Lemmas A.6-A.8 and the fact that $\left\|\hat{\boldsymbol{\theta}}^{\text{QIF}} - \boldsymbol{\theta}^0\right\|_2 = O_p\left(n^{-1/2}\right)$.

A.5 Proof of Theorem 4

The proof of this theorem consists of three steps. Step I establishes the convergence rate of $\{(\hat{\boldsymbol{\beta}}^{\text{PQIF}})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{\text{PQIF}})^{\mathrm{T}}\}$; Step II shows the sparsity of $\left\{\left(\hat{\boldsymbol{\beta}}^{\text{PQIF}}\right)^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{\text{PQIF}})^{\mathrm{T}}\right\}$; Step III demonstrates the asymptotic distribution of the penalized estimators.

Step I. Let $\kappa_n = n^{-1/2} + a_n + c_n$, where $a_n = \max_{2\le l\le d_1}\left\{\left|p'_{\lambda_{1l}}\left(|\beta_l^0|\right)\right|, \beta_l^0 \neq 0\right\}$, $c_n = \max_{1\le s\le d_2}\left\{\left|p'_{\lambda_{2s}}\left(|\alpha_s^0|\right)\right|, \alpha_s^0 \neq 0\right\}$, and β_l^0 and α_s^0 are the l-th and s-th elements of $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$, respectively, for $l = 2, \ldots, d_1$ and $s = 1, \ldots, d_2$. In addition, let $\tilde{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^{(1)0} + \kappa_n v_1$, $\tilde{\beta}_1 = \sqrt{1 - \left\|\tilde{\boldsymbol{\beta}}^{(1)}\right\|_2^2}$, $\tilde{\boldsymbol{\beta}} = \left\{\tilde{\beta}_1, \left(\tilde{\boldsymbol{\beta}}^{(1)}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$, and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^0 + \kappa_n v_2$, where $v_1 = (v_{12}, \ldots, v_{1d_1})^{\mathrm{T}}$, $v_2 = (v_{21}, \ldots, v_{2d_2})^{\mathrm{T}}$, and $\|v_1\|_2 = \|v_2\|_2 = C$ for some positive constant C. Thus,

$$Q_n^*\left(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}\right) - Q_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \begin{pmatrix} \tilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)0} \\ \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0 \end{pmatrix}^{\mathrm{T}} \dot{Q}_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) \\ + \frac{1}{2}\begin{pmatrix} \tilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0 \\ \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0 \end{pmatrix}^{\mathrm{T}} \ddot{Q}_n^*\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*\right) \begin{pmatrix} \tilde{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^0 \\ \tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0 \end{pmatrix}, \tag{A.8}$$

for some $\left\{(\boldsymbol{\beta}^*)^{\mathrm{T}}, (\boldsymbol{\alpha}^*)^{\mathrm{T}}\right\}^{\mathrm{T}}$ that lies between $\left\{(\boldsymbol{\beta}^0)^{\mathrm{T}}, (\boldsymbol{\alpha}^0)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\left(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}, \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$. Applying (A.5) and

(A.7), we obtain that

$$\dot{Q}_n^*\left(\boldsymbol{\theta}^0\right) = -2\tilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right)^{\mathrm{T}} \tilde{\Xi}_n\left(\boldsymbol{\theta}^0\right)^{-1} \frac{1}{n}\sum_{i=1}^{n}\left\{\begin{array}{c} D_i\left(\boldsymbol{\beta}^0\right)\Lambda_1 e_i \\ \vdots \\ D_i\left(\boldsymbol{\beta}^0\right)\Lambda_k e_i \end{array}\right\}\{1+o_p(1)\} = O_p\left(n^{1/2}\right),$$

$$\text{and } \ddot{Q}_n^*\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*\right) = 2\tilde{\dot{\Omega}}_n\left(\theta^0\right)^{\mathrm{T}} \tilde{\Xi}_n\left(\boldsymbol{\theta}^0\right)^{-1} \tilde{\dot{\Omega}}_n\left(\boldsymbol{\theta}^0\right) + o_p(1) = O_p(n).$$

