Ann Inst Stat Math (2016) 68:725–764
DOI 10.1007/s10463-015-0513-1

Parameter estimation for a generalized semiparametric model with repeated measurements

Shujie Ma · Zijian Huang · Chih-Ling Tsai

Received: 21 March 2014 / Revised: 6 November 2014 / Published online: 17 March 2015

Abstract In this paper, we propose a flexible generalized semiparametric model for repeated measurements by combining generalized partially linear single-index models with varying coefficient models. The proposed model is a useful analytic tool to explore dynamic patterns which naturally exist in longitudinal data and also study possible nonlinear relationships between the response and covariates. We then employ the quadratic inference function and develop an estimation procedure to estimate unknown regression parameters and nonparametric functions. To select variables and estimate parameters simultaneously, we further obtain penalized estimators. Moreover, we establish theoretical properties of the parametric and nonparametric estimators. Both simulations and an empirical example are presented to illustrate the use of the proposed model.

Keywords Quadratic inference function · Partially linear single-index model · Penalized estimator · Repeated measurements · Varying coefficient model · B-splines

S. Ma's research was partially supported by NSF Grant DMS 1306972.

S. Ma (✉) · Z. Huang
Department of Statistics, University of California-Riverside, Riverside, CA 92521, USA
e-mail: shujie.ma@ucr.edu

Z. Huang
e-mail: zhuan006@ucr.edu

C.-L. Tsai
Graduate School of Management, University of California-Davis, Davis, CA 95616, USA
e-mail: cltsai@ucdavis.edu

1 Introduction

In the last three decades, advanced computing and telecommunication technologies have enabled researchers to collect data effectively and accurately. Hence, it is not surprising that the collected data can be complex and the analysis of such data is challenging. For example, in the regression context, the response variable can be discrete with repeated measurements, the relationship between the mean of the response variable and covariates can be nonlinear, and the coefficients of explanatory variables can be dynamic. This motivates us to propose a model that can simultaneously account for these characteristics.

To take into account discrete responses and nonlinearity, Carroll et al. (1997) proposed generalized partially linear single-index models (GPLSIM). These models encompass several important models, e.g., single-index models (Brillinger 1983; Horowitz 1998; Cui et al. 2011), generalized linear models (McCullagh and Nelder 1989), partially linear models (Speckman 1998; Härdle et al. 2000), generalized partially linear models (Boente et al. 2006), and partially linear single-index models (Yu and Ruppert 2002; Xia and Härdle 2006; Ma and Zhu 2013). The above references mainly focus on parameter estimation. Recently, researchers have employed penalized procedures (e.g., LASSO, Tibshirani 1996; SCAD, Fan and Li 2001) to simultaneously select variables and estimate parameters for those models (e.g., Xie and Huang 2009; Liang et al. 2010; Zhang et al. 2010; Zeng et al. 2012).

Although the GPLSIM has played an important role in data analysis, it does not allow regression coefficients to be dynamic. To this end, Cleveland et al. (1991) and Hastie and Tibshirani (1993) proposed varying coefficient models, which have been applied in diverse fields, such as biological science, economics, finance, medicine, and social science. Further extensions to broad models are developed; see, for example, generalized varying coefficient models (Cai et al. 2000a), semi-varying coefficient models (Zhang et al. 2002), survival models (Fan et al. 2006) and the newly proposed varying index coefficient model (Ma and Song 2014). It is also noteworthy that an analog to the varying coefficient structure has been studied in the field of time series (e.g., see Chen and Tsay 1993; Cai et al. 2000b). An excellent review paper on varying coefficient models can be found in Fan and Zhang (2008).

To better understand the performance of a response variable for each individual subject, a number of GPLSIMs as well as varying coefficient models have been extended to take into account repeated measurements (or longitudinal data or clustered data). Accordingly, various parameter estimation and model selection procedures are proposed (e.g., see Lin and Ying 2001; Davis 2002; Diggle et al. 2002; Huang et al. 2002; Wang 2003; Fan and Li 2004; Fan and Huang 2005; Lin and Carroll 2006; Wang et al. 2008; Ma 2012; Xu and Zhu 2012). To obtain parameter estimation in repeated measurements, one needs to incorporate the correlation structure. Among available approaches, Qu and Li (2006) employed quadratic inference function (QIF) in Qu et al. (2000) to directly incorporate correlations into their varying coefficient models without estimating nuisance parameters associated with correlations. Recently, Zhou and Qu (2012) adopted the QIF approach to obtain estimation and selection of correlation structure.

In this paper, we introduce a generalized semiparametric model for repeated measurements by combining the GPLSIM with varying coefficient models. The proposed model is a useful analytic tool to investigate dynamic patterns of slope functions with some covariates such as time which naturally exist in longitudinal data as well as to capture possible nonlinear relationships between the response and covariates. Moreover, it contains many existing known parametric and nonparametric models as special cases, and thus it can be used for different types of data. Since each of GPLSIM and varying coefficient models has its own special feature, it is not surprising that obtaining parameter estimators and their theoretical properties becomes more challenging. For the sake of estimation, we first approximate the nonparametric function and coefficient functions by their corresponding linear combinations of spline basis functions. We then propose a profile QIF procedure to obtain parameter estimates. It is worth noting that the profile procedure induces a single objective function of the parameters, which allows us to consider the penalization method for variable estimations and selections. The resulting penalized estimators of the nonzero coefficients are asymptotically normal and have the oracle property.

The rest of this paper is organized as follows. Section 2 introduces the model structure and notation. Section 3 presents the estimation procedure and demonstrates the consistency and asymptotic normality of parametric estimators as well as the consistency of nonparametric estimators. Section 4 proposes penalized estimators and shows their oracle properties. Simulation studies and an empirical example are presented in Sect. 5. We conclude this article with discussions in Sect. 6, and technical proofs are relegated in the Appendix.

2 A generalized semiparametric model

To introduce the generalized semiparametric model by unifying partially linear single-index and varying coefficient models with repeated measurements, we denote $(Y_{ij}, X_{ij}, Z_{ij}, T_{ij})$ as the j-th repeated observation for the i-th subject (or experimental unit) for $1 \le i \le n$ and $1 \le j \le m_i$, where Y_{ij} is the response variable and it is independent of other subjects, $X_{ij} = \left(X_{ij,1}, \ldots, X_{ij,p}\right)^{\mathrm{T}}$ and $Z_{ij} = \left(Z_{ij,1}, \ldots, Z_{ij,d}\right)^{\mathrm{T}}$ are p-dimensional and d-dimensional vectors of covariates, respectively, and T_{ij} represents a single predictor. Let $\mathbb{C}_{ij} = \left(X_{ij}^{\mathrm{T}}, Z_{ij}^{\mathrm{T}}, T_{ij}\right)^{\mathrm{T}}$ be the collection of covariates for the j-th observation of the i-th subject. We then consider the marginal model and assume that $E\left(Y_{ij} \left|\mathbb{C}_{ij}\right.\right) = \mu_{ij}$, where the marginal mean μ_{ij} depends on $\mathbb{C}_{ij}$ through a known monotonic and differentiable link function ϑ. This leads to the predictor function

$$\eta_{ij} = \vartheta(\mu_{ij}) = g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) + \sum\nolimits_{l=1}^{d_1} \alpha_l\left(T_{ij}\right) Z_{ij,l} + \sum\nolimits_{l=d_1+1}^{d} \alpha_l Z_{ij,l},$$
$$j = 1, \ldots, m_i, i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^{\mathrm{T}}$ is a p-dimensional index parameter, $\alpha_l(\cdot)$, $l = 1, \ldots, d_1$, are unknown smooth functions, α_l, $l = d_1 + 1, \ldots, d$, are coefficients, and $g(\cdot)$ is an unknown differentiable function of $U_{ij}(\boldsymbol{\beta}) = X_{ij}^{\mathrm{T}}\boldsymbol{\beta}$. For identifiability, we assume that $\boldsymbol{\beta}$ belongs to the parameter space:

$$\Theta = \left\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1, \beta_1 > 0, \boldsymbol{\beta} \in R^p\right\}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm of any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$ such that $\|\boldsymbol{\zeta}\| = (|\zeta_1|^2 + \cdots + |\zeta_s|^2)^{1/2}$.

Model (1) contains many existing models as special cases. When $\alpha_l\left(T_{ij}\right) = \alpha_l$ for $1 \le l \le d_1$, α_l are unknown constants, and $m_i = 1$ for $i = 1, \ldots, n$, it leads to a generalized partially linear single-index model (Carroll et al. 1997); when $g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) = 0$, it yields a semiparametric varying coefficient partially linear model (Fan and Huang 2005); when $p = 1$ and $\alpha_l\left(T_{ij}\right) = \alpha_l$ for $1 \le l \le d_1$, it results to a generalized partially linear model (Härdle et al. 2000); when $g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) = 0$ and $\alpha_l = 0$ for $l = d_1 + 1, \ldots, d$, it gives a generalized varying coefficient model (Hastie and Tibshirani 1993; Cai et al. 2000a). It is worth noting that model (1) is different from the varying index coefficient model proposed by Ma and Song (2014), since the latter aims to assess nonlinear interaction effects of index variables with other covariates on the response in the cross-sectional data setting.

3 Parameter estimates

3.1 The approximation of predictor function

In this subsection, we approximate the unknown functions $g(\cdot)$ and $\alpha_l(\cdot)$ in (1) by B-splines described as follows. Based on the given knots, we define the sets of q-th order B-spline basis functions $B_1(u) = \left\{B_{1,J}(u) : 1 \le J \le N+q\right\}^{\mathrm{T}}$ and $B_2(t) = \left\{B_{2,J}(t) : 1 \le J \le N+q\right\}^{\mathrm{T}}$ (see de Boor 2001), where N is the number of interior knots with the distance between neighboring knots satisfying the conditions given in Zhou et al. (1998). Then, the unknown function g in (1) can be approximated by a linear combination of the B-spline functions such that $g(u) \approx \sum_{J=1}^{N+q} \gamma_{J,0} B_{1,J}(u) = B_1(u)^{\mathrm{T}}\boldsymbol{\gamma}_0$ with a set of coefficients $\boldsymbol{\gamma}_0 = \left(\gamma_{1,0}, \ldots, \gamma_{N+q,0}\right)^{\mathrm{T}}$. Analogously, $\alpha_l(t)$ in (1) can be approximated by $\alpha_l(t) \approx \sum_{J=1}^{N+q} \gamma_{J,l} B_{2,J}(t) = B_2(t)^{\mathrm{T}}\boldsymbol{\gamma}_l$, where $\boldsymbol{\gamma}_l = \left(\gamma_{1,l}, \ldots, \gamma_{N+q,l}\right)^{\mathrm{T}}$. Accordingly, we obtain an approximation of the predictor function η_{ij}, which is

$$\begin{aligned}\widetilde{\eta}_{ij} = &\sum_{J=1}^{N+q} \gamma_{J,0} B_{1,J}\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right) + \sum_{l=1}^{d_1}\sum_{J=1}^{N+q} \gamma_{J,l} B_{2,J}\left(T_{ij}\right) Z_{ij,l} \\ &+ \sum_{l=d_1+1}^{d} \alpha_l Z_{ij,l}.\end{aligned} \tag{3}$$

From the above equation, we propose a two-step estimation procedure to estimate parametric and nonparametric components in the following two subsections, respectively.

3.2 The profile QIF estimators of parametric vectors

Let $Y_i = \left(Y_{i1}, \ldots, Y_{im_i}\right)^{\mathrm{T}}$ and $\mu_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \left(\mu_{i1}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right), \ldots, \mu_{im_i}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right)^{\mathrm{T}}$. Then, denote $\widetilde{\mu}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \left(\widetilde{\mu}_{i1}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right), \ldots, \widetilde{\mu}_{im_i}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right)^{\mathrm{T}}$, where $\widetilde{\mu}_{ij}\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \vartheta^{-1}\left(\widetilde{\eta}_{ij}\right), \boldsymbol{\alpha} = \left(\alpha_{d_1+1}, \ldots, \alpha_d\right)^{\mathrm{T}}$ is a $d_2 \times 1$ vector, $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_0^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{d_1}^{\mathrm{T}}\right)^{\mathrm{T}}$ is a $(1 + d_1)\, J_n \times 1$ vector, and $J_n = N + q$. For given $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, the quasi-likelihood estimator of $\boldsymbol{\gamma}$ is the solution of the following estimating equations,

$$\sum\nolimits_{i=1}^{n} \widetilde{\mu}_i'\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \mathbf{V}_i^{-1}\left(Y_i - \widetilde{\mu}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right) = 0, \tag{4}$$

where $\widetilde{\mu}_i'\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \left[\left(\widetilde{\mu}_{i1}', \ldots, \widetilde{\mu}_{im_i}'\right)\right]_{(1+d_1)J_n \times m_i}$, $\widetilde{\mu}_{ij}' = \partial \widetilde{\mu}_{ij} / \partial \boldsymbol{\gamma}$ for $j = 1, \ldots, m_i$, and $\mathbf{V}_i$ is the $m_i \times m_i$ covariance matrix of Y_i. Since $\mathbf{V}_i$ is often unknown in practice, we adopt the approach in Liang and Zeger (1986) and assume that $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i\left(\boldsymbol{\varsigma}\right) \mathbf{A}_i^{1/2} / \phi$, where $\mathbf{R}_i\left(\boldsymbol{\varsigma}\right)$ is the working correlation matrix of Y_i, $\boldsymbol{\varsigma}$ is a vector of nuisance parameters, and $\mathbf{A}_i$ is an $m_i \times m_i$ diagonal matrix with the marginal variance of Y_{ij} as its j-th diagonal element. However, the working correlation structure may be misspecified. Hence, we further apply the quadratic inference function (QIF) in Qu et al. (2000) to efficiently incorporate the within-cluster correlation structure. For the sake of simplicity, we assume that cluster sizes are equal, i.e., $m_i = m < \infty$, and let $\mathbf{R}$ be a common working correlation matrix. When the cluster sizes are unequal, our estimation procedure given below can be modified via the same technique proposed by Xue et al. (2010). Following the QIF approach, the inverse of $\mathbf{R}$ can be approximated by a linear combination of κ basis matrices, i.e.,

$$\mathbf{R}^{-1} \approx a_1 \mathbf{M}_1 + \cdots + a_\kappa \mathbf{M}_\kappa, \tag{5}$$

where $\mathbf{M}_1 = \mathbf{I}$ (the identity matrix) and $\mathbf{M}_k$ are known symmetric basis matrices for $1 \leq k \leq \kappa$.

We next construct the QIF to obtain parameter estimators. To this end, consider the estimating function of $\boldsymbol{\gamma}$, for the given $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$:

$$\begin{aligned} \widetilde{\phi}_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) &= n^{-1} \sum\nolimits_{i=1}^{n} \widetilde{\phi}_{in}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \left\{\widetilde{\phi}_{n,1}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)^{\mathrm{T}}, \ldots, \widetilde{\phi}_{n,\kappa}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)^{\mathrm{T}}\right\}^{\mathrm{T}} \\ &= n^{-1} \left\{ \begin{array}{c} \sum_{i=1}^{n} \mathbf{Q}_i\left(\boldsymbol{\beta}\right)^{\mathrm{T}} \widetilde{\Delta}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \mathbf{A}_i^{-1/2} \mathbf{M}_1 \mathbf{A}_i^{-1/2}\left(Y_i - \widetilde{\mu}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right) \\ \vdots \\ \sum_{i=1}^{n} \mathbf{Q}_i\left(\boldsymbol{\beta}\right)^{\mathrm{T}} \widetilde{\Delta}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \mathbf{A}_i^{-1/2} \mathbf{M}_\kappa \mathbf{A}_i^{-1/2}\left(Y_i - \widetilde{\mu}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right) \end{array} \right\}_{\kappa J_n (1+d_1) \times 1}, \end{aligned} \tag{6}$$

where $\widetilde{\Delta}_i\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \operatorname{diag}\left(\widetilde{\nu}_{i1}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right), \ldots, \widetilde{\nu}_{im}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right)\right)$, $\widetilde{\nu}_{ij}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \partial \widetilde{\mu}_{ij}\left(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) / \partial \widetilde{\eta}_{ij}$,

$\mathbf{Q}_i(\boldsymbol{\beta}) = \left(Q_{i1}(\boldsymbol{\beta})^{\mathrm{T}}, \ \ldots, Q_{im_i}(\boldsymbol{\beta})^{\mathrm{T}}\right)^{\mathrm{T}}$, and

$$Q_{ij}(\boldsymbol{\beta}) = \left[B_1\left(U_{ij}(\boldsymbol{\beta})\right)^{\mathrm{T}}, \left\{B_2\left(T_{ij}\right)^{\mathrm{T}} Z_{ij,l} : 1 \le l \le d_1\right\}\right]^{\mathrm{T}}_{J_n(1+d_1)\times 1}.$$

Then, define the QIF to be

$$\widetilde{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \widetilde{\phi}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}} \widetilde{C}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \widetilde{\phi}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{7}$$

where $\widetilde{C}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = n^{-2}\sum_{i=1}^n \widetilde{\phi}_{in}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})\,\widetilde{\phi}_{in}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}$. Accordingly, the QIF estimator of $\boldsymbol{\gamma}$ is

$$\begin{aligned}\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \left[\left\{\widetilde{\boldsymbol{\gamma}}_0(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\gamma}}_{d_1}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}\right\}^{\mathrm{T}}\right] \\ &= \arg \min_{\boldsymbol{\gamma}\in R^{(1+d_1)J_n}} \widetilde{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}).\end{aligned}$$

As a result, the QIF estimators of $g(\cdot)$, $g'(\cdot)$ (the first derivative of g), and $\alpha_l(\cdot)$ are $\widetilde{g}(u; \boldsymbol{\beta}, \boldsymbol{\alpha}) = B_1(u)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_0(\boldsymbol{\beta}, \boldsymbol{\alpha})$, $\widetilde{\alpha}_l(t; \boldsymbol{\beta}, \boldsymbol{\alpha}) = B_2(t)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_l(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and $\widetilde{g}'(u; \boldsymbol{\beta}, \boldsymbol{\alpha}) = B_1'(u)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_0(\boldsymbol{\beta}, \boldsymbol{\alpha})$, respectively, where $B_1'(u)$ is the first derivative of $B_1(u)$. By replacing $g(\cdot)$ and $\alpha_l(\cdot)$ with $\widetilde{g}(\cdot; \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\widetilde{\alpha}_l(\cdot; \boldsymbol{\beta}, \boldsymbol{\alpha})$ in (3), we obtain

$$\widehat{\eta}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \widetilde{g}\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}; \boldsymbol{\beta}, \boldsymbol{\alpha}\right) + \sum_{l=1}^{d_1} \widetilde{\alpha}_l\left(T_{ij}; \boldsymbol{\beta}, \boldsymbol{\alpha}\right) Z_{ij,l} + \sum_{l=d_1+1}^{d} \alpha_l Z_{ij,l}. \tag{8}$$

Before estimating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we use the assumptions of $||\boldsymbol{\beta}|| = 1$ and $\beta_1 > 0$ in (2) to reform the space of $\boldsymbol{\beta}$ given below, which ensures identifiability.

$$\left\{\left(\left(1 - \sum_{s=2}^{p}\beta_s^2\right)^{1/2}, \beta_2, \ldots, \beta_p\right)^{\mathrm{T}} : \sum_{s=2}^{p}\beta_s^2 < 1\right\}.$$

Denote $\widehat{\eta}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \{\widehat{\eta}_{i1}(\boldsymbol{\beta}, \boldsymbol{\alpha}), \ldots, \widehat{\eta}_{im}(\boldsymbol{\beta}, \boldsymbol{\alpha})\}^{\mathrm{T}}$ and its gradient with respect to $\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}$ by $\widehat{\mathbf{D}}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \nabla\widehat{\eta}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \left\{\frac{\partial\widehat{\eta}_{i1}(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}}, \ldots, \frac{\partial\widehat{\eta}_{im}(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}}\right\}^{\mathrm{T}}_{m\times(p-1+d_2)}$, where $\boldsymbol{\beta}_{-1} = \left(\beta_2, \ldots, \beta_p\right)^{\mathrm{T}}$. Consider the profiled estimating function of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$,

$$\begin{aligned}\psi_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^n \psi_{in}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \left\{\psi_{n,1}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}, \ldots, \psi_{n,\kappa}(\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}}\right\}^{\mathrm{T}} \\ &= n^{-1}\left\{\begin{array}{c}\sum_{i=1}^n \widehat{\mathbf{D}}_i^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha})\,\Delta_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\,\mathbf{A}_i^{-1/2}\mathbf{M}_1\mathbf{A}_i^{-1/2}\left(Y_i - \widehat{\mu}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\right) \\ \vdots \\ \sum_{i=1}^n \widehat{\mathbf{D}}_i^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\alpha})\,\Delta_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\,\mathbf{A}_i^{-1/2}\mathbf{M}_\kappa\mathbf{A}_i^{-1/2}\left(Y_i - \widehat{\mu}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})\right)\end{array}\right\}_{\kappa(p-1+d_2)\times 1},\end{aligned}$$

where $\widehat{\mu}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \{\widehat{\mu}_{i1}(\boldsymbol{\beta}, \boldsymbol{\alpha}), \ldots, \widehat{\mu}_{im}(\boldsymbol{\beta}, \boldsymbol{\alpha})\}^{\mathrm{T}}$, $\widehat{\mu}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \vartheta^{-1}\left\{\widehat{\eta}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})\right\}$, $\Delta_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathrm{Diag}(\widehat{\nu}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}))$, and $\widehat{\nu}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \partial\widehat{\mu}_{ij}(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial\widehat{\eta}_{ij}$. Then, the profiled

QIF estimator of $\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}$ is

$$\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} = \arg \min_{(\boldsymbol{\beta}_{-1}, \boldsymbol{\alpha})} Q_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha}),$$

where

$$Q_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha}) = \psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}} \Psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1} \psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{9}$$

and $\Psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha}) = n^{-2} \sum_{i=1}^n \psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha})^{\mathrm{T}} \psi_n^* (\boldsymbol{\beta}, \boldsymbol{\alpha})$. Using the fact that $\beta_1 = \sqrt{1 - \left\|\boldsymbol{\beta}_{-1}\right\|^2}$, we also obtain the estimator $\widehat{\beta}_1$. The detailed procedure for computing $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ is given in Sect. 4.2.

To study the asymptotic properties of the parametric estimators, we need to introduce a few quantities evaluated at the true parameter values. To this end, let $\boldsymbol{\beta}^0 = \left(\beta_1^0, \boldsymbol{\beta}_{-1}^{0\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0$ be the true parameter vectors, $\boldsymbol{\beta}_{-1}^0 = \left(\beta_2^0, \ldots, \beta_p^0\right)^{\mathrm{T}}$, and $\mathbf{J}^0 = \frac{\partial \boldsymbol{\beta}^0}{\partial \boldsymbol{\beta}^{(1)}}$ given below be the Jacobian matrix of size $p \times (p-1)$.

$$\mathbf{J}^0 = \begin{pmatrix} -\boldsymbol{\beta}_{-1}^{0\mathrm{T}} / \sqrt{1 - \left\|\boldsymbol{\beta}_{-1}^0\right\|^2} \\ \mathbf{I}_{p-1} \end{pmatrix}_{p \times (p-1)}.$$

For $1 \le s \le p$ and $0 \le l \le d_1$, let $\boldsymbol{\xi}_{s,l} = \left(\xi_{s,1,l}, \ldots, \xi_{s,N+q,l}\right)^{\mathrm{T}}$ be a $J_n \times 1$ vector of parameters. Let $\boldsymbol{\xi}_s = \left\{\left(\boldsymbol{\xi}_{s,0}^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_{s,d_1}^{\mathrm{T}}\right)^{\mathrm{T}}\right\}_{(1+d_1) J_n \times 1}$. For $1 \le s \le p$, we further define

$$\omega_{n,s} \left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}_s\right) = n^{-1} \sum_{i=1}^n \begin{bmatrix} \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Lambda_1 \Delta_i \left\{X_{ij,s} - \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \boldsymbol{\xi}_s\right\} \\ \vdots \\ \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Lambda_\kappa \Delta_i \left\{X_{ij,s} - \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \boldsymbol{\xi}_s\right\} \end{bmatrix}_{\kappa J_n (1+d_1) \times 1} \quad \text{and}$$

$$\Xi_n \left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \frac{1}{n^2} \times \sum_{i=1}^n \begin{bmatrix} \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Gamma_{1,1} \Delta_i \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right) & \cdots & \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Gamma_{1,\kappa} \Delta_i \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right) \\ \vdots & \ddots & \vdots \\ \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Gamma_{\kappa,1} \Delta_i \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right) & \cdots & \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Gamma_{\kappa,\kappa} \Delta_i \mathbf{Q}_i \left(\boldsymbol{\beta}^0\right) \end{bmatrix}_{\kappa J_n (1+d_1) \times \kappa J_n (1+d_1)}, \tag{10}$$

where $\Lambda_k = \mathbf{A}_i^{-1/2}\mathbf{M}_k\mathbf{A}_i^{-1/2}$, $\Gamma_{k,k'} = \Lambda_k \mathbf{V}_i \Lambda_{k'}$ for $1 \le k, k' \le \kappa$, and Δ_i, $\mathbf{V}_i$, and Λ_k are evaluated at $(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0)$. Then, we obtain the estimate of $\boldsymbol{\xi}_s$,

$$\widehat{\boldsymbol{\xi}}_s = \arg\min_{\boldsymbol{\xi}_s \in R^{(1+d_1)J_n}} \left\{\omega_{n,s}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}_s\right)^{\mathrm{T}} \Xi_n \left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{-1} \omega_{n,s}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}_s\right)\right\}. \tag{11}$$

In addition, replace $X_{ij,s}$ and $\boldsymbol{\xi}_s$ in $\omega_{n,s}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}_s\right)$ by $Z_{ij,l}$ and $\boldsymbol{\tau}_l$, respectively, which yields $\omega_{n,l}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\tau}_l\right)$ for $d_1 + 1 \le l \le d$. Adapting (11), we obtain the estimate $\widehat{\boldsymbol{\tau}}_l$.

Define $\widehat{X}_{ij,s} = X_{ij,s} - Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\boldsymbol{\xi}}_s$, $\widehat{X}_{ij} = \left(\widehat{X}_{ij,1}, \ldots, \widehat{X}_{ij,p}\right)^{\mathrm{T}}$, $\widehat{Z}_{ij,l} = Z_{ij,l} - Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\boldsymbol{\tau}}_l$, and $\widehat{Z}_{ij}^{(2)} = \left(\widehat{Z}_{ij,d_1+1}, \ldots, \widehat{Z}_{ij,d}\right)^{\mathrm{T}}$. In Lemma 3 of Appendix, we demonstrate that

$$\frac{\partial\widehat{\eta}_{ij}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\ \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}} = \left\{\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0; \boldsymbol{\beta}^0,\ \boldsymbol{\alpha}^0\right)\widehat{X}_{ij}^{\mathrm{T}}\mathbf{J}^0, \widehat{Z}_{ij}^{(2)\mathrm{T}}\right\}^{\mathrm{T}}\left\{1 + o_p(1)\right\}.$$

Accordingly,

$$\begin{aligned}&\widehat{\mathbf{D}}_i\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\\ &= \left[\left\{\widetilde{g}'\left(X_{i1}^{\mathrm{T}}\boldsymbol{\beta}^0; \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\widehat{X}_{i1}, \ldots, \widetilde{g}'\left(X_{im}^{\mathrm{T}}\boldsymbol{\beta}^0; \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\widehat{X}_{im}\right\}^{\mathrm{T}}\right.\\ &\quad\left.\times\, \mathbf{J}^0, \left(\widehat{Z}_{i1}^{(2)}, \ldots, \widehat{Z}_{im}^{(2)}\right)^{\mathrm{T}}\right]\left\{1 + o_p(1)\right\}.\end{aligned}$$

Define $D_{ij}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \left\{g'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0\right)\widehat{X}_{ij}^{\mathrm{T}}\mathbf{J}^{\mathbf{0}}, \widehat{Z}_{ij}^{(2)\mathrm{T}}\right\}^{\mathrm{T}}$,

$$\dot{\psi}_n\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right) = E\begin{Bmatrix}\mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Lambda_1\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i\\ \vdots\\ \mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Lambda_\kappa\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i\end{Bmatrix} \quad \text{and}$$

$$\begin{aligned}&\Psi_n\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\\ &= E\begin{Bmatrix}\mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Gamma_{1,1}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i & \cdots & \mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Gamma_{1,\kappa}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i\\ \vdots & \ddots & \vdots\\ \mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Gamma_{\kappa,1}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i & \cdots & \mathbf{D}_i^{\mathrm{T}}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\Gamma_{\kappa,\kappa}\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}\right)\mathbf{D}_i\end{Bmatrix},\end{aligned}$$

where $\mathbf{D}_i = \mathbf{D}_i\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \left(D_{i1}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right), \ldots, D_{im}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right)^{\mathrm{T}}$. Then, the asymptotic properties of parametric estimators are given below.

Theorem 1 *Assume that conditions* (C1)–(C5) *in the Appendix hold,* $N^4 n^{-1} = o(1)$, *and* $N^{-4r+2}n = o(1)$ *with* $r > 3/2$ *in condition* (C2). *Then, we have*

$$\left\| \left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\beta}_{-1}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}} \right\| = o_p\left(1\right), \textit{ and, as } n \to \infty,$$

$$\sqrt{n}\left(\Sigma_n^{(1)}\right)^{1/2}\left(\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\beta}_{-1}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}}\right) \to N\left(0, \mathbf{I}_{p-1+d_2}\right),$$

where $\Sigma_n^{(1)} = \dot{\psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{\mathrm{T}} \Psi_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) \dot{\psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$ *and* $\mathbf{I}_a$ *denotes the identity matrix with dimension* $a \times a$.

Let $\boldsymbol{\Upsilon} = \begin{pmatrix} \mathbf{J}_{p\times(p-1)} & \mathbf{0}_{p\times d_2} \\ \mathbf{0}_{d_2\times(p-1)} & \mathbf{I}_{d_2\times d_2} \end{pmatrix}$. The above theorem, together with the multivariate delta method, establishes the asymptotic normality of parametric estimators, $\sqrt{n}\Sigma_n^{1/2}\left(\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\beta}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}}\right) \to N\left(0, \mathbf{I}_{p+d_2}\right)$, as $n \to \infty$, where $\Sigma_n = \Upsilon \Sigma_n^{(1)} \boldsymbol{\Upsilon}^{\mathrm{T}}$. It is also worth noting that the resulting estimators are not semiparametric efficient since we assume that the true correlation structure is unknown and the working correlation may be misspecified.

3.3 The QIF estimator of nonparametric functions

After obtaining the parametric estimators $\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$, we replace $\left(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}$ by $\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$ in $\widetilde{Q}_n\left(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ of (7). This allows us to find the estimator

$$\widehat{\boldsymbol{\gamma}} = \left(\widehat{\boldsymbol{\gamma}}_0^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\gamma}}_{d_1}^{\mathrm{T}}\right)^{\mathrm{T}} = \arg \min_{\boldsymbol{\gamma}\in\mathbf{R}^{(1+d_1)J_n}} \widetilde{Q}_n\left(\boldsymbol{\gamma}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}\right).$$

Accordingly, the estimators of nonparametric functions $g\left(\cdot\right)$ and $\alpha_l\left(\cdot\right)$ are $\widehat{g}\left(u\right) = B_1\left(u\right)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_0$ and $\widehat{\alpha}_l\left(t\right) = B_2\left(t\right)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_l$, respectively. Next, we present the L_2 convergence rates of $\widehat{g}$ and $\widehat{\alpha}_l$. With a slight abuse of notation in using $\|\cdot\|$, let $\|\phi\| = \left\{\int_{\mathcal{S}} \phi\left(t\right)^2 dt\right\}^{1/2}$ be the L_2 norm of any square integrable real-valued function $\phi\left(t\right)$ on its support $\mathcal{S}$.

Theorem 2 *Assume that* $N^4 n^{-1} = o\left(1\right)$ *and* $N^{-2-2r} n = o\left(1\right)$ *with* $r > 3/2$ *in condition* (C2). *Then, under conditions* (C1)–(C5), *we have* $\|\widehat{g}\left(\cdot\right) - g\left(\cdot\right)\| = O_p\left(\sqrt{N/n} + N^{-r}\right)$ *and* $\|\widehat{\alpha}_l\left(\cdot\right) - \alpha_l\left(\cdot\right)\| = O_p\left(\sqrt{N/n} + N^{-r}\right)$.

The optimal order requirements in the above theorem are achieved when the number of interior knots N is chosen to be $N \asymp n^{1/(2r+1)}$. As a result, the estimators $\widehat{g}$ and $\widehat{\alpha}_l$ of the nonparametric functions g and α_l have the optimal convergence rate $O_p\left(N^{-r/(2r+1)}\right)$.