As a result, the first term on the right-hand side of (A.8) is of order $O_p\{Cn^{1/2}(n^{-1/2}+a_n+c_n)\}$ and the second term $\asymp C^2 n(n^{-1/2}+a_n+c_n)^2$. From the Taylor expansion and the Cauchy-Schwarz inequality, as $n \to \infty$, we further have that

$$\begin{aligned} &\left| n\sum_{l=2}^{d_1} p_{\lambda_{1l}}\left(\left|\tilde{\beta}_l\right|\right) - n\sum_{l=2}^{d_1} p_{\lambda_{1l}}\left(\left|\beta_l^0\right|\right) + n\sum_{s=1}^{d_2} p_{\lambda_{2s}}\left(\left|\tilde{\alpha}_s\right|\right) - n\sum_{s=1}^{d_2} p_{\lambda_{2s}}\left(\left|\alpha_s^0\right|\right)\right| \\ \leq\ & n\left(\sqrt{d_{10}}\kappa_n a_n \left\|v_1\right\|_2 + \kappa_n^2 b_n \left\|v_1\right\|_2^2 + \sqrt{d_{20}}\kappa_n c_n \left\|v_2\right\|_2 + \kappa_n^2 d_n \left\|v_2\right\|_2^2\right) \\ \leq\ & nC\kappa_n^2\left(\sqrt{d_{10}} + b_n C + \sqrt{d_{20}} + d_n C\right), \end{aligned} \tag{A.9}$$

where $b_n = \max_{2\leq l\leq d_1}\left\{\left|p''_{\lambda_{1l}}\left(\left|\beta_l^0\right|\right)\right|, \beta_l^0 \neq 0\right\}$ and $d_n = \max_{1\leq s\leq d_2}\left\{\left|p''_{\lambda_{2s}}\left(\left|\alpha_s^0\right|\right)\right|, \alpha_l^0 \neq 0\right\}$. When $b_n \to 0$, $d_n \to 0$, and C is sufficiently large, the second term on the right-hand side of (A.8) dominates its first term and (A.9). Thus, for any give $\nu > 0$, there exists a large constant C such that,

$$P\left\{\inf_{V_{12}} \mathcal{L}\left(\boldsymbol{\beta}^0 + \kappa_n v_1, \boldsymbol{\alpha}^0 + \kappa_n v_2\right) > \mathcal{L}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\} \geq 1-\nu,$$

as $n \to \infty$, where $V_{12} = \{(v_1, v_2) : \|v_1\| = C \text{ and } \|v_2\| = C\}$. Accordingly, the rate of convergence of $\left\{\left(\hat{\boldsymbol{\beta}}^{\text{PQIF}}\right)^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{\text{PQIF}})^{\mathrm{T}}\right\}$ is $O_p\left(n^{-1/2}+a_n+c_n\right)$. Moreover, under the assumptions that $\lambda_{1l} \to 0$ and $\lambda_{2s} \to 0$ for all $1 \leq l \leq d_1$ and $1 \leq s \leq d_2$, we have that $a_n = 0$ and $c_n = 0$. Consequently, the rate of convergence of $\left\{\left(\hat{\boldsymbol{\beta}}^{\text{PQIF}}\right)^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}^{\text{PQIF}})^{\mathrm{T}}\right\}$ is $O_p\left(n^{-1/2}\right)$.

Step II. Let $\boldsymbol{\beta}_1 = \left\{\beta_1, \left(\boldsymbol{\beta}_1^{(1)}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1$ satisfy $\left\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^0\right\| = O_p\left(n^{-1/2}\right)$ and $\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_1^0\| = O_p\left(n^{-1/2}\right)$, respectively. We then show, with probability tending to 1, that

$$\mathcal{L}\left\{\left(\begin{array}{c}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{array}\right), \left(\begin{array}{c}\boldsymbol{\alpha}_1 \\ \mathbf{0}\end{array}\right)\right\} = \min_{\mathcal{C}} \mathcal{L}\left\{\left(\begin{array}{c}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{array}\right), \left(\begin{array}{c}\boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2\end{array}\right)\right\}, \tag{A.10}$$

as $n \to \infty$, where $\mathcal{C} = \left\{\|\boldsymbol{\beta}_2\| \leq C^* n^{-1/2}, \|\boldsymbol{\alpha}_2\| \leq C^* n^{-1/2}\right\}$ and C^* is a positive constant. To this end, consider $\beta_l \in \left(-C^* n^{-1/2}, C^* n^{-1/2}\right)$ for $l = d_{10}+1, \ldots, d_1$. When $\beta_l \neq 0$, one has

$\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_l = \partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_l + n p'_{\lambda_{2l}}(|\beta_l|) \operatorname{sgn}(\beta_l)$, where

$$\partial \left\{ \frac{1}{2} Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) \right\} / \partial \beta_l = \frac{1}{n} \sum_{i=1}^{n} \left[\left[\beta_l \tilde{g}'(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \left\{ X_{ij,1} - \hat{E}(X_{ij,1} | \boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \right\} + \right.\right.$$

$$\left.\left. \sum_{l'=2}^{d_1} \tilde{g}'(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \left\{ X_{ij,l'} - \hat{E}(X_{ij,l'} | \boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \right\} \right]_{j=1}^{m} \boldsymbol{\Lambda}_r \hat{D}_i(\boldsymbol{\beta})^{\mathrm{T}} \right]_{1 \le r \le k}^{\mathrm{T}} \Xi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \Omega_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[\left[\beta_l \tilde{g}'(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \left\{ X_{ij,1} - \hat{E}(X_{ij,1} | \boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \right\} + \right.\right.$$

$$\left.\left. \sum_{l'=2}^{d_1} \tilde{g}'(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \left\{ X_{ij,l'} - \hat{E}(X_{ij,l'} | \boldsymbol{\beta}^{\mathrm{T}} X_{ij}) \right\} \right]_{j=1}^{m} \boldsymbol{\Lambda}_r \hat{D}_i(\boldsymbol{\beta})^{\mathrm{T}} \right]_{1 \le r \le k}^{\mathrm{T}} \Xi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \times$$

$$\frac{1}{n} \sum_{i=1}^{n} \left[\hat{D}_i(\boldsymbol{\beta}) \boldsymbol{\Lambda}_r \hat{D}_i(\boldsymbol{\beta})^{\mathrm{T}} \begin{pmatrix} \boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(1)0} \\ \boldsymbol{\alpha} - \boldsymbol{\alpha}^0 \end{pmatrix} \left(1 + O_p\left(n^{-1/2}\right)\right) - \hat{D}_i(\boldsymbol{\beta}) \boldsymbol{\Lambda}_r e_i \right]_{r=1}^{k}.$$

Since $\left\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\right\| = O_p\left(n^{-1/2}\right)$ and $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^0\| = O_p\left(n^{-1/2}\right)$, we have that $n^{-1}\partial\left\{\frac{1}{2}Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})\right\}/\partial \beta_l$ is of order $O_p\left(n^{-1/2}\right)$. Thus,

$$\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_l = n \lambda_{1l} \left\{ \lambda_{1l}^{-1} p'_{\lambda_{1l}}(|\beta_l|) \operatorname{sgn}(\beta_l) + O_p\left(n^{-1/2} \lambda_{1l}^{-1}\right) \right\}.$$

Using the results of $\liminf_{n \to \infty} \liminf_{\beta_l \to 0^+} \lambda_{1l}^{-1} p'_{\lambda_{1l}}(|\beta_l|) > 0$ and $n^{-1/2}\lambda_{1l}^{-1} \to 0$, we further obtain $\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_l > 0$ for $\beta_l > 0$ and $\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_l < 0$ for $\beta_l < 0$. Analogously, we can show that $\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \alpha_s > 0$ for $\alpha_s > 0$ and $\partial \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \alpha_s < 0$ for $\alpha_s < 0$. Consequently, the minimum of $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is attained at $\boldsymbol{\beta}_2 = 0, \boldsymbol{\alpha}_2 = 0$, which proves (A.10). This, together with the result of Step I, implies that, with probability tending to 1, $\hat{\boldsymbol{\beta}}_2^{\text{PQIF}} = 0$ and $\hat{\boldsymbol{\alpha}}_2^{\text{PQIF}} = 0$, as $n \to \infty$. This completes the proof of part (i) in Theorem 4.