4 Penalized QIF estimation

4.1 Penalized estimators

In data analysis, the true model is often unknown. Hence, researchers have employed the penalized approach to simultaneously select relevant variables and estimate unknown parameters for partially linear single-index models (see, e.g., Xie and Huang 2009; Liang et al. 2010) and varying coefficients models (see, e.g., Li and Liang 2008; Wang et al. 2008; Wang and Xia 2009). This motivates us to propose a penalized QIF method for the proposed generalized semiparametric model. Without loss of generality, we assume that the correct model in (1) has the true regression coefficients $\boldsymbol{\beta}^0 = \left(\boldsymbol{\beta}_{(1)}^{0\mathrm{T}}, \boldsymbol{\beta}_{(2)}^{0\mathrm{T}}\right)^{\mathrm{T}}$ and $\boldsymbol{\alpha}^0 = \left(\boldsymbol{\alpha}_{(1)}^{0\mathrm{T}}, \boldsymbol{\alpha}_{(2)}^{0\mathrm{T}}\right)^{\mathrm{T}}$, where $\boldsymbol{\beta}_{(1)}^0 = \left[\beta_1^0, \left\{\left(\boldsymbol{\beta}_{(1),-1}^0\right)_{(p_1-1)\times 1}\right\}^{\mathrm{T}}\right]^{\mathrm{T}}$ is the $p_1 \times 1$ vector of non-zeros, $\boldsymbol{\beta}_{(2)}^0$ is the $(p - p_1) \times 1$ vector of zeros, $\boldsymbol{\alpha}_{(1)}^0$ is the $d_{20} \times 1$ vector of non-zeros, and $\boldsymbol{\alpha}_{(2)}^0$ is the $(d_2 - d_{20}) \times 1$ vector of zeros. Their corresponding covariates are given as $X_{ij} = \left[\left\{\left(X_{ij}^{(1)}\right)_{p_1\times 1}\right\}^{\mathrm{T}}, \left\{\left(X_{ij}^{(2)}\right)_{(p-p_1)\times 1}\right\}^{\mathrm{T}}\right]^{\mathrm{T}}$, $Z_{ij}^{(2)} = \left[\left\{\left(Z_{ij}^{(21)}\right)_{d_{20}\times 1}\right\}^{\mathrm{T}}, \left\{\left(Z_{ij}^{(22)}\right)_{(d_2-d_{20})\times 1}\right\}^{\mathrm{T}}\right]^{\mathrm{T}}$.

To find the penalized parametric estimators, we propose the penalized QIF,

$$\mathcal{L}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2} Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) + n \sum\nolimits_{s=2}^{p} p_{\lambda_{n1}}(|\beta_s|) + n \sum\nolimits_{l=d_1+1}^{d} p_{\lambda_{n2}}(|\alpha_l|), \quad (12)$$

where $Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the unpenalized objective function defined in (9) and $p_{\lambda_n}(\cdot)$ is a penalty function with a regularization parameter λ_n. There are various penalty functions available in the literature, such as the L_1 and L_2 penalties, which yield the LASSO-type (Tibshirani 1996) and ridge-type estimators (Goldstein and Smith 1974), 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 $(t)_+ = tI(t > 0)$. By minimizing $\mathcal{L}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})$, we obtain the penalized QIF estimators $\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{PQIF}} = \left(\left(\widehat{\boldsymbol{\beta}}_{(1),-1}^{\mathrm{PQIF}}\right)^{\mathrm{T}}, \left(\widehat{\boldsymbol{\beta}}_{(2)}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right)^{\mathrm{T}}$ of $\boldsymbol{\beta}_{-1} = \left(\left(\boldsymbol{\beta}_{(1),-1}\right)^{\mathrm{T}}, \left(\boldsymbol{\beta}_2\right)^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}} = \left(\left(\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}\right)^{\mathrm{T}}, \left(\widehat{\boldsymbol{\alpha}}_{(2)}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right)^{\mathrm{T}}$ of $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_{(1)}^{\mathrm{T}}, \boldsymbol{\alpha}_{(2)}^{\mathrm{T}}\right)^{\mathrm{T}}$.

To study asymptotic properties of penalized estimators, we follow the same approach for obtaining $\widehat{X}_{ij}$ and $\widehat{Z}_{ij}^{(2)}$ in Sect. 3 to get $\widehat{X}_{ij}^{(1)}$ and $\widehat{Z}_{ij}^{(21)}$. Let $\mathbf{D}_{1i}\left(\boldsymbol{\beta}_{(1)}^0\right) =$

$\left(D_{1,i1}\left(\boldsymbol{\beta}^0_{(1)}\right),\ldots,D_{1,im}\left(\boldsymbol{\beta}^0_{(1)}\right)\right)^{\mathrm{T}}$ and $\Delta_{1i}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)=\mathrm{Diag}(\widehat{\nu}_{1,ij}(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)}))$, where $D_{1,ij}\left(\boldsymbol{\beta}^0_{(1)}\right) = \left\{g'\left(\boldsymbol{\beta}^{0\mathrm{T}}_{(1)}X^{(1)}_{ij}\right)\widehat{X}^{(1)\mathrm{T}}_{ij}\mathbf{J}^0_1,\widehat{Z}^{(21)\mathrm{T}}_{ij}\right\}^{\mathrm{T}}$, $\widehat{\mu}_{1,ij}\left(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)}\right) = \vartheta^{-1}\left\{\widehat{\eta}_{1,ij}\left(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)}\right)\right\}$, $\widehat{\nu}_{1,ij}(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)})=\partial\widehat{\mu}_{1,ij}(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)})/\partial\widehat{\eta}_{1,ij}(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)})$, $\widehat{\eta}_{1,ij}(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)}) = \widetilde{g}\left(\boldsymbol{\beta}^{0\mathrm{T}}_{(1)}X^{(1)}_{ij}\right)+\sum_{l=1}^{d_1}\widetilde{\alpha}_l\left(T_{ij}\right)Z_{ij,l}+\boldsymbol{\alpha}^{0\mathrm{T}}_{(1)}Z^{(21)}_{ij}$, and $\mathbf{J}^0_1 = \begin{pmatrix}-\boldsymbol{\beta}^{0\mathrm{T}}_{(1),-1}/\sqrt{1-\left\|\boldsymbol{\beta}^0_{(1),-1}\right\|^2}\\ \mathbf{I}_{d_{10}-1}\end{pmatrix}$. In addition, let $\dot{\psi}_{n1}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)$ and $\Psi_{n1}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)$ be defined in the same manner as $\dot{\psi}_n\left(\boldsymbol{\beta}^{\mathbf{0}},\boldsymbol{\alpha}^{\mathbf{0}}\right)$ and $\Psi_n\left(\boldsymbol{\beta}^{\mathbf{0}},\boldsymbol{\alpha}^{\mathbf{0}}\right)$ in Sect. 3 by replacing their $\mathbf{D}_i\left(\boldsymbol{\beta}^0\right)$ and $\Delta_i\left(\boldsymbol{\beta}^{\mathbf{0}},\boldsymbol{\alpha}^{\mathbf{0}}\right)$ with $\mathbf{D}_{1i}\left(\boldsymbol{\beta}^0_{(1)}\right)$ and $\Delta_{1i}\left(\boldsymbol{\beta}^0_{(1)},\ \boldsymbol{\alpha}^0_{(1)}\right)$, respectively. Then, we establish the following oracle properties of the penalized estimators.

Theorem 3 *Assume that* $N^4n^{-1}=o(1)$, $N^{-4r+2}n=o(1)$ *with* $r>3/2$ *in condition* (C2), *and the tuning parameters satisfy* $\lambda_{n1}\to 0$, $\lambda_{n2}\to 0$, $n^{1/2}\lambda_{n1}\to\infty$ *and* $n^{1/2}\lambda_{n2}\to\infty$. *Then, under conditions* (C1)–(C5), *the penalized estimators satisfy*: (1) (*sparsity*) $P\left(\left\{\left(\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}_{(2)}\right)^{\mathrm{T}},\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}_{(2)}\right)^{T}\right\}^{\mathrm{T}}=\mathbf{0}\right)\to 1$; *and* (2) (*asymptotic normality*)

$$\sqrt{n}\left(\Sigma^{(1)}_{n1}\right)^{1/2}\left\{\left\{\left(\widehat{\boldsymbol{\beta}}^{PQIF}_{(1),-1}\right)^{\mathrm{T}},\left(\widehat{\boldsymbol{\alpha}}^{PQIF}_{(1)}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}-\left(\boldsymbol{\beta}^{0T}_{(1),-1},\boldsymbol{\alpha}^{0T}_{(1)}\right)^{\mathrm{T}}\right\}\to N\left(0,\mathbf{I}_{(p_1+d_{20}-1)}\right),$$

where $\Sigma^{(1)}_{n1}=\dot{\psi}_{n1}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)^{\mathrm{T}}\Psi_{n1}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)\dot{\psi}_{n1}\left(\boldsymbol{\beta}^0_{(1)},\boldsymbol{\alpha}^0_{(1)}\right)$.

Let $\boldsymbol{\Upsilon}_1=\begin{pmatrix}\mathbf{J^0_1}_{p_1\times(p_1-1)} & \mathbf{0}_{p_1\times d_{20}}\\ \mathbf{0}_{d_{20}\times(p_1-1)} & \mathbf{I}_{d_{20}\times d_{20}}\end{pmatrix}$. The above theorem, together with the multivariate delta method, leads to the asymptotic normality of penalized parametric estimators, $\sqrt{n}\Sigma^{1/2}_{n1}\left(\left\{\left(\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}_{(1)}\right)^{\mathrm{T}},\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}_{(1)}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}-\left(\boldsymbol{\beta}^{0\mathrm{T}}_{(1)},\boldsymbol{\alpha}^{0\mathrm{T}}_{(1)}\right)^{\mathrm{T}}\right)\to N\left(0,\mathbf{I}_{p_1+d_{20}}\right)$, as $n\to\infty$, where $\Sigma_{n1}=\boldsymbol{\Upsilon}_1\Sigma^{(1)}_{n1}\boldsymbol{\Upsilon}^{\mathrm{T}}_1$.

We next study the asymptotic properties of the penalized nonparametric estimators. To this end, assume that $\alpha_l(\cdot)\equiv 0$ for $(d_{10}+1)\le l\le d_1$ in the true model. By the density assumption of T_{ij} in Condition (C1) of the Appendix, we obtain that $\widetilde{\alpha}_l(\cdot)=0$ if and only if $E\left\{\widetilde{\alpha}_l\left(T_{ij}\right)^2\right\}=0$. In addition, $\alpha_l(t)\approx\widetilde{\alpha}_l(t)=B_2(t)^{\mathrm{T}}\boldsymbol{\gamma}_l$. This motivates us to consider the empirical L_2 norm as a metric, that is, $\|\widetilde{\alpha}_l\|=\left\|\boldsymbol{\gamma}_l\right\|_{\mathbf{W}_n}=\left(\boldsymbol{\gamma}^{\mathrm{T}}_l\mathbf{W}_n\boldsymbol{\gamma}_l\right)^{1/2}$, where $\mathbf{W}_n=n^{-1}_{\mathrm{T}}\sum_{i=1}^n\sum_{j=1}^m B_2\left(T_{ij}\right)B_2\left(T_{ij}\right)^{\mathrm{T}}$ and $n_{\mathrm{T}}=nm$. Using this metric and replacing $(\boldsymbol{\beta},\boldsymbol{\alpha})$ in $Q_n(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{\alpha})$ by its $\sqrt{n}$ consistent estimator $(\widehat{\boldsymbol{\beta}}^*,\widehat{\boldsymbol{\alpha}}^*)$ (e.g., $(\widehat{\beta}^{\mathrm{PQIF}},\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}})$), we adopt Wang et al.'s (2007) group-penalized approach and propose the following penalized QIF for spline coefficients,

$$\mathcal{L}_n(\boldsymbol{\gamma})=\frac{1}{2}Q_n\left(\boldsymbol{\gamma},\widehat{\boldsymbol{\beta}}^*,\ \widehat{\boldsymbol{\alpha}}^*\right)+n\sum_{l=1}^{d_1}p_{\lambda_{n3}}\left(\left\|\boldsymbol{\gamma}_l\right\|_{\mathbf{W}_n}\right). \tag{13}$$

The resulting penalized estimator of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF}} = \left\{ \left(\widehat{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}} \right)^{\mathrm{T}}, 0 \leq l \leq d_1 \right\}^{\mathrm{T}} = \arg\min_{\boldsymbol{\gamma}} \left(\mathcal{L}_n \left(\boldsymbol{\gamma} \right) \right).$$

Subsequently, we obtain the estimators of $g(u)$ and $\alpha_l(t)$, which are $\widehat{g}^{\mathrm{PQIF}}(u) = B_1(u)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_0^{\mathrm{PQIF}}$ and $\widehat{\alpha}_l^{\mathrm{PQIF}}(t) = B_2(t)^{\mathrm{T}} \, \widehat{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}}$. Then, we demonstrate the following asymptotic properties of nonparametric estimators.

Theorem 4 *Assume that* $\lambda_{n3} \to 0$ *and* $\lambda_{n3} n^{r/(2r+1)} \to \infty$ *with* $r > 3/2$ *in condition* (C2). *Then, under conditions* (C1)–(C5), $\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF}}$ *satisfies* (1) (*sparsity*) $P(\widehat{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}} = 0) \to 1$ *for* $d_{10} + 1 \leq l \leq d_1$; *and* (2) ($L_2$ *rate of convergence*) $\left\| \widehat{g}^{\mathrm{PQIF}}(\cdot) - g(\cdot) \right\| = O_p\left(N^{-r/(2r+1)}\right)$ *and* $\left\| \widehat{\alpha}_l^{\mathrm{PQIF}}(\cdot) - \alpha_l(\cdot) \right\| = O_p\left(N^{-r/(2r+1)}\right)$ *for* $1 \leq l \leq d_{10}$, *where* $N \asymp n^{1/(2r+1)}$.

Theorem 4 indicates that, under some regularity conditions, the penalized QIF estimator has the same optimal convergence rate as the unpenalized estimator. In addition, the penalized procedure is able to correctly select relevant B-spline coefficients with probability approaching 1.

4.2 Estimation algorithm

The algorithm for obtaining unpenalized estimators is a special case of the procedure to calculate penalized estimators. Hence, we only focus on the penalized estimates. To this end, we consider three possible scenarios: (i) $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are penalized, while $\boldsymbol{\gamma}$ is unpenalized; (ii) $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unpenalized, but $\boldsymbol{\gamma}$ is penalized; (iii) $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\gamma}$ are penalized. In the first scenario, let $\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i\right)$ and $\widehat{\boldsymbol{\gamma}}^i$ be the i-th iterative estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\boldsymbol{\gamma}$, respectively. For given $\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i\right)$, we employ (7) to obtain the estimator $\widehat{\boldsymbol{\gamma}}^{i+1}$ of $\boldsymbol{\gamma}$ at the $(i+1)$ th step. That is,

$$\widehat{\boldsymbol{\gamma}}^{i+1} = \widehat{\boldsymbol{\gamma}}^i - \ddot{\widetilde{Q}}_n \left(\widehat{\boldsymbol{\gamma}}^i, \widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i \right)^{-1} \dot{\widetilde{Q}}_n \left(\widehat{\boldsymbol{\gamma}}^i, \widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i \right), \tag{14}$$

where $\dot{\widetilde{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \partial \widetilde{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \boldsymbol{\gamma}$ and $\ddot{\widetilde{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \partial^2 \widetilde{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\mathrm{T}}$.

Based on $\widehat{\boldsymbol{\gamma}}^{i+1}$, we next obtain the $(i+1)$-th iterative estimators of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. To this end, we use the fact that $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a function of γ and then denote $Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in (9) as $Q_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right)$ and its associated component $\widehat{\mathbf{D}}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ results to

$$\widehat{\mathbf{D}}_i \left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1} \right) = \left[\left\{ B_1' \left(X_{i1}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^i \right)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_0^{i+1} \widehat{X}_{i1}, \ldots, B_1' \left(X_{im}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^i \right)^{\mathrm{T}} \widehat{\boldsymbol{\gamma}}_0^{i+1} \widehat{X}_{im} \right\}^{\mathrm{T}} \right.$$
$$\left. \times \mathbf{J}^{\mathbf{0}}, \left(\widehat{Z}_{i1}^{(2)}, \ldots, \widehat{Z}_{im}^{(2)} \right)^{\mathrm{T}} \right].$$

For the sake of simplicity, let $\theta = \left(\boldsymbol{\beta}_{-1}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}$, and denote $\dot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \widehat{\boldsymbol{\gamma}}) = \partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \widehat{\boldsymbol{\gamma}}) / \partial \boldsymbol{\theta}$ and $\ddot{Q}_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \widehat{\boldsymbol{\gamma}}) = \partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha}, \widehat{\boldsymbol{\gamma}}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}$.