Step III. Lastly, we demonstrate the asymptotic normality of $\hat{\boldsymbol{\beta}}_1^{\text{PQIF}}$ and $\hat{\boldsymbol{\alpha}}_1^{\text{PQIF}}$. Define

$$\begin{aligned} R_{\lambda_1} &= \left\{ p'_{\lambda_{12}}\left(\left|\beta_2^0\right|\right) \operatorname{sgn}\left(\beta_2^0\right), \ldots, p'_{\lambda d_{10}}\left(\left|\beta_{d_{10}}^0\right|\right) \operatorname{sgn}\left(\beta_{d_{10}}^0\right) \right\}^{\mathrm{T}}, \\ \Sigma_{\lambda_1} &= \operatorname{diag}\left\{ p''_{\lambda_{12}}\left(\left|\beta_2^0\right|\right), \ldots, p''_{\lambda d_{10}}\left(\left|\beta_{d_{10}}^0\right|\right) \right\}, \\ R_{\lambda_2} &= \left\{ p'_{\lambda_{21}}\left(\left|\alpha_1^0\right|\right) \operatorname{sgn}\left(\alpha_1^0\right), \ldots, p'_{\lambda d_{20}}\left(\left|\alpha_{d_{20}}^0\right|\right) \operatorname{sgn}\left(\alpha_{d_{20}}^0\right) \right\}^{\mathrm{T}}, \\ \Sigma_{\lambda_2} &= \operatorname{diag}\left\{ p''_{\lambda_{21}}\left(\left|\alpha_1^0\right|\right), \ldots, p''_{\lambda d_{20}}\left(\left|\alpha_{d_{20}}^0\right|\right) \right\}, \end{aligned} \tag{A.11}$$

where β_l^0 and α_s^0 are the l-th and s-th elements of $\boldsymbol{\beta}_1^0$ and $\boldsymbol{\alpha}_1^0$, respectively, for $2 \le l \le d_1$ and $1 \le s \le d_2$. Furthermore, define

$$\hat{D}_{1i}(\boldsymbol{\beta}_1) = \left[\left\{ \tilde{g}'\left(\boldsymbol{\beta}_1^{\mathrm{T}} X_{i1}^1\right) \hat{X}_{i1}^1, \ldots, \tilde{g}'\left(\boldsymbol{\beta}_1^{\mathrm{T}} X_{im}^1\right) \hat{X}_{im}^1 \right\}^{\mathrm{T}} \boldsymbol{J}_1, \left(\hat{Z}_{i1}^1, \ldots, \hat{Z}_{im}^1\right)^{\mathrm{T}} \right]^{\mathrm{T}},$$

where $\hat{X}_{ij}^1 = X_{ij}^1 - \hat{E}\left(X_{ij}^1 \left|\boldsymbol{\beta}_1^{\mathrm{T}} X_{ij}^1\right.\right)$ and $\hat{Z}_{ij}^1 = Z_{ij}^1 - \hat{E}\left(Z_{ij}^1 \left|\boldsymbol{\beta}_1^{\mathrm{T}} X_{ij}^1\right.\right)$. By (3.1), $\hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}}$ and $\hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}}$ satisfy

$$\mathbf{0} = \begin{pmatrix} \frac{\partial \mathcal{L}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}, \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}}\right)}{\partial \hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}}} \\ \frac{\partial \mathcal{L}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}, \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}}\right)}{\partial \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}}} \end{pmatrix} = \frac{1}{2}\dot{Q}_{1n}^*\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}, \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}}\right) + \tag{A.12}$$

$$n\begin{pmatrix} R_{\lambda_1} \\ R_{\lambda_2} \end{pmatrix} + n\begin{pmatrix} \Sigma_{\lambda_1} & 0 \\ 0 & \Sigma_{\lambda_2} \end{pmatrix}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}} - \boldsymbol{\beta}_1^{(1)0} \\ \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}} - \boldsymbol{\alpha}_1^0 \end{pmatrix} + O_p\left(n^{-1}\right),$$

where the first derivative $\dot{Q}_{1n}^*(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1) = \Omega_{1n}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)^{\mathrm{T}}\, \Xi_{1n}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)^{-1}\, \Omega_{1n}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$. The quantities $\Omega_{1n}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$ and $\Xi_{1n}(\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1)$ are defined in the same manner as $\Omega_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\Xi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in (2.9) by replacing $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$ and $\hat{D}_i$ with $\boldsymbol{X}_{i,1} = (X_{i1}^1, \ldots, X_{im}^1)_{m\times d_{10}}^{\mathrm{T}}$, $\boldsymbol{Z}_{i,1} = (Z_{i1}^1, \ldots, Z_{im}^1)_{m\times d_{20}}^{\mathrm{T}}$, and $\hat{D}_{1i}(\boldsymbol{\beta}_1)$, respectively. Let $\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}} = \{(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}})^{\mathrm{T}}, (\hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}})^{\mathrm{T}}\}^{\mathrm{T}}$. Then,

$$\begin{aligned} \dot{Q}_{1n}^*\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right) &= 2\dot{\Omega}_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{\mathrm{T}} \Xi_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{-1} \Omega_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right) + O_p\left(n^{-1}\right) \\ &= 2\dot{\Omega}_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{\mathrm{T}} \Xi_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{-1} \dot{\Omega}_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)\begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}} - \boldsymbol{\beta}_1^{(1)0} \\ \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}} - \boldsymbol{\alpha}_1^0 \end{pmatrix} - \\ &\quad 2\dot{\Omega}_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{\mathrm{T}} \Xi_{1n}\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{-1} \frac{1}{n}\sum_{i=1}^n \left\{\begin{array}{c} D_{1i}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}\right)\Lambda_1 e_i \\ \vdots \\ D_{1i}\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}\right)\Lambda_k e_i \end{array}\right\} + O_p\left(n^{-1}\right). \end{aligned} \tag{A.13}$$

By (A.12) and (A.13), we have that

$$\left\{\dot{\Omega}_n\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{\mathrm{T}} \Xi_n\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{-1} \dot{\Omega}_n\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right) + n\begin{pmatrix} \Sigma_{\lambda_1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\lambda_2} \end{pmatrix}\right\}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}} - \boldsymbol{\beta}_1^{(1)0} \\ \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}} - \boldsymbol{\alpha}_1^0 \end{pmatrix}$$

$$+n\begin{pmatrix} R_{\lambda_1} \\ R_{\lambda_2} \end{pmatrix} = \dot{\Omega}_n\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{\mathrm{T}} \Xi_n\left(\hat{\boldsymbol{\theta}}_1^{\mathrm{PQIF}}\right)^{-1} \frac{1}{n}\sum_{i=1}^n \left\{\begin{array}{c} D_i\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}\right)\Lambda_1 e_i \\ \vdots \\ D_i\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{PQIF}}\right)\Lambda_k e_i \end{array}\right\} + O_p\left(n^{-1}\right).$$

Let $\Sigma_\lambda = \begin{pmatrix} \Sigma_{\lambda_1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\lambda_2} \end{pmatrix}$. Applying the Lindeberg-Feller Central Limit Theorem and Condition (C5), we obtain that

$$\sqrt{n}\left(\Sigma_{\boldsymbol{\theta}_1^{(1)0}} + \Sigma_\lambda\right)\left\{\begin{pmatrix} \hat{\boldsymbol{\beta}}_1^{(1)\mathrm{PQIF}} - \boldsymbol{\beta}_1^{(1)0} \\ \hat{\boldsymbol{\alpha}}_1^{\mathrm{PQIF}} - \boldsymbol{\alpha}_1^0 \end{pmatrix} + \left(\Sigma_{\boldsymbol{\theta}_1^{(1)0}} + \Sigma_\lambda\right)^{-1}\begin{pmatrix} R_{\lambda_1} \\ R_{\lambda_2} \end{pmatrix}\right\} \to N\left\{\mathbf{0}, \Sigma_{\boldsymbol{\theta}_1^{(1)0}}\right\}.$$

Finally, using the fact that $\sqrt{n}\Sigma_\lambda = \sqrt{n}R_{\lambda_1} = \sqrt{n}R_{\lambda_2} = 0$, under the assumptions that $\lambda_{1l} \to 0$ and $\lambda_{2s} \to 0$ for all $1 \le l \le d_1$ and $1 \le s \le d_2$, we complete the proof of part (ii) in Theorem 4.

Table 1: Variable selection and estimation results for $\boldsymbol{\beta}^0$ with the exchangeable (EXCH), AR(1), and independent (IND) working correlation structures for $\sigma = 0.2, 0.5$. The columns of C, O, and U present the percentage of correct-fitting, over-fitting, and under-fitting, respectively. The column NC (and correspondingly NI) reports the average number of truly nonzero (zero) coefficients that are correctly (incorrectly) set to nonzero. The columns PQIF and ORACLE show the squared root values of the estimated MSEs of the penalized and oracle estimates.