To obtain the penalized estimate of $\boldsymbol{\theta}$, we adopt the approach of Fan and Li (2001) and obtain the locally quadratic approximation of $2\mathcal{L}_n^*\left(\boldsymbol{\beta}^{i+1}, \boldsymbol{\alpha}^{i+1}, \widehat{\boldsymbol{\gamma}}^{i+1}\right)$ in (12) as follows:

$$
\begin{aligned}
& Q_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right) + \dot{Q}_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right)^{\mathrm{T}}\left(\boldsymbol{\theta}^{i+1} - \widehat{\boldsymbol{\theta}}^i\right) \\
& + 2^{-1}\left(\boldsymbol{\theta}^{i+1} - \widehat{\boldsymbol{\theta}}^i\right)^{\mathrm{T}} \ddot{Q}_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right)\left(\boldsymbol{\theta}^{i+1} - \widehat{\boldsymbol{\theta}}^i\right)^{\mathrm{T}} \\
& + 2n \sum_{s=2}^{p} p_{\lambda_{n1}}\left(\left|\widehat{\beta}_s^i\right|\right) + n\left(\boldsymbol{\beta}_{-1}^{i+1}\right)^{\mathrm{T}} \Phi_{\lambda_{n1}}\left(\widehat{\boldsymbol{\beta}}_{-1}^i\right) \boldsymbol{\beta}_{-1}^{i+1} - n\left(\widehat{\boldsymbol{\beta}}_{-1}^i\right)^{\mathrm{T}} \Phi_{\lambda_{n1}}\left(\widehat{\boldsymbol{\beta}}_{-1}^i\right) \widehat{\boldsymbol{\beta}}_{-1}^i \\
& + 2n \sum_{l=d_1+1}^{d} p_{\lambda_{n2}}\left(\left|\widehat{\alpha}_l^i\right|\right) + n\left(\boldsymbol{\alpha}^{i+1}\right)^{\mathrm{T}} \Phi_{\lambda_{n2}}\left(\widehat{\boldsymbol{\alpha}}^i\right) \boldsymbol{\alpha}^{i+1} - n\left(\widehat{\boldsymbol{\alpha}}^i\right)^{\mathrm{T}} \Phi_{\lambda_{n2}}\left(\widehat{\boldsymbol{\alpha}}^i\right) \widehat{\boldsymbol{\alpha}}^i,
\end{aligned}
$$

where

$$
\begin{aligned}
\Phi_{\lambda_{n1}}\left(\boldsymbol{\beta}_{-1}\right) &= \operatorname{diag}\left\{p_{\lambda_{n1}}'(|\beta_2|)/|\beta_2|, \ldots, p_{\lambda_{n1}}'\left(|\beta_p|\right)/|\beta_p|\right\}, \\
\Phi_{\lambda_{n2}}(\boldsymbol{\alpha}) &= \operatorname{diag}\left\{p_{\lambda_{n2}}'\left(|\alpha_{d_1+1}|\right)/|\alpha_{d_1+1}|, \ldots, p_{\lambda_{n2}}'(|\alpha_d|)/|\alpha_d|\right\}.
\end{aligned}
$$

Minimizing the above function with respect to θ^{i+1}, we obtain that

$$
\widehat{\boldsymbol{\theta}}^{i+1} = \widehat{\boldsymbol{\theta}}^i - \left\{\ddot{Q}_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right) + 2n\Phi\left(\widehat{\boldsymbol{\theta}}^i\right)\right\}^{-1}\left\{\dot{Q}_n^*\left(\widehat{\boldsymbol{\beta}}^i, \widehat{\boldsymbol{\alpha}}^i, \widehat{\boldsymbol{\gamma}}^{i+1}\right) + 2n\Phi\left(\widehat{\boldsymbol{\theta}}^i\right)\widehat{\boldsymbol{\theta}}^i\right\}, \tag{15}
$$

where $\Phi(\boldsymbol{\theta}) = \begin{pmatrix} \Phi_{\lambda_{n1}}\left(\boldsymbol{\beta}^{(1)}\right) & \mathbf{0}_{(p-1)\times d_2} \\ \mathbf{0}_{d_2\times(p-1)} & \Phi_{\lambda_{n2}}(\boldsymbol{\alpha}) \end{pmatrix}$. Subsequently, $\widehat{\beta}_1^{i+1} = \left(1 - \left\|\widehat{\boldsymbol{\beta}}_{-1}^{i+1}\right\|^2\right)^{1/2}$. If the i-th iterative penalized estimate $\widehat{\beta}_s^i$ is close to zero (i.e., $|\widehat{\beta}_s^i| < \epsilon_1^*$ for a small threshold value ϵ_1^*), we set $\widehat{\beta}_s^{i+1} = 0$. The iteration is stopped at the $(i+1)^{\text{th}}$ step if $\left\|\widehat{\boldsymbol{\theta}}^{i+1} - \widehat{\boldsymbol{\theta}}^i\right\| < \delta_1^*$ and $\left\|\widehat{\boldsymbol{\gamma}}^{i+1} - \widehat{\boldsymbol{\gamma}}^i\right\| < \delta_1^*$ for a small threshold value δ_1^*. Accordingly, the penalized estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}} = \widehat{\boldsymbol{\beta}}^{i+1}$ and $\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}} = \widehat{\boldsymbol{\alpha}}^{i+1}$. It is noteworthy that unpenalized QIF estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\boldsymbol{\gamma}}$ can be obtained iteratively from Eqs. (14) and (15) by setting $\Phi(\widehat{\boldsymbol{\theta}}^i) = 0$ in (15).

In the second scenario, we can show that the unpenalized QIF estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are $\sqrt{n}$-consistent. Hence, we use them to replace $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\alpha}}^*$ in Eq. (13), and then employ the same techniques as those used for obtaining Eq. (15) to yield the penalized estimator $\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i+1}$ at the $(i+1)$ th step given below.

$$
\begin{aligned}
\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i+1} = \widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i} &- \left\{\ddot{Q}_n\left(\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i}, \widehat{\boldsymbol{\beta}}^{\mathrm{QIF}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{QIF}}\right) + 2n\Phi_{\lambda_{n3}}\left(\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i}\right)\right\}^{-1} \\
&\times \left\{\dot{Q}_n\left(\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i}, \widehat{\boldsymbol{\beta}}^{\mathrm{QIF}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{QIF}}\right) + 2n\Phi_{\lambda_{n3}}\left(\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i}\right)\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF},i}\right\},
\end{aligned} \tag{16}
$$

where

$$\Phi_{\lambda_{n3}}(\boldsymbol{\gamma}) = \text{diag}\left\{p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_1\|_{\mathbf{W}n}\right)/\|\boldsymbol{\gamma}_1\|_{\mathbf{W}n}, \ldots, p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_p\|_{\mathbf{W}n}\right)/\|\boldsymbol{\gamma}_p\|_{\mathbf{W}n}\right\}.$$

If the i-th iterative penalized estimator $\widehat{\boldsymbol{\gamma}}^{\text{PQIF},i}$ is close to zero (i.e., $\|\widehat{\boldsymbol{\gamma}}^{\text{PQIF},i}\|_{\mathbf{W}n} < \epsilon_2^*$ for a small threshold value ϵ_2^*), we set $\widehat{\boldsymbol{\gamma}}^{\text{PQIF},i+1} = \mathbf{0}$. The iteration stops when $\|\widehat{\boldsymbol{\gamma}}^{\text{PQIF},i+1} - \widehat{\boldsymbol{\gamma}}^{\text{PQIF},i}\| < \delta_2^*$ for a small threshold value δ_2^*, which leads to $\widehat{\boldsymbol{\gamma}}^{\text{PQIF}} = \widehat{\boldsymbol{\gamma}}^{\text{PQIF},i+1}$.

In the third scenario, we are able to demonstrate that the penalized estimators, $\widehat{\boldsymbol{\beta}}^{\text{PQIF}}$ and $\widehat{\boldsymbol{\alpha}}^{\text{PQIF}}$, obtained from the first scenario are consistent. Thus, we substitute $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\alpha}}^*$ in Eq. (13) with these estimators. Afterwards, we adopt the same procedure as given in Eq. (16) by replacing its $\widehat{\boldsymbol{\beta}}^{\text{QIF}}$ and $\widehat{\boldsymbol{\alpha}}^{\text{QIF}}$ with $\widehat{\boldsymbol{\beta}}^{\text{PQIF}}$ and $\widehat{\boldsymbol{\alpha}}^{\text{PQIF}}$, respectively, to obtain $\widehat{\boldsymbol{\gamma}}^{\text{PQIF}}$.

To facilitate computations, we recommend using the unpenalized estimators as initial estimators in the iterative equations, (14), (15), and (16). It is worth noting that the tuning parameters are unknown in those equations, and we adapt Wang et al.'s (2007) BIC criterion to choose the tuning parameters λ_{n1}, λ_{n2} and λ_{n3} in the penalized QIF procedure. They are

$$\text{BIC}(\lambda_{n1}, \lambda_{n2}) = \mathcal{L}_n^*\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}\right) + \log(n) \times \left(\widehat{p}_1 - 1 + \widehat{d}_{21}\right) \quad \text{and}$$
$$\text{BIC}(\lambda_{n3}) = \mathcal{L}_n(\widehat{\boldsymbol{\gamma}}) + \log(n) \times \left\{J_n\left(1 + \widehat{d}_{10}\right)\right\},$$

where $\widehat{d}_{21}$ and $\widehat{p}_1$ are the number of nonzero components in $\widehat{\boldsymbol{\alpha}}^{\text{PQIF}}$ and $\widehat{\boldsymbol{\beta}}^{\text{PQIF}}$, and $\widehat{d}_{10}$ is the number of nonzero estimated functions $\widehat{\alpha}_l^{\text{PQIF}}(\cdot)$. Accordingly, $\left(\widehat{\lambda}_{n1}, \widehat{\lambda}_{n2}\right) = \arg\min_{(\lambda_{n1},\lambda_{n2})}\text{BIC}(\lambda_{n1}, \lambda_{n2})$ and $\widehat{\lambda}_{n3} = \arg\min_{\lambda_{n3}}\text{BIC}(\lambda_{n3})$. In our numerical studies given below, we use cubic splines with $q = 4$ to estimate nonparametric functions. In addition, the number of interior knots is set at $N = \left[n^{1/(2q+1)}\right] + 1$, which is of the optimal order and $[a]$ denotes the greatest integer less than or equal to a. In the empirical implementations, we use the minimal and maximal values of $X_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ and T_{ij} as the two boundary points to generate B-spline basis functions $B_{1,J}(u)$ and $B_{2,J}(t)$, respectively.

5 Numerical examples

5.1 Simulation studies

In this subsection, we conduct two Monte Carlo studies to evaluate the finite sample performance of the proposed estimators. The first two example focus on scenario (i) $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are penalized, while $\boldsymbol{\gamma}$ used for computing nonparametric functions is unpenalized. In contrast, the third example addresses scenario (ii) $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unpenalized, but $\boldsymbol{\gamma}$ is penalized.

Example 1 Within each cluster, the correlated binary responses Y_{ij} are generated from a marginal logit model,

$$\text{logit}\, P\left(Y_{ij}=1\left|X_{ij}, Z_{ij}, T_{ij}\right.\right)=g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0\right)$$
$$+\sum\nolimits_{l=1}^{2}\alpha_l\left(T_{ij}\right)Z_{ij,l}+\sum\nolimits_{l=3}^{6}\alpha_l^0 Z_{ij,l}, \tag{17}$$

where $g(U)=0.5\cos(2\pi U)$, $\alpha_1(T)=0.1\cos(2\pi T)$, $\alpha_2(T)=0.1\sin(2\pi T)$, $\boldsymbol{\beta}^0=\frac{1}{\sqrt{22}}(3,3,2,0,0,0,0)^{\mathrm{T}}$, $\boldsymbol{\alpha}^0=\left(\alpha_3,\alpha_{4,}\alpha_5,\alpha_6\right)^{\mathrm{T}}=(-0.5,0,0,0.4)^{\mathrm{T}}$, $j=1,\ldots,5$, $i=1,\ldots,n$, and $n=200$ and 500. We then use the algorithm in Emrich and Piedmonte (1991) to generate correlated binary responses with an exchangeable correlation structure and the correlation parameter is 0.3 within each cluster. Furthermore, covariates $X_{ij}=\left(X_{ij,1}\ldots,X_{ij,7}\right)^{\mathrm{T}}$ are independently generated from uniform[0, 1], T_{ij} are randomly simulated from uniform[0, 1], and $\left(Z_{ij,1}\ldots,Z_{ij,6}\right)^{\mathrm{T}}$ are independently generated from $N\left(0,0.5^2\right)$. To assess the robustness of covariance setting, we consider three different working correlation structures: independent (IND), exchangeable (EX), and AR(1), although the data are simulated from the exchangeable setting.

To examine the selection performance of parametric components, we conduct 200 realizations and report the proportions of parameters correctly fitted (C), overfitted (O), and underfitted (U) as well as the average of true positives (TP), i.e., the average number of covariates being correctly selected from all possible candidates, and the average number of false positives (FP), i.e., the average number of covariates being incorrectly selected from all possible candidates. To evaluate the estimation accuracy, we compare the SCAD-penalized QIF (PQIF) estimate with the ORACLE estimate obtained by assuming that we know the zero components in $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$. The assessment measure is the median of squared errors (MSE) defined as the median of $\left\|\widehat{\boldsymbol{\beta}}_{(k)}^{\mathrm{PQIF}}-\boldsymbol{\beta}^0\right\|^2$ and the median of $\left\|\widehat{\boldsymbol{\alpha}}_{(k)}^{\mathrm{PQIF}}-\boldsymbol{\alpha}^0\right\|^2$ in 200 realizations, where $\widehat{\boldsymbol{\beta}}_{(k)}^{\mathrm{PQIF}}$ and $\widehat{\boldsymbol{\alpha}}_{(k)}^{\mathrm{PQIF}}$ are the PQIF estimates of $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ calculated in the k^{th} realization.

Tables 1 and 2 report variable selection and estimation results for $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$, respectively. Both tables show that the proportions of correctly fitted models increase and the proportions of overfitted and underfitted models decrease when the sample size becomes larger. In addition, the number of true positives is closer to the correct number of nonzero parameters and the number of false positives decreases to zero, as the sample size increases. Moreover, the difference between the PQIF and ORACLE estimates measured by MSE becomes negligible as the sample size increases. The above findings support the theoretical results. It is noteworthy that the three working correlation structures yield similar performance, although EX is the correct structure. This indicates that the PQIF estimators are robust even though the working correlation is misspecified.

To evaluate the performance of the estimates of the nonparametric functions, we next define the integrated squared error (ISE) of the estimated functions $\widehat{g}$, $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$, given as

Table 1 Variable selection and estimation results for $\boldsymbol{\beta}^0$ with the exchangeable (EX), AR(1), and independent (IND) working correlation structures in Example 1

n		Variable selection and parameter estimation						
		C	O	U	TP	FP	PQIF	ORACLE
	EX	0.800	0.120	0.080	2.920	0.190	0.0209	0.0184
200	AR(1)	0.745	0.140	0.115	2.885	0.190	0.0231	0.0184
	IND	0.720	0.175	0.105	2.895	0.265	0.0302	0.0182
	EX	1.000	0.000	0.000	3.000	0.000	0.0062	0.0062
500	AR(1)	1.000	0.000	0.000	3.000	0.000	0.0062	0.0062
	IND	0.995	0.005	0.000	3.000	0.005	0.0063	0.0063

The symbols C, O, and U denote the proportion of correct fitting, overfitting, and underfitting, respectively. The TP and FP denote the average number of true positives and false positives, respectively. The columns PQIF and ORACLE report the median of squared errors (MSEs) of the penalized and oracle estimates

Table 2 Variable selection and estimation results for α^0 with the exchangeable (EX), AR(1), and independent (IND) working correlation structures in Example 1

n		Variable selection and parameter estimate						
		C	O	U	TP	FP	PQIF	ORACLE
	EX	0.805	0.055	0.140	1.895	0.095	0.0285	0.0228
200	AR(1)	0.795	0.050	0.155	1.895	0.105	0.0291	0.0224
	IND	0.765	0.090	0.145	1.920	0.160	0.0274	0.0224
	EX	0.980	0.014	0.006	1.990	0.014	0.0103	0.0103
500	AR(1)	0.970	0.020	0.010	1.995	0.025	0.0107	0.0101
	IND	0.955	0.020	0.025	1.985	0.030	0.0108	0.0105

The symbols C, O, and U denote the proportion of correct fitting, overfitting, and underfitting, respectively. The TP and FP denote the average numbers of true positives and false positives, respectively. The columns PQIF and ORACLE report the median of squared errors (MSEs) of the penalized and oracle estimates

$$\text{ISE}\,(\widehat{g}) = (nm)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\widehat{g}\left(X_{ij}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}\right) - g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^{0}\right)\right\}^{2},$$
$$\text{ISE}\,(\widehat{\alpha}_l) = (nm)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\widehat{\alpha}_l\left(T_{ij}\right) - \alpha_l\left(T_{ij}\right)\right\}^{2}, l = 1, 2.$$

When $n = 200$, the averages of the ISEs for $\widehat{g}$, $\widehat{\alpha}_1$, and $\widehat{\alpha}_2$ across 200 realizations are 0.100, 0.207 and 0.215, respectively. As the sample size increases to 500, the corresponding averages of the ISEs decrease to 0.02, 0.048 and 0.048, which corroborates the theoretical results in Theorem 2.