			Variable selection and parameter estimation						
σ	n		C	O	U	NC	NI	PQIF	ORACLE
0.2	100	EX	0.902	0.098	0.000	3	0.104	0.0187	0.0158
		AR(1)	0.854	0.146	0.000	3	0.164	0.0217	0.0179
		IND	0.818	0.182	0.000	3	0.190	0.0265	0.0221
	200	EX	0.986	0.014	0.000	3	0.014	0.0118	0.0110
		AR(1)	0.976	0.024	0.002	3	0.024	0.0130	0.0122
		IND	0.980	0.020	0.000	3	0.020	0.0167	0.0158
	500	EX	1.000	0.000	0.000	3	0.000	0.0071	0.0071
		AR(1)	1.000	0.000	0.000	3	0.000	0.0077	0.0077
		IND	1.000	0.000	0.000	3	0.000	0.0100	0.0095
0.5	100	EX	0.866	0.134	0.000	3	0.144	0.0545	0.0424
		AR(1)	0.846	0.154	0.000	3	0.174	0.0643	0.0559
		IND	0.806	0.194	0.000	3	0.210	0.0733	0.0605
	200	EX	0.926	0.074	0.000	3	0.076	0.0310	0.0272
		AR(1)	0.910	0.090	0.002	3	0.096	0.0341	0.0308
		IND	0.896	0.104	0.000	3	0.108	0.0392	0.0366
	500	EX	0.990	0.010	0.000	3	0.010	0.0195	0.0173
		AR(1)	0.986	0.014	0.000	3	0.014	0.0209	0.0192
		IND	0.982	0.018	0.000	3	0.018	0.0232	0.0221

Table 2: Variable selection and estimation results for $\boldsymbol{\alpha}^0$ with the exchangeable (EX), AR(1), and independent (IND) working correlation structures for $\sigma = 0.2, 0.5$. The columns of C, O, and U present the percentage of correct-fitting, over-fitting, and under-fitting, respectively. The column NC (and correspondingly NI) reports the average number of truly nonzero (zero) coefficients that are correctly (incorrectly) set to nonzero. The columns PQIF and ORACLE show the squared root values of the estimated MSEs of the penalized and oracle estimates.

			Variable selection and parameter estimate						
σ	n		C	O	U	NC	NI	PQIF	ORACLE
		EX	0.906	0.094	0.000	2	0.102	0.0152	0.0141
	100	AR(1)	0.854	0.146	0.000	2	0.084	0.0176	0.0164
		IND	0.918	0.082	0.000	2	0.084	0.0210	0.0200
		EX	0.994	0.006	0.000	2	0.006	0.0105	0.0100
0.2	200	AR(1)	0.980	0.020	0.002	2	0.020	0.0114	0.0110
		IND	0.972	0.028	0.000	2	0.030	0.0141	0.0141
		EX	1.000	0.000	0.000	2	0.000	0.0063	0.0063
	500	AR(1)	1.000	0.000	0.000	2	0.000	0.0071	0.0071
		IND	0.996	0.004	0.000	2	0.004	0.0089	0.0089
		EX	0.902	0.098	0.000	2	0.108	0.0387	0.0332
	100	AR(1)	0.822	0.178	0.000	2	0.182	0.0412	0.0374
		IND	0.876	0.124	0.000	2	0.140	0.0534	0.0471
		EX	0.968	0.032	0.000	2	0.032	0.0259	0.0255
0.5	200	AR(1)	0.944	0.056	0.002	2	0.078	0.0294	0.0276
		IND	0.920	0.080	0.000	2	0.080	0.0433	0.0342
		EX	0.992	0.008	0.000	2	0.008	0.0173	0.0173
	500	AR(1)	0.994	0.006	0.000	2	0.006	0.0182	0.0179
		IND	0.994	0.006	0.000	2	0.006	0.0182	0.0179

Table 3: The PQIF estimates for debt maturity example with standard errors (SE) and the corresponding p-values across the EX, AR(1), and IND working correlation structures.

Variables	EXCH			AR(1)			IND		
	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
LEVERAGE (X_1)	0.867	0.006	< 0.001	0.846	0.003	< 0.001	0.916	0.003	< 0.001
ASSETMAT (X_2)	0.478	0.012	< 0.001	0.516	0.006	< 0.001	0.379	0.014	< 0.001
MV/BV (X_3)	0.073	0.012	< 0.001	0.066	0.012	< 0.001	0.039	0.014	0.003
SIZE (X_4)				0.052	0.013	< 0.001	0.058	0.013	< 0.001
GTAXRATE (X_6)	0.113	0.011	< 0.001	0.027	0.011	0.007			
VAR (X_8)	-0.033	0.017	0.026	-0.102	0.016	< 0.001	-0.107	0.012	< 0.001
LOWBOND (Z_1)	-0.328	0.030	< 0.001	-0.348	0.028	< 0.001	-0.385	0.031	< 0.001
HIGHBOND (Z_2)	-0.106	0.047	0.012	-0.101	0.044	0.011	-0.125	0.047	< 0.001