Example 2 This example addresses the case where the covariates are correlated and some are discrete. To this end, we generate the response observations using model (17) with the same true parameters, nonparametric functions, and distribution of variable T_{ij} as those given in Example 1. In addition, the covariates $\left(Z_{ij,1}\ldots, Z_{ij,7}\right)^{\mathrm{T}}$ are simulated from a multivariate normal distribution with mean zero, marginal variance 0.5^2,

Table 3 Variable selection and estimation results for $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ with the exchangeable (EX), AR(1), and independent (IND) working correlation structures in Example 2

		Variable selection and parameter estimation						
		C	O	U	TP	FP	PQIF	ORACLE
	EX	0.975	0.002	0.023	2.973	0.009	0.0108	0.0104
$\boldsymbol{\beta}^0$	AR(1)	0.970	0.005	0.025	2.973	0.015	0.0136	0.0128
	IND	0.960	0.015	0.025	2.973	0.045	0.0170	0.0155
	EX	0.910	0.080	0.010	1.991	0.109	0.0132	0.0126
$\boldsymbol{\alpha}^0$	AR(1)	0.905	0.085	0.010	1.986	0.120	0.0145	0.0137
	IND	0.900	0.085	0.015	1.982	0.127	0.0168	0.0141

The symbols C, O, and U denote the proportion of correct fitting, overfitting, and underfitting, respectively. The TP and FP denote the average number of true positives and false positives, respectively. The columns PQIF and ORACLE report the median of squared errors (MSEs) of the penalized and oracle estimates

and AR(1) correlation matrix with autocorrelation coefficient 0.3, while the covariate $Z_{ij,8}$ is generated from Bernoulli(0.5). Moreover, the covariates $\left(X_{ij,1}\ldots, X_{ij,7}\right)^{\mathrm{T}}$ are simulated from the same distribution as that of $\left(Z_{ij,1}\ldots, Z_{ij,7}\right)^{\mathrm{T}}$. To assess the robustness of estimates against the working correlation, we consider three different working correlation structures: independent (IND), exchangeable (EX), and AR(1), whereas the data are simulated from the exchangeable setting.

Table 3 presents variable selection and estimation results for $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ with $n = 500$ in 200 realizations. They indicate that the proportions of correctly fitted models are closer to one and the proportions of overfitted and underfitted models are closer to zero. In addition, the number of true positives is closer to the correct number of nonzero parameters and the number of false positives is small. Moreover, the MSE values of the PQIF and ORACLE estimates are similar, which confirms our theoretical results.

Example 3 In this example, within each cluster, the correlated binary responses Y_{ij} are generated from a marginal logit model,

$$\operatorname{logit} P\left(Y_{ij}=1 \,\middle|\, X_{ij}, Z_{ij}, T_{ij}\right)=g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0\right)+\sum\nolimits_{l=1}^{5}\alpha_l\left(T_{ij}\right)Z_{ij,l}+\sum\nolimits_{l=6}^{7}\alpha_l^0 Z_{ij,l},$$

where $g(U) = 0.5\cos(2\pi U)$, $\alpha_1(T) = 0.7\cos(2\pi T)$, $\alpha_2(T) = 0.7\sin(2\pi T)$, $\alpha_l(\cdot) = 0$ for $3 \le l \le 5$, $\boldsymbol{\beta}^0 = \frac{1}{\sqrt{22}}(3, 3, 2)^{\mathrm{T}}$, $\alpha^0 = (\alpha_6, \alpha_7)^{\mathrm{T}} = (-0.5, 0.4)^{\mathrm{T}}$, $j = 1, \ldots, 5, i = 1, \ldots, n$, and $n = 200$ and 500. In addition, the binary responses are generated from an exchangeable correlation structure with the correlation parameter 0.15. Moreover, covariates X_{ij}, T_{ij} and Z_{ij} are independently simulated from the same distributions as given in Example 1.

To assess the selection performance for varying coefficient components, we conduct 200 realizations. Table 4 reports the selection and estimation results for the varying coefficients with covariates $\left(Z_{ij,1},\ldots, Z_{ij,5}\right)^{\mathrm{T}}$. It shows that the proportions of correct fittings are close to 1 (above 95 %) for all the three correlation structures for $n = 500$,

Table 4 Variable selection and estimation results for the varying coefficient functions $\alpha_l(T)$ with the exchangeable (EX), AR(1), and independent (IND) working correlation structures in Example 3

n		Variable selection and estimation				
		C	O	U	TP	FP
	EX	0.525	0.280	0.195	1.809	0.418
200	AR(1)	0.510	0.290	0.200	1.785	0.425
	IND	0.495	0.290	0.215	1.755	0.427
	EX	0.955	0.010	0.035	1.955	0.010
500	AR(1)	0.945	0.000	0.055	1.945	0.000
	IND	0.955	0.015	0.030	1.955	0.015

The symbols C, O, and U denote the proportion of correct fitting, overfitting, and underfitting, respectively. The TP and FP denote the average of true positives and the average of false positives

while they are relatively low for $n = 200$. The high proportion of correct fitting in the large sample size corroborates the model selection consistency established in Theorem 4. In addition, the number of true positives gets closer to 2, and the number of false positives decreases to zero, as the sample size n increases.

In addition to varying coefficient components, we next study the performance of parametric components. Table 5 shows the MSEs of the parameter estimates and the empirical coverage probabilities of the 95 % confidence intervals for the parametric components. All three working correlation structures result in similar average MSE values for both parameter estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Furthermore, the MSE values decrease as n increases, which confirms the consistency property of the parameter estimates. Moreover, the empirical coverage probabilities get closer to the nominal coverage level of 95 % as n increases, which corroborates the asymptotic normality of the parameter estimators. Next, we assess the overall model fitting. To this end, we define the model error (ME) as the average of the squared difference of the estimated and true conditional means of Y_{ij}. Figure 1 depicts the boxplots of the model errors by comparing the PQIF and oracle (OR) estimates, where OR is computed by assuming the true model is known a priori. It is not surprising that the model errors of the oracle estimates are smaller than those of the PQIF estimates. As the sample size gets large, however, the model errors of PQIF and OR are very similar. It is also noteworthy that the model errors are small even though $n = 200$, which demonstrates the accuracy of PQIF estimates.

Remark To study the performance of the proposed estimation and selection methods in scenario (iii), we generate data from the following model:

$$\text{logit}\, P\left(Y_{ij}=1\,\middle|X_{ij}, Z_{ij}, T_{ij}\right)=g\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0\right)+\sum\nolimits_{l=1}^{5}\alpha_l\left(T_{ij}\right)Z_{ij,l}+\sum\nolimits_{l=6}^{9}\alpha_l^0 Z_{ij,l},$$

where $\boldsymbol{\beta}^0 = \frac{1}{\sqrt{22}}(3, 3, 2, 0, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\alpha}^0 = (\alpha_6, \alpha_7, \alpha_8, \alpha_9)^{\mathrm{T}} = (-0.5, 0, 0, 0.4)^{\mathrm{T}}$, and $\alpha_l(T)$ are defined as Example 2 for $1 \le l \le 5$. In addition, covariates X_{ij}, T_{ij} and

Table 5 The average MSEs of the parameter estimates for $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ and $\alpha = (\alpha_6, \alpha_7)^T$ and the empirical coverage probabilities (CP) of the 95 % confidence intervals for parameters $(\beta_1, \beta_2, \beta_3)$ and (α_6, α_7) based on 200 realizations in Example 3

n		MSE		CP				
		$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$	β_1	β_2	β_3	α_6	α_7
	EX	0.0168	0.0267	0.855	0.865	0.835	0.915	0.935
200	AR(1)	0.0170	0.0269	0.855	0.825	0.865	0.915	0.925
	IND	0.0171	0.0271	0.865	0.865	0.860	0.910	0.940
	EX	0.0047	0.0073	0.955	0.935	0.925	0.920	0.940
500	AR(1)	0.0050	0.0078	0.955	0.920	0.920	0.915	0.945
	IND	0.0051	0.0078	0.960	0.935	0.925	0.920	0.950

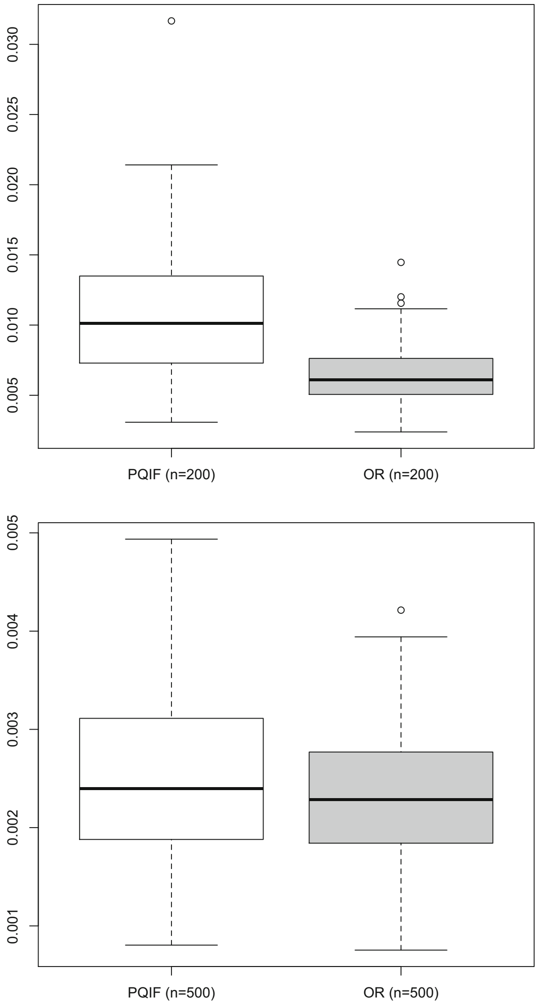

Fig. 1 Boxplots of the model errors calculated from the PQIF and oracle (OR) estimates with the EX, AR(1) and IND working correlation structures for $n = 200$ (*top panel*) and $n = 500$ (*bottom panel*)

Z_{ij} are independently simulated from the same distributions as given in Example 2, and Y_{ij} have the same correlation structure as given in Example 2. Since Monte Carlo results show similar findings as those in Examples 1 and 2, we do not present them here.

5.2 Empirical example

Following Klein et al. (1984), we consider a data set from the Wisconsin epidemiologic study of diabetic retinopathy (WESDR). The aim of this study is to investigate the risk factors for diabetic retinopathy. The response is a binary variable indicating the presence of diabetic retinopathy in each of two eyes from 720 individuals in the study. In addition, the data set contains 13 risk factors including: eye refractive error, eye intraocular pressure, age at diabetes diagnosis (years), duration of diabetes (years), glycosylated hemoglobin level, systolic blood pressure, diastolic blood pressure, body mass index , pulse rate (beats/30 s), sex (male = 1, female = 2), proteinuria (absent = 0, present = 1), doses of insulin per day taken by the patient, and type of county of residence (urban = 1, rural = 2).

Based on a preliminary fitting of the data to a logistic linear regression model, we found that there exist significant interaction effects between the logarithm of diabetes' duration, respectively, with *glycosylated hemoglobin level*, *systolic blood pressure*, and *diastolic blood pressure*, where the logarithmic transformation of diabetes duration is used to amend its right skewness. This motivates us to consider $Z_{ij,1}$ = *glycosylated hemoglobin level*, $Z_{ij,2}$ = *systolic blood pressure*, and $Z_{ij,3}$ = *diastolic blood pressure* as the covariates associated with their corresponding varying coefficients $\alpha_l\left(T_{ij}\right)$ ($l = 1, 2, 3$), where T_{ij} = *logarithm of diabetes duration*. We then assign the rest of the continuous variables to be index covariates such that $X_{ij,1}$ = *age at diagnosis of diabetes* , $X_{ij,2}$ = *body mass index* , $X_{ij,3}$ = *eye refractive error*, $X_{ij,4}$ = *eye intraocular pressure* , and $X_{ij,5}$ = *pulse rate*. The remaining categorical variables, $Z_{ij,4}$ = *sex*, $Z_{ij,5}$ = *proteinuria*, $Z_{ij,6}$ = *doses of insulin*, and $Z_{ij,7}$ = *type of county of residence*, are used as the covariates in the linear part with constant coefficients. As a result, we fit the data with the following equation,

$$\eta_{ij} = \text{logit}(\mu_{ij}) = g\left(X_{ij,1}\beta_1 + \cdots + X_{ij,5}\beta_5\right) + \sum\nolimits_{l=1}^{3} \alpha_l\left(T_{ij}\right) Z_{ij,l} + \sum\nolimits_{l=4}^{7} \alpha_l Z_{ij,l}, \quad (18)$$

where $j = 1, 2$, $i = 1, \ldots, 720$. It is worth noting that we only consider IND and EX correlation structures since there are two repeated measurements for each subject and the results are the same for EX and AR(1) structures. In addition, all continuous variables are centered and standardized for parameter estimation.

By applying the penalized QIF method in Sect. 4.1, two index variables ($X_{ij,1}$ = *age at diabetes diagnosis* and $X_{ij,2}$ = *body mass index*) and one linear variable ($Z_{ij,5}$ = *proteinuria*) are selected under the IND and EX working correlation structures. Table 6 presents the parameter estimates (EST) and their standard errors (SE) for the selected variables. The resulting Wald test statistics show that these variables are significant

Table 6 The PQIF estimates (EST) and their associated standard errors (SE) of regression coefficients for the selected variables, respectively, under the IND and EX working correlation structures for the Wisconsin epidemiologic study

		β_1	β_2	α_5
IND	EST	0.303	0.953	0.307
	SD	0.132	0.042	0.099
EX	EST	0.447	0.894	0.311
	SD	0.151	0.075	0.103

at the 5 % level. Furthermore, the estimated coefficient of *proteinuria* (0.307 in IND and 0.311 in EX) indicates that the presence of diabetic retinopathy is approximately $\exp(0.31) = 1.35$ times as frequent among proteinuria than among no-proteinuria, after adjusting for the other variables in the model.

Next, we plot the estimated index functions $\widehat{g}(\cdot)$ against the variables of *age at diabetes diagnosis* and *body mass index*, respectively, by setting the rest of their corresponding index components to zero. Figure 2 depicts the estimated functions $\widehat{g}(\cdot)$ under the IND and EX working correlation structures. The function $\widehat{g}(\cdot)$ displays a quadratic pattern over the *body mass index*, which is consistent with the findings of Barnhart and Williamson (1998) and Lian et al. (2013). For example, under the EX structure, the value of $\widehat{g}(\cdot)$ above 0 indicates that the presence of diabetic retinopathy is higher when *body mass index* lies between 2.854 and 6.397 than in the tail regions (i.e., 2.042 to 2.854 and 6.397 to 7.228). It is interesting to note that $\widehat{g}(\cdot)$ also exhibits a quadratic pattern across the variable of *age at diabetes diagnosis*, and the value of $\widehat{g}(\cdot)$ above 0 shows that the presence of diabetic retinopathy is higher when age ranges between 5.1 and 25.1 than in the tail regions (i.e., 0.4 to 5.1 and 25.1 to 29.9). Accordingly, it is not surprising that the plot of $\widehat{g}(\cdot)$ versus the index exhibits a quadratic shape. In sum, the diabetic retinopathy risk is highest in this study among people with middle values for body mass and middle values for age at diagnosis of diabetes.

We finally present the graphs of the estimated varying coefficient functions $\widehat{\alpha}_l(\cdot)$ $(l = 1, 2, 3)$ against *logarithm of diabetes duration* under the EX structure. Since the plots under the IND structure are similar to those under the EX structure, we omit them. The varying coefficient functions in Fig. 3 exhibit strong nonlinear patterns. Specifically, $\widehat{\alpha}_1(\cdot)$ and $\widehat{\alpha}_2(\cdot)$ indicate that coefficients are largest when the diabetes duration is shortest, while $\widehat{\alpha}_3(\cdot)$ has the largest coefficient around the middle values of diabetes duration. Consequently, the associated coefficients for the variables *glycosylated hemoglobin level*, *systolic blood pressure*, and *diastolic blood pressure* are not constant across different durations.

6 Discussion

In this paper, we introduce a generalized semiparametric model emerging from generalized partially linear single-index models and varying coefficient models with

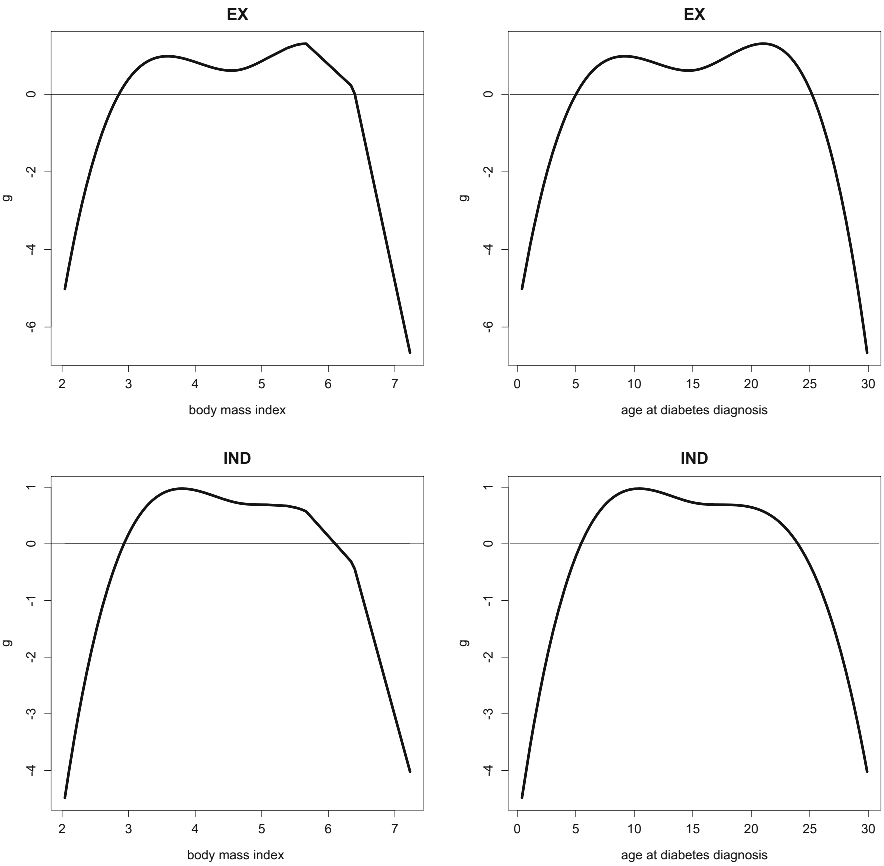

Fig. 2 Plots of $\widehat{g}(\cdot)$ against the variables *body mass index* and *age at diabetes diagnosis*, respectively, under the IND and EX working correlation structures using the Wisconsin epidemiologic study

repeated measurements. For model estimation, we propose the profile QIF estimator for the regression parameters and the QIF spline estimators for the index function and varying coefficient functions. For model selections, penalized and group-penalized estimation procedures are employed, respectively, for parametric and nonparametric functions. In addition, asymptotic consistency is studied for the resulting estimators, and asymptotic normality is further established for the parametric estimators for conducting statistical inference such as Wald test. Moreover, we demonstrate the oracle properties of the penalized estimators. Monte Carlo studies indicate that the proposed estimators perform well.

In practice, there are a few possible approaches to fit the data with model (1). Based on our limited experience, we propose the following procedures. First, place continuous variables into the single-index component and put discrete variables into either the varying coefficient component or the linear component. Second, for continuous variables, plot the estimated mean of the response variable (or the estimated single-index function) against each of them. If the plots of those variables do not depict

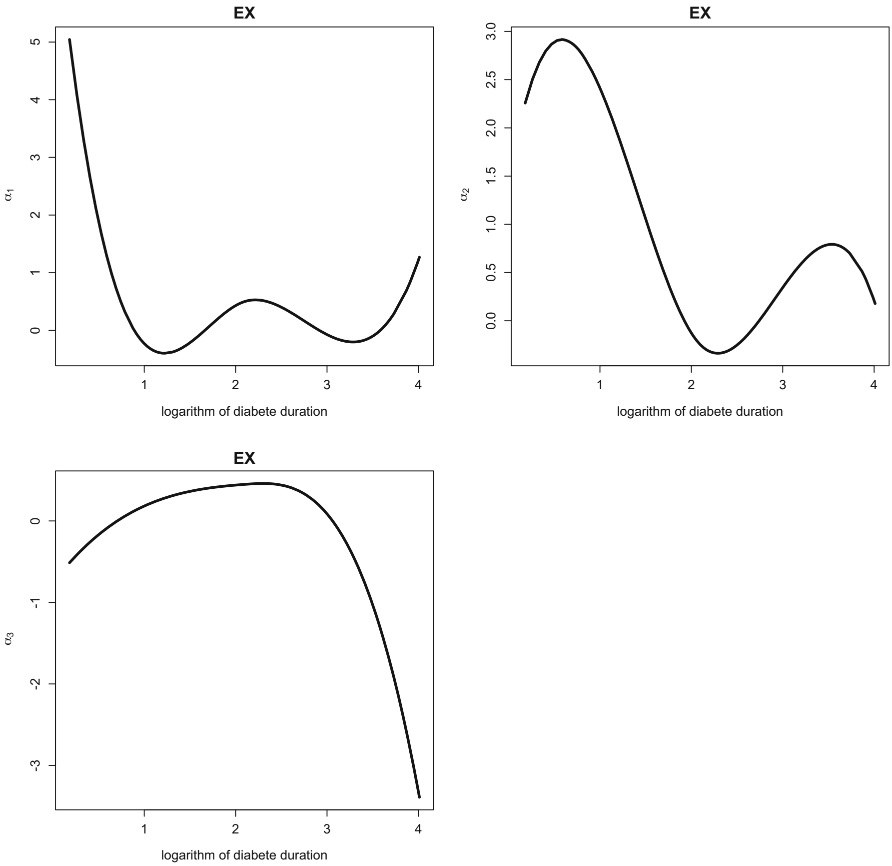

Fig. 3 Plots of $\widehat{\alpha}_l\,(\cdot)$, $l = 1, 2, 3$, against the *logarithm of diabetes duration* under the EX working correlation structure using the Wisconsin epidemiologic study

the nonlinear pattern, one can put them into either the varying coefficient component or the linear component. Third, choose the varying coefficient index, which exhibits possible interaction effects with those variables assigned in the varying coefficient component.

To extend applications of the proposed generalized semiparametric model, we identify five future research topics. The first is to generalize the penalized quadratic inference function so that one is able to estimate and select the mean components and correlation components, simultaneously. The second is to make inferences by testing the parametric and nonparametric components. The third is to adapt the approach of Stute and Zhu (2005) and then develop a test for assessing the appropriateness of model (1). The fourth is to allow the nonparametric component to be a non-smooth function. Finally, we propose applying the proposed model to the areas of quantile regression and survival analysis. We believe that these efforts would broaden the usefulness of the proposed model.

Appendix

We begin this appendix by introducing necessary notations used in the following proofs of theorems. For any positive numbers a_n and b_n, let $a_n \asymp b_n$ denote that $\lim_{n\to\infty} a_n/b_n = c$, where c is a positive constant, and let $a_n \sim b_n$ denote that $\lim_{n\to\infty} a_n/b_n = 1$. In addition, let $C^{(r)}(S) = \left\{\phi \left|\phi^{(r)} \in C(S)\right.\right\}$ be the set of the r-th order smooth functions ϕ on the support $\mathcal{S}$. For any vector $\zeta = (\zeta_1, \ldots, \zeta_s)^{\mathrm{T}} \in R^s$, denote $\|\zeta\|_\infty = \max(|\zeta_1| + \cdots + |\zeta_s|)$, and for any symmetric matrix $\mathbf{A}$, denote its L_r norm as $\|\mathbf{A}\|_r = \max_{\zeta\in R^s,\zeta\neq 0} \|\mathbf{A}\zeta\|_r \|\zeta\|_r^{-1}$. Moreover, for any matrix $\mathbf{A} = \left(A_{ij}\right)_{i=1,j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le s} \sum_{j=1}^{t} \left|A_{ij}\right|$ and $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$. To develop the theoretical results of the proposed estimators, we next present the following technical conditions.

Regularity conditions

- (C1) The density function $f_{\boldsymbol{\beta}^{\mathrm{T}} X_{ij}}\left(\boldsymbol{\beta}^{\mathrm{T}} x_{ij}\right)$ of random variable $\boldsymbol{\beta}^{\mathrm{T}} X_{ij}$ is bounded away from 0 on the support of $\boldsymbol{\beta}^{\mathrm{T}} X_{ij}$ for $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$, and the density function $f_{T_{ij}}(t)$ of random variable T_{ij} is bounded away from 0 on the support of T_{ij}.
- (C2) The true functions $g(u)$ and $\alpha_l(t)$ satisfy $g(u) \in C^{(r)}(S_U)$ and $\alpha_l(t) \in C^{(r)}(S_T)$ for $l = 1, \ldots, d_1$ and given integer $r > 3/2$, where $\mathcal{S}_U$ and $\mathcal{S}_T$ are the compact support sets of $U_{ij}\left(\boldsymbol{\beta}^0\right)$ and T_{ij}, respectively. In addition, the order of spline functions satisfies $q \ge r$.
- (C3) The eigenvalues of $\mathbf{M}_k$, $1 \le k \le \kappa$ are bounded away from 0 and infinity. Let $\Gamma = \left(\Gamma_{k,k'}\right)_{k,k'=1}^{\kappa} = \left(\Gamma_{j,j',k,k'}\right)_{j,j'=1,k,k'=1}^{m,\kappa}$. For any $1 \le j \le m$, and any given vector $a = (a_k)_{k=1}^{\kappa} \in R^\kappa$, there exist constants $0 < c_\Gamma < C_\Gamma < \infty$, such that $c_\Gamma \sum_{k=1}^{\kappa} a_k^2 \le \sum_{k,k'=1}^{\kappa} a_k a_{k'} \Gamma_{j,j,k,k'} \le C_\Gamma \sum_{k=1}^{\kappa} a_k^2$.
- (C4) The eigenvalues of $E\left((1, Z_{ij}^{(1)\mathrm{T}})^{\mathrm{T}}(1, Z_{ij}^{(1)\mathrm{T}}) \left|U_{ij}(\boldsymbol{\beta}^0) = u, T_{ij} = t\right.\right)$ are uniformly bounded away from 0 and ∞ for all $u\in S_U$ and $t \in S_T$, where $Z_{ij}^{(1)} = \left(Z_{ij,1}, \ldots, Z_{ij,d_1}\right)^{\mathrm{T}}$.
- (C5) The eigenvalues of $\dot{\psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$ and $\Psi_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$ are bounded away from 0 and infinity.

Conditions (C1) and (C2), which are given in Zhou et al. (1998), are typical assumptions in the nonparametric smoothing literature. Conditions (C3)–(C5) are needed for the convergence rates of the parametric and nonparametric estimators as well as the existence of asymptotic variances of the parametric estimators. It is worth noting that Condition (C1) ensures that the density functions are bounded away from 0 in their supports. In practice, we do not know the true support, and we may use minimum and maximum as the bounded values of the support. In addition, the parameter estimators and their asymptotic properties may not be valid in the case that Conditions (C2)–(C5) are not satisfied.

Proofs of Theorems 1 and 2

Before proving both theorems, we demonstrate the three lemmas given below.

Lemma 1 *Under Conditions* (C1) *and* (C4)*, for any* $\mathbf{a} \in R^{J_n}$ *, there exist constants* $0 < c_1 < C_1 < \infty$ *such that for* $\forall \boldsymbol{\beta} \in \Theta$ *and for sufficiently large n,*

$$c_1 N^{-1} \|\mathbf{a}\|^2 \leq \mathbf{a}^T E \left\{ \mathbf{Q}_i (\boldsymbol{\beta})^{\mathrm{T}} \mathbf{Q}_i (\boldsymbol{\beta}) \right\} \mathbf{a} \leq C_1 N^{-1} \|\mathbf{a}\|^2 , \tag{19}$$

and

$$\begin{aligned} &\max_{1 \leq J, J' \leq N+q} \left| n^{-1} \sum\nolimits_{i=1}^{n} B_{1,J} (U_i (\boldsymbol{\beta}))^{\mathrm{T}} B_{1,J'} (U_i (\boldsymbol{\beta})) \right. \\ &\quad \left. - E \left\{ B_{1,J} (U_i (\boldsymbol{\beta}))^{\mathrm{T}} B_{1,J'} (U_i (\boldsymbol{\beta})) \right\} \right| \\ &= O_{a.s.} \left\{ \sqrt{(\log n) / (nN)} \right\}, \\ &\max_{1 \leq J, J' \leq N+q} \left| n^{-1} \sum\nolimits_{i=1}^{n} B_{2,J} (T_i)^T B_{2,J'} (T_i) - E \left\{ B_{2,J} (T_i)^T B_{2,J'} (T_i) \right\} \right| \\ &= O_{a.s.} \left\{ \sqrt{(\log n) / (nN)} \right\}, \end{aligned} \tag{20}$$

where

$$\begin{aligned} B_J (U_i (\boldsymbol{\beta})) &= \left[\left\{ B_{1,J} (U_{i1} (\boldsymbol{\beta})), \ldots, B_{1,J} (U_{im} (\boldsymbol{\beta})) \right\}^{\mathrm{T}} \right]_{m \times 1}, \\ B_{2,J} (T_i) &= \left[\left\{ B_{2,J} (T_{i1}), \ldots, B_{2,J} (T_{im}) \right\}^{\mathrm{T}} \right]_{m \times 1}. \end{aligned}$$

Proof By Theorem 5.4.2 of DeVore and Lorentz (1993) and Condition (C1), we have that, for sufficiently large n and for any $\alpha = (\alpha_1, \ldots \alpha_{J_n})^{\mathrm{T}} \in R^{J_n}$, there exist constants $0 < c_1^* < C_1^* < \infty$ and $0 < c_2^* < C_2^* < \infty$ such that

$$c_1^* N^{-1} \sum\nolimits_{J=1}^{J_n} \alpha_J^2 \leq E \left\{ \sum\nolimits_{J=1}^{J_n} \alpha_J B_{1,J} (U_{ij}(\boldsymbol{\beta}^0)) \right\}^2 \leq C_1^* N^{-1} \sum\nolimits_{J=1}^{J_n} \alpha_J^2 \ \text{ and} \tag{21}$$

$$c_2^* N^{-1} \sum\nolimits_{J=1}^{J_n} \alpha_J^2 \leq E \left\{ \sum\nolimits_{J=1}^{J_n} \alpha_J B_{2,J} \left(T_{ij} \right) \right\}^2 \leq C_2^* N^{-1} \sum\nolimits_{J=1}^{J_n} \alpha_J^2. \tag{22}$$

In addition, Condition (C4) implies that, for any $\vartheta = (\vartheta_0, \vartheta_1, \ldots \vartheta_{d_1})^{\mathrm{T}} \in R^{d_1+1}$, there exist constants $0 < c_3^* < C_3^* < \infty$ such that

$$c_3^* \sum\nolimits_{l=0}^{d_1} \vartheta_l^2 \leq E \left\{ \left(\vartheta_0 + \sum\nolimits_{l=1}^{d_1} \vartheta_l Z_{ij,l})^2 | U_{ij}(\boldsymbol{\beta}^0 \right), T_{ij} \right\} \leq C_3^* \sum\nolimits_{l=0}^{d_1} \vartheta_l^2. \tag{23}$$