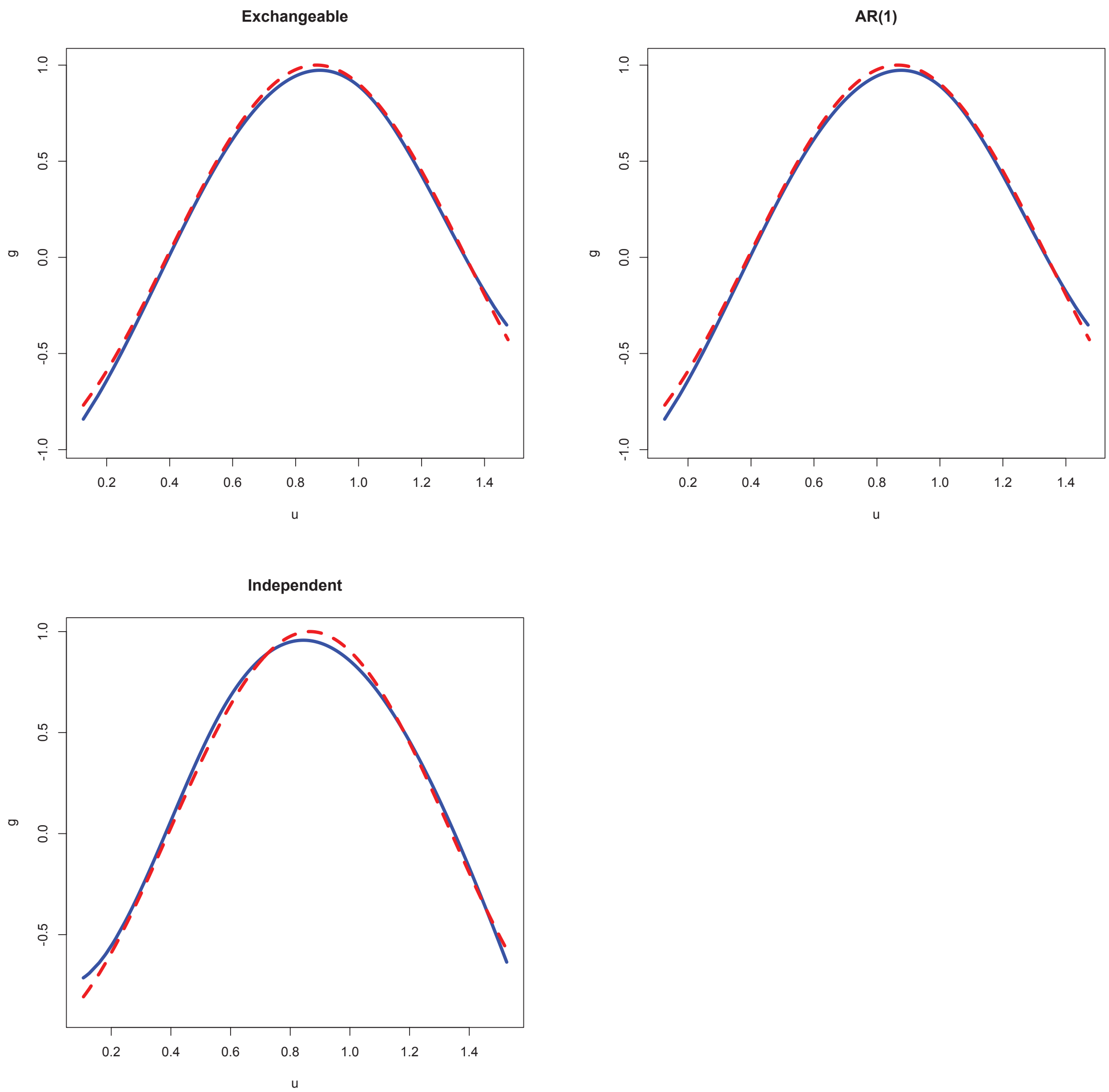

Figure 1: Three graphs of $\hat{g}\,(\hat{u})$ versus $\hat{u}$ (solid curves), together with the true function $g(u)$ (dashed curve), for the exchangeable, AR(1), and independent working correlation structures, when the sample size $n = 100$.

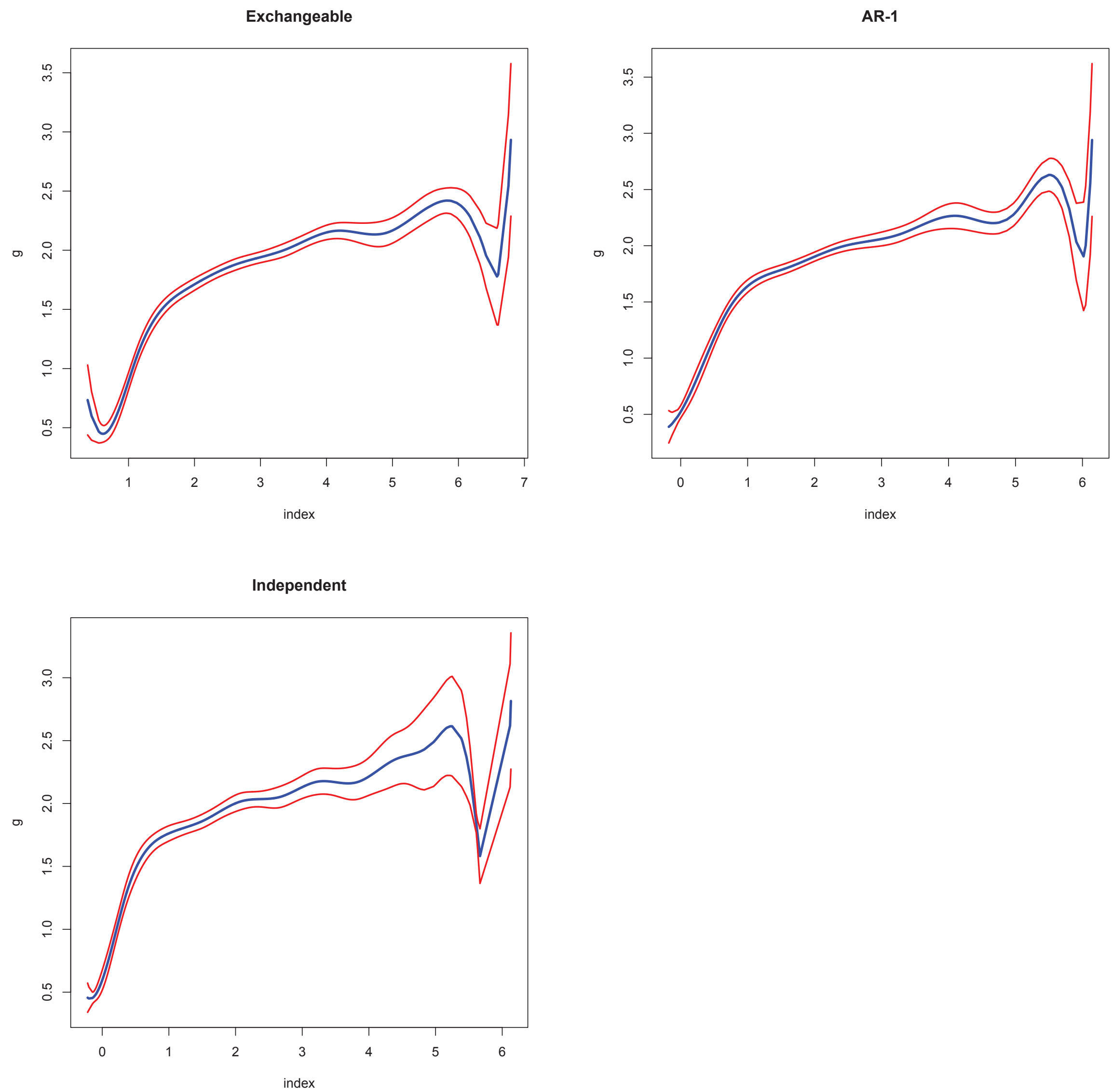

Figure 2: Three graphs of $\hat{g}(\hat{u})$ versus $\hat{u}$ (solid curves), together with their corresponding 95% confidence intervals (upper and lower thin curves), for the exchangeable, AR(1), and independent working correlation structures.