Let $\mathbf{a} = \left(a_{J,l} : 1 \le J \le J_n, 0 \le l \le d_1\right)$. After algebraic simplification, we have

$$\begin{aligned}
&\mathbf{a}^{\mathrm{T}} E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\} \mathbf{a} \\
&= \sum\nolimits_{j=1}^{m} \mathbf{a}^{\mathrm{T}} E\left\{Q_{ij}\left(\boldsymbol{\beta}^0\right) Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\right\} \mathbf{a} \\
&= \sum\nolimits_{j=1}^{m} E\left[\left\{\sum\nolimits_{J=1}^{J_n} a_{J,0} B_{1,J}(U_{ij}(\boldsymbol{\beta}^0))\right\} + \sum\nolimits_{l=1}^{d_1}\left\{\sum\nolimits_{J=1}^{J_n} a_{J,l} B_{2,J}\left(T_{ij}\right)\right\} Z_{ij,l}\right]^2 \\
&\le \sum\nolimits_{j=1}^{m} C_3^*\left[E\left\{\sum\nolimits_{J=1}^{J_n} a_{J,0} B_{1,J}(U_{ij}(\boldsymbol{\beta}^0))\right\}^2 + \sum\nolimits_{l=1}^{d_1} E\left\{\sum\nolimits_{J=1}^{J_n} a_{J,l} B_{2,J}\left(T_{ij}\right)\right\}^2\right] \\
&\le \sum\nolimits_{j=1}^{m} C_3^*\left\{C_1^* N^{-1} \sum\nolimits_{J=1}^{J_n} \alpha_J^2 + C_1^* N^{-1} \sum\nolimits_{l=1}^{d_1} \sum\nolimits_{J=1}^{J_n} a_{J,l}^2\right\},
\end{aligned}$$

where the first inequality of the above equation follows from (23) and the second inequality follows from (21) and (22). By letting $C_1 = mC_3^* C_1^*$, we then obtain that

$$\mathbf{a}^{\mathrm{T}} E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\} \mathbf{a} \le C_1 N^{-1} \|\mathbf{a}\|^2.$$

Applying a similar approach, we can show that

$$\mathbf{a}^{\mathrm{T}} E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\} \mathbf{a} \ge c_1 N^{-1} \|\mathbf{a}\|^2.$$

This completes the proof of (19), and the result of (20) can be obtained by Bernstein's inequality from Bosq (1998). □

Lemma 2 *Under Conditions* (C1)–(C4)*, we have (1)* $\left|\widetilde{g}\left(u, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) - g\left(u\right)\right| = O_p\left(\sqrt{N/n} + N^{-r}\right)$ *and* $\left|\widetilde{g}'\left(u, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) - g'\left(u\right)\right| = O_p\left(\sqrt{N^3/n} + N^{-r+1}\right)$ *uniformly in* $u \in S_U$*; and (2)* $|\widetilde{\alpha}_l\left(t, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) - \alpha_l\left(t\right)| = O_p\left(\sqrt{N/n} + N^{-r}\right)$ *and* $\left|\widetilde{\alpha}_l'\left(t, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) - \alpha_l'\left(t\right)\right| = O_p\left(\sqrt{N^3/n} + N^{-r+1}\right)$ *uniformly in* $t \in S_T$*, for* $1 \le l \le d_1$.

Proof For the sake of simplicity and with a slight abuse of notations, we denote $\widetilde{\boldsymbol{\gamma}} = \left(\widetilde{\boldsymbol{\gamma}}_0^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{\gamma}}_{d_1}^{\mathrm{T}}\right)^{\mathrm{T}} = \widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$, $\widetilde{\phi}_n^0 = \left\{\left(\widetilde{\phi}_{n,1}^0\right)^{\mathrm{T}}, \ldots, \left(\widetilde{\phi}_{n,\kappa}^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}} = \widetilde{\phi}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$, and $\widetilde{C}_n^0 = \widetilde{C}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)$. According to the result on page 149 of de Boor (2001), for g and α_l satisfying Condition (C2), there exists $\boldsymbol{\gamma}_l^0 \in R^{J_n}$ such that

$$\sup_{u \in S_U}\left|g\left(u\right) - g^0\left(u\right)\right| = O\left(N^{-r}\right) \text{ and } \sup_{t \in S_T}\left|\alpha_l\left(t\right) - \alpha_l^0\left(t\right)\right| = O\left(N^{-r}\right), \quad (24)$$

where $g^0\left(u\right) = B_1\left(u\right)^{\mathrm{T}} \boldsymbol{\gamma}_0^0$ and $\alpha_l^0\left(t\right) = B_2\left(t\right)^{\mathrm{T}} \boldsymbol{\gamma}_l^0$. Let $\boldsymbol{\gamma}^0 = \left(\boldsymbol{\gamma}_0^{0\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{d_1}^{0\mathrm{T}}\right)^{\mathrm{T}}$, and we then show that $\left\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right\|_\infty = o_{\mathrm{a.s.}}(1)$. By the same arguments as given in

Qu and Li (2006), we know that the global minimum for $\widetilde{Q}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)$ given in (7) exists. As a result, we only need to demonstrate that the minimizer $\widetilde{\boldsymbol{\gamma}}$ remains inside of $\mathcal{S}_{\boldsymbol{\gamma}^0}$, where $\mathcal{S}_{\boldsymbol{\gamma}^0}$ is any neighborhood of $\boldsymbol{\gamma}^0$.

Let

$$\widetilde{\varrho}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)=\left\|n^{-1}\left\{E\widetilde{C}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\}^{-1/2}\left\{E\widetilde{\phi}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\}\right\|;$$

it is noteworthy that $\widetilde{\varrho}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)$ is a continuous function in $\boldsymbol{\gamma}$. By (19) and (24), we have

$$\left\|EC_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{1/2}\right\|_2\asymp\kappa\left\|\left[E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\}\right]^{1/2}\right\|_2\asymp N^{-1/2}$$

and $\left\|E\phi_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|_2=O(N^{-r-1/2})$. Therefore,

$$\begin{aligned}\varrho_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)&\le\left\|EC_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{1/2}\right\|_2^{-1}\left\|E\phi_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|_2\\&=O(N^{1/2}N^{-r-1/2})=o\left(1\right).\end{aligned}$$

Assume that $\widetilde{\boldsymbol{\gamma}}\in\mathcal{S}^C_{\boldsymbol{\gamma}^0}$, where $\mathcal{S}^C_{\boldsymbol{\gamma}^0}$ is complement of $\mathcal{S}_{\boldsymbol{\gamma}^0}$. Then, there exists a constant $0<C<\infty$, such that

$$\widetilde{\varrho}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)=\left\|n^{-1}\left\{E\widetilde{C}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\}^{-1/2}\left\{E\widetilde{\phi}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\}\right\|>C.\quad(25)$$

Since $\widetilde{\boldsymbol{\gamma}}$ is the minimizer of $\widetilde{Q}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)$, we have that

$$\begin{aligned}&\left\|n^{-1}\widetilde{C}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1/2}\widetilde{\phi}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|\\&\quad\le\left\|n^{-1}\widetilde{C}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1/2}\widetilde{\phi}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|.\end{aligned}$$

By the strong law of large numbers, we further obtain

$$\begin{aligned}&\left\|n^{-1}\widetilde{C}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1/2}\widetilde{\phi}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|\\&\quad\to\left\|n^{-1}\left(E\widetilde{C}^0_n\right)^{-1/2}\left(E\widetilde{\phi}^0_n\right)\right\|=o\left(1\right),\end{aligned}$$

almost surely. Thus, $\left\|n^{-1}\widetilde{C}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1/2}\widetilde{\phi}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\|=o\left(1\right)$. Recall that

$$\widetilde{\phi}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)=n^{-1}\sum\nolimits_{i=1}^n\widetilde{\phi}_{in}\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)$$

as given in (6). It is also worth noting that $\widetilde{\phi}_{in}\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)$ is a continuous function of $\boldsymbol{\gamma}$, and, for all $\boldsymbol{\gamma}\in\mathcal{S}^C_{\boldsymbol{\gamma}^0}$, there exists $0<C^*<\infty$ such that

$$\begin{aligned}&\left|\left|\widetilde{\phi}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right|\right|\\&\le n^{-1}\{\textstyle\sum_{i=1}^n\widetilde{\phi}_{in}\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\widetilde{\phi}_{in}\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\}^{1/2}\\&\le C^*n^{-1}(n\kappa J_n(1+d_1))^{1/2}.\end{aligned}$$

Then, by the uniform law of large numbers, we have

$$\sup_{\boldsymbol{\gamma}\in\mathcal{S}^C_{\boldsymbol{\gamma}^0}}||\widetilde{\phi}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)-E\widetilde{\phi}_n\left(\boldsymbol{\gamma},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)||=o_{\text{a.s.}}(1).$$

This, together with the continuous mapping theorem, leads to

$$\left|n^{-1}\widetilde{C}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1/2}\widetilde{\phi}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)-\widetilde{\varrho}_n\left(\widetilde{\boldsymbol{\gamma}},\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right|=o_{\text{a.s.}}(1),$$

which contradicts with (25). Consequently, $\widetilde{\boldsymbol{\gamma}}$ remains inside of $\mathcal{S}_{\boldsymbol{\gamma}^0}$.

Using the above result and the Taylor expansion, we have

$$\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0=-\left\{\partial^2\widetilde{Q}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}\right\}^{-1}\left\{\partial\widetilde{Q}_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\gamma}\right\}\left(1+o_p(1)\right).$$

Let

$$\Omega_n=\begin{pmatrix}\Omega_{n,1}\\\vdots\\\Omega_{n,\kappa}\end{pmatrix}=n^{-1}\sum_{i=1}^n\begin{bmatrix}\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Lambda_1\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\\\vdots\\\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Lambda_\kappa\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\end{bmatrix}_{\kappa J_n(1+d_1)\times J_n(1+d_1)}\quad\text{and}$$

$$\begin{aligned}\Xi_n&=n^{-2}\\&\times\sum_{i=1}^n\begin{bmatrix}\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Gamma_{1,1}\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)&\cdots&\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Gamma_{1,\kappa}\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\\\vdots&\ddots&\vdots\\\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Gamma_{\kappa,1}\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)&\cdots&\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Gamma_{\kappa,\kappa}\Delta_i\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\end{bmatrix}_{\kappa J_n(1+d_1)\times\kappa J_n(1+d_1)}.\end{aligned}$$

By (24) and the weak law of large numbers, we have $C_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)=\Xi_n\left(1+o_p(1)\right)$. Thus,

$$\begin{aligned}\partial Q_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\alpha^0\right)/\partial\boldsymbol{\gamma}&=2\left\{\partial\phi_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}}/\partial\boldsymbol{\gamma}\right\}\Xi_n^{-1}\phi_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\\&\quad\times\left(1+o_p(1)\right)=-2\Omega_n^{\mathrm{T}}\Xi_n^{-1}\phi_n^0\left(1+o_p(1)\right)\quad\text{and}\\\partial^2Q_n\left(\boldsymbol{\gamma}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}&=2\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\left(1+o_p(1)\right).\end{aligned}\tag{26}$$

As a result,

$$\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0 = \left(n^{-1}\Omega_n^{\mathrm{T}} \Xi_n^{-1} \Omega_n\right)^{-1} \left(n^{-1}\Omega_n^{\mathrm{T}} \Xi_n^{-1} \phi_n^0\right)\left(1 + o_p(1)\right). \tag{27}$$

By (19), (20) and Condition (C3), it can be shown that, with probability approaching 1, $\|n\Xi_n\|_2 \asymp N^{-1}$ and $\sup_{1\le k\le\kappa}\left\|\Omega_{n,k}\right\|_2 \asymp N^{-1}$, and thus $\left\|n^{-1}\Xi_n^{-1}\right\|_2 \asymp N$. Moreover, by (19),

$$\begin{aligned}\left\|E\left(\Omega_n\right)^{\mathrm{T}} E\left(\Omega_n\right)\right\|_2 &= \left\|\sum_{k=1}^{\kappa} E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \Delta_i \Lambda_k \Delta_i \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\}^{\otimes 2}\right\|_2 \\ &\asymp \kappa \left\|E\left\{\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)\right\}\right\|_2^2 \asymp N^{-2}.\end{aligned}$$

This, together with (20), implies that, with probability approaching 1, $\left\|\Omega_n^{\mathrm{T}}\Omega_n\right\|_2 \asymp N^{-2}$. Accordingly, with probability approaching 1,

$$\left\|n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right\|_2 \asymp N\left\|\Omega_n^{\mathrm{T}}\Omega_n\right\|_2 \asymp N^{-1} \text{ and } \left\|\left(n^{-1}\Omega_n^{\mathrm{T}}\ \Xi_n^{-1}\Omega_n\right)^{-1}\right\|_2 \asymp N. \tag{28}$$

Next, let $\mu_i = (\mu_{i1}, \ldots, \mu_{im})^{\mathrm{T}}$. By (6), $\phi_{n,k}^0$ can be decomposed into $\phi_{n,k,e}^0 + \phi_{n,k,\mu}^0$, where

$$\begin{aligned}\phi_{n,k,e}^0 &= n^{-1}\sum_{i=1}^{n} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \widetilde{\Delta}_i\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\Lambda_k\left(Y_i - \mu_i\right),\\ \phi_{n,k,\mu}^0 &= n^{-1}\sum_{i=1}^{n} \mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}} \widetilde{\Delta}_i\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\Lambda_k\left\{\mu_i - \widetilde{\mu}_i\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\}.\end{aligned}$$

Denote $\phi_{n,e}^0 = \left\{\left(\phi_{n,1,e}^0\right)^{\mathrm{T}}, \ldots, \left(\phi_{n,\kappa,e}^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\phi_{n,\mu}^0 = \left\{\left(\phi_{n,1,\mu}^0\right)^{\mathrm{T}}, \ldots, \left(\phi_{n,\kappa,\mu}^0\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$. Accordingly, $\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0 = \left(\widetilde{\boldsymbol{\gamma}}_e + \widetilde{\boldsymbol{\gamma}}_\mu\right)\left(1 + o_p(1)\right)$, where $\widetilde{\boldsymbol{\gamma}}_e = \left(n^{-1}\Omega_n^{\mathrm{T}} \Xi_n^{-1}\Omega_n\right)^{-1}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\phi_{n,e}^0\right)$ and $\widetilde{\boldsymbol{\gamma}}_\mu = \left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right)^{-1}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\phi_{n,\mu}^0\right)$. Let $\mathbb{C} = \left(\mathbb{C}_{ij}^{\mathrm{T}}, 1 \le j \le m, 1 \le i \le n\right)^{\mathrm{T}}$. Then, for any vector $\mathbf{a} \in R^{J_n(1+d_1)}$ with $\|\mathbf{a}\| = 1$, $E\left(\mathbf{a}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_e\right) = 0$, and (28) leads to

$$E\left\{\left(\mathbf{a}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_e\right)^2 | \mathbb{C}\right\} \asymp \mathbf{a}^{\mathrm{T}}\left(\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right)^{-1}\mathbf{a} \asymp Nn^{-1}.$$

Thus, by the weak law of large numbers, $\left|\mathbf{a}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_e\right| = O_p\left(N^{1/2}n^{-1/2}\right)$. Furthermore, with probability approaching 1, there exists a constant $0 < C < \infty$, such that

$$\left|\mathbf{a}^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_\mu\right| \le C\left\|\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right)^{-1}\right\|_2 \left\|n^{-1}\Xi_n^{-1}\right\|_2 \sup_{1\le k\le\kappa}\left\|\Omega_{n,k}\right\|_2 \left\|E\left\{Q_{ij}\left(\boldsymbol{\beta}^0\right)\right\}\right\| O\left(N^{-r}\right) = O\left(N^{-r}\right).$$

The above results imply $\left|\mathbf{a}^{\mathrm{T}}\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{0}\right)\right|=O_p\left(N^{1/2}n^{-1/2}+N^{-r}\right)$. This, in conjunction with (24), ensures that $\left|\widetilde{g}\left(u,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)-g\left(u\right)\right|=O_p\left(N^{1/2}n^{-1/2}+N^{-r}\right)$ uniformly for every $u\in S_U$ and $\left|\widetilde{\alpha}_l\left(t,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)-\alpha_l\left(t\right)\right|=O_p\left(N^{1/2}n^{-1/2}+N^{-r}\right)$ uniformly for every $t\in S_T$.

To show the second part of the lemma, we employ the results on page 116 of de Boor (2001) and obtain that $\widetilde{g}'\left(u,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)=B_1^{*}\left(u\right)^{\mathrm{T}}\mathbf{D}_1\widetilde{\boldsymbol{\gamma}}_0\left(\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)$ and $\widetilde{\alpha}_l'\left(t,\ \boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)=B_2^{*}\left(t\right)^{\mathrm{T}}\mathbf{D}_1\widetilde{\boldsymbol{\gamma}}_l\left(\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)$, where $B_1^{*}\left(u\right)=\left\{B_{1,J}^{*}\left(u\right):1\leq J\leq N+q-1\right\}^{\mathrm{T}}$ is the $(q-1)$-th order B-spline basis, and $\mathbf{D}_1=\left[(q-1)\left\{\left(-\mathbf{D},\mathbf{0}_{(J_n-1)}\right)+\left(\mathbf{0}_{(J_n-1)},\mathbf{D}\right)\right\}\right]_{(J_n-1)\times J_n}$, $\mathbf{D}=\mathrm{diag}\left(d_1,\ldots,d_{N+q-1}\right)$, $d_J=\left(\xi_{q-1+J}-\xi_J\right)^{-1}$ for $1\leq J\leq N+q-1$, and $\mathbf{0}_{(J_n-1)}$ is the $(N-1)$ dimensional vector with "0 " elements, and $B_2^{*}\left(t\right)$ is defined in the same way. It is easy to prove that $\|\mathbf{D}_1\|_{\infty}=O\left(N\right)$. Applying similar techniques to those used in the proofs for $\widetilde{g}\left(u,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)$ and $\widetilde{\alpha}_l\left(t,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)$, we have that $\left|\widetilde{g}'\left(u,\boldsymbol{\beta}^{0}\ ,\boldsymbol{\alpha}^{0}\right)-g'\left(u\right)\right|=O_p\left(\sqrt{N^3/n}+N^{-r+1}\right)$ uniformly in $u\in S_U$ and $\left|\widetilde{\alpha}_l'\left(t,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)-\alpha_l'\left(t\right)\right|=O_p\left(\sqrt{N^3/n}+N^{-r+1}\right)$ uniformly in $t\in S_T$, for $1\leq l\leq d_1$, which completes the proof. □

Lemma 3 *Under Conditions* (C1)–(C4)*, we have that*

$$\frac{\partial\widehat{\eta}_{ij}\left(\boldsymbol{\beta}^{0},\alpha^{0}\right)}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\ \alpha^{\mathrm{T}}\right)^{\mathrm{T}}}=\left\{\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^{0},\boldsymbol{\beta}^{0}\ ,\alpha^{0}\right)\widehat{X}_{ij}^{\mathrm{T}}\mathbf{J}^{\mathbf{0}},\ \widehat{Z}_{ij}^{(2)T}\right\}^{\mathrm{T}}+O_p\left(N^{-r+1}+N^{3/2}n^{-1/2}\right).$$

Proof By (8), we obtain

$$\frac{\partial\widehat{\eta}_{ij}\left(\boldsymbol{\beta}^{\mathbf{0}},\boldsymbol{\alpha}^{\mathbf{0}}\right)}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\boldsymbol{\alpha}^{\ \mathrm{T}}\right)^{\mathrm{T}}}=\begin{bmatrix}\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^{0}\ ,\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)\widehat{X}_{ij}^{\mathrm{T}}\mathbf{J}^{\mathbf{0}}+\left\{Q_{ij}\left(\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\partial\widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)/\partial\boldsymbol{\beta}_{-1}^{\mathrm{T}}\right)\right\}^{\mathrm{T}}\\ Z_{ij}^{(2)}+\left\{Q_{ij}\left(\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\partial\widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta},\boldsymbol{\alpha}\right)/\partial\boldsymbol{\alpha}^{\mathrm{T}}\right)\right\}^{\mathrm{T}}\end{bmatrix}.$$

From (27), it can be shown that

$$\begin{aligned}&Q_{ij}\left(\boldsymbol{\beta}^{0}\right)^{\mathrm{T}}\left(\partial\widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}^{0},\boldsymbol{\alpha}^{0}\right)/\partial\boldsymbol{\beta}_{-1}^{\mathrm{T}}\right)\\&=-Q_{ij}\left(\boldsymbol{\beta}^{0}\right)^{\mathrm{T}}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right)^{-1}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\right)\\&\quad\times n^{-1}\sum\nolimits_{i=1}^{n}\mathbf{Q}_i\left(\boldsymbol{\beta}^{0}\right)^{\mathrm{T}}\Delta_i\Lambda_k\Delta_i\left\{g^{0\prime}\left(U_{ij}\left(\boldsymbol{\beta}^{0}\right)\right)X_{ij},1\leq j\leq m\right\}^{\mathrm{T}}\mathbf{J}^{\mathbf{0}}\\&\quad+O_p\left(N^{1/2}n^{-1/2}+N^{-r}\right)\\&=-\left\{Q_{ij}\left(\boldsymbol{\beta}^{0}\right)^{\mathrm{T}}\widehat{\boldsymbol{\eta}}_s,1\leq s\leq p\right\}\mathbf{J}^{\mathbf{0}}+O_p\left(N^{1/2}n^{-1/2}+N^{-r}\right),\end{aligned}$$

where $g^{0\prime}\left(U_{ij}\left(\boldsymbol{\beta}^0\right)\right)=B_1'\left(U_{ij}\left(\boldsymbol{\beta}^0\right)\right)^{\mathrm{T}}\boldsymbol{\gamma}_0^0$ and

$$\widehat{\boldsymbol{\zeta}}_s=\left(n^{-1}\Omega_n^{\mathrm{T}}\,\Xi_n^{-1}\Omega_n\right)^{-1}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\right)\times n^{-1}\sum_{i=1}^n\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Lambda_k\Delta_i\left\{g^{0\prime}\left(U_{ij}\left(\boldsymbol{\beta}^0\right)\right)X_{ij,s},\,1\le j\le m\right\}^{\mathrm{T}}.$$

Furthermore, by Lemma 2, we have that

$$Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\boldsymbol{\zeta}}_s=\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\vartheta}_s+O_p\left(N^{-r+1}+N^{3/2}n^{-1/2}\right),$$

where

$$\widehat{\vartheta}_s=\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\right)^{-1}\left(n^{-1}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\right)n^{-1}\sum_{i=1}^n\mathbf{Q}_i\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\Delta_i\Lambda_k\Delta_iX_{i\cdot,s},\tag{29}$$

and $X_{i\cdot,s}=\left\{X_{ij,s},\,1\le j\le m\right\}^{\mathrm{T}}$. Thus,

$$\begin{aligned}&Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\left(\partial\widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\beta}_{-1}^{\mathrm{T}}\right)\\&=-\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0,\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\left\{Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\vartheta}_s:1\le s\le p\right\}\mathbf{J}^{\mathbf{0}}\\&\quad+O_p\left(N^{-r+1}+N^{3/2}n^{-1/2}\right).\end{aligned}$$

Analogously, we can demonstrate that

$$\begin{aligned}&Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\left(\partial\widetilde{\boldsymbol{\gamma}}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\alpha}^{\mathrm{T}}\right)\\&=-\left\{Q_{ij}\left(\boldsymbol{\beta}^0\right)^{\mathrm{T}}\widehat{\boldsymbol{\eta}}_l:d_1+1\le l\le d\right\}+O_p\left(N^{-r+1}+N^{3/2}n^{-1/2}\right).\end{aligned}$$

Accordingly,

$$\frac{\partial\widehat{\eta}_{ij}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)}{\partial\left(\boldsymbol{\beta}_{-1}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}}=\left\{\widetilde{g}'\left(X_{ij}^{\mathrm{T}}\boldsymbol{\beta}^0,\boldsymbol{\beta}^0\,,\alpha^0\right)\widehat{X}_{ij}^{\mathrm{T}}\mathbf{J}^{\mathbf{0}},\widehat{Z}_{ij}^{(2)\mathrm{T}}\right\}^{\mathrm{T}}+O_p\left(N^{-r+1}+N^{3/2}n^{-1/2}\right),$$

which completes the proof. □

Proof of Theorem 1 Let $\widehat{\boldsymbol{\theta}}=\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{T}},\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\theta^0=\left(\boldsymbol{\beta}_{-1}^{0\mathrm{T}},\boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}}$. Let $\mathcal{S}(\theta^0)$ be any open set that includes θ^0. We use the same technique given in the proofs of Lemma A.2 to show that $\widehat{\boldsymbol{\theta}}$ remains inside of $\mathcal{S}(\theta^0)$, so that $||\widehat{\boldsymbol{\theta}}-\theta^0||=o_{\mathrm{a.s.}}(1)$. In the following, we demonstrate the asymptotic normality of $\widehat{\boldsymbol{\theta}}$. By the Taylor expansion, we have

$$\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0=-\left\{\partial^2Q_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}\right\}^{-1}\left\{\partial Q_n^*\left(\boldsymbol{\beta}^0,\alpha^0\right)/\partial\boldsymbol{\theta}\right\}\left\{1+o_p\left(1\right)\right\}.$$

Define $\dot{\psi}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha}) = \left\{\dot{\psi}_{n,1}^*(\boldsymbol{\beta},\boldsymbol{\alpha}), \ldots, \dot{\psi}_{n,\kappa}^*(\boldsymbol{\beta},\boldsymbol{\alpha})\right\}^{\mathrm{T}}$, where

$$\dot{\psi}_{n,k}^*(\boldsymbol{\beta},\boldsymbol{\alpha}) = n^{-1}\sum\nolimits_{i=1}^{n} \widehat{\mathbf{D}}_i^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Lambda_k\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,\widehat{\mathbf{D}}_i(\boldsymbol{\beta},\boldsymbol{\alpha}).$$

By the definition of $Q_n^*\left(\boldsymbol{\beta}^0,\alpha^0\right)$ given in (9), it can be shown that

$$\partial Q_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\theta} = -2\dot{\psi}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\Psi_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1}\psi_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) + O_p\left(n^{-1}\right)$$
$$\text{and } \partial^2 Q_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}} = 2\dot{\psi}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\Psi_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1}\dot{\psi}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) + o_p(1).$$

By Lemmas 2 and 3, we have that

$$\dot{\psi}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) = \dot{\psi}_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) + O_p\left(N^{3/2}n^{-1/2} + N^{-r+1}\right) \text{ and} \tag{30}$$
$$n\Psi_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) = \Psi_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) + O_p\left(N^{3/2}n^{-1/2} + N^{-r+1}\right). \tag{31}$$

The above results imply that

$$\partial^2 Q_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}} = 2n\dot{\psi}_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}}\Psi_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1}\dot{\psi}_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) + o_p(n). \tag{32}$$

We next define $\psi_n(\boldsymbol{\beta},\boldsymbol{\alpha}) = \left\{\psi_{n,1}(\boldsymbol{\beta},\boldsymbol{\alpha})^{\mathrm{T}}, \ldots, \psi_{n,\kappa}(\boldsymbol{\beta},\boldsymbol{\alpha})^{\mathrm{T}}\right\}^{\mathrm{T}}$, where

$$\psi_{n,k}(\boldsymbol{\beta},\boldsymbol{\alpha}) = n^{-1}\sum\nolimits_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Lambda_k\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,(Y_i - \mu_i).$$

Then, for $N^4 n^{-1} = o(1)$, $N^{-4r+2}n = o(1)$ with $r > 3/2$, and $1 \le k \le \kappa$, we employ Lemma 2 and obtain

$$\begin{aligned}
&\psi_{n,k}^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right) - \psi_{n,k}\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\\
&= n^{-1}\sum\nolimits_{i=1}^{n}\left\{\widehat{\mathbf{D}}_i^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha}) - \mathbf{D}_i^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha})\right\}\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Lambda_k\left(Y_i - \widehat{\mu}_i(\boldsymbol{\beta},\boldsymbol{\alpha})\right)\\
&\quad + n^{-1}\sum\nolimits_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Delta_i(\boldsymbol{\beta},\boldsymbol{\alpha})\,\Lambda_k\left(\mu_i(\boldsymbol{\beta},\boldsymbol{\alpha}) - \widehat{\mu}_i(\boldsymbol{\beta},\boldsymbol{\alpha})\right)\\
&= O_p\left(N^{3/2}n^{-1/2} + N^{-r+1}\right)O_p\left(n^{-1/2} + N^{1/2}n^{-1/2} + N^{-r}\right)\\
&\quad + O_p\left(n^{-1/2}\right)O_p\left(N^{1/2}n^{-1/2} + N^{-r}\right)\\
&= o_p\left(n^{-1/2}\right).
\end{aligned} \tag{33}$$

By (30), (31) and (33),

$$\partial Q_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)/\partial\boldsymbol{\theta} = -2n\dot{\psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{\mathrm{T}} \Psi_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{-1} \psi_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) + o_p\left(n^{1/2}\right). \tag{34}$$

This, together with (32), leads to

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = \left\{\widetilde{\dot{\psi}}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{\mathrm{T}} \widetilde{\Psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{-1} \widetilde{\dot{\psi}}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\}^{-1} \times \left\{\widetilde{\dot{\psi}}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{\mathrm{T}} \widetilde{\Psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)^{-1} \widetilde{\psi}_n\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\} + o_p\left(n^{-1/2}\right).$$

By the Lindeberg–Feller Central Limit Theorem and Condition (C5), we then obtain the asymptotic normality of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0$ presented in Theorem 1. □

Proof of Theorem 2 Applying Lemma 2 and the fact that $\left\| \left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}} - \left(\boldsymbol{\beta}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}} \right\| = O_p\left(n^{-1/2}\right)$, we are able to prove this theorem straightforwardly. □

Proof of Theorem 3

We consider the three steps given below to show the oracle properties of the PQIF estimators.

Step I Find the convergence rate of $\left\{\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{PQIF}}\right)^{\mathrm{T}}, \left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$. Let $\widetilde{\boldsymbol{\beta}}_{-1} = \boldsymbol{\beta}_{-1}^0 + n^{-1/2}\mathbf{v}_{-1} = \left(\widetilde{\beta}_2, \ldots, \widetilde{\beta}_p\right)^{\mathrm{T}}$, $\widetilde{\beta}_1 = \sqrt{1 - \|\widetilde{\boldsymbol{\beta}}_{-1}\|^2}$, $\widetilde{\boldsymbol{\beta}} = \left(\widetilde{\beta}_1, \widetilde{\boldsymbol{\beta}}_{-1}^{\mathrm{T}}\right)^{\mathrm{T}}$, and $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^0 + n^{-1/2}\mathbf{w} = \left(\widetilde{\alpha}_{d_1+1}, \ldots, \widetilde{\alpha}_d\right)^{\mathrm{T}}$, where $\mathbf{v} = \left(v_1, \mathbf{v}_{-1}^{\mathrm{T}}\right)^{\mathrm{T}} = \left(v_1, \ldots, v_p\right)^{\mathrm{T}}$, $\mathbf{w} = \left(w_{d_1+1}, \ldots, w_d\right)^{\mathrm{T}}$, and $\|\mathbf{v}\|_2 = \|\mathbf{w}\|_2 = C$ for some positive constant C. Denote $\widetilde{\boldsymbol{\theta}} = \left(\widetilde{\boldsymbol{\beta}}_{-1}^{\mathrm{T}}, \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$, $\theta = \left(\boldsymbol{\beta}_{-1}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}\right)^{\mathrm{T}}$, $\dot{Q}_n^*\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \partial Q_n^*\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)/\partial\boldsymbol{\theta}$, and $\ddot{Q}_n^*\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right) = \partial^2 Q_n^*\left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}$. Then,

$$Q_n^*\left(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}\right) - Q_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^{\mathrm{T}} \dot{Q}_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) + \frac{1}{2}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^{\mathrm{T}} \times \ddot{Q}_n^*\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*\right)\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right), \tag{35}$$

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 $\left(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\right)^{\mathrm{T}}$. By (32) and (34), we have, with probability approaching 1,

$$\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^{\mathrm{T}} \ddot{Q}_n^*\left(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*\right)\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right) = \left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^{\mathrm{T}} \ddot{Q}_n^*\left(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right) + O\left(C^3 n^{-1/2}\right)$$

$$
\begin{aligned}
&\asymp 2n\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right)^{\mathrm{T}} \dot{\psi}_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{\mathrm{T}} \Psi_n\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)^{-1} \dot{\psi}_n\left(\boldsymbol{\beta}^0,\alpha^0\right)\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right)\\
&\quad + o\left(C^2\right)+O\left(C^3 n^{-1/2}\right)\\
&\asymp C^2+O\left(C^3 n^{-1/2}\right)
\end{aligned}
$$

and $\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^0\right)^{\mathrm{T}} \dot{Q}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)=O_p(C)$.

Next, let $a_n=\max_{2\le s\le p}\left\{\left|p'_{\lambda_{n1}}\left(|\beta_s^0|\right)\right|, \beta_s^0\neq 0\right\}$, $b_n=\max_{2\le s\le p}\left\{\left|p''_{\lambda_{n1}}\left(|\beta_s^0|\right)\right|, \beta_s^0\neq 0\right\}$, $c_n=\max_{d_1+1\le l\le d}\left\{\left|p'_{\lambda_{n2}}\left(\left|\alpha_l^0\right|\right)\right|, \alpha_l^0\neq 0\right\}$, and $d_n=\max_{d_1+1\le l\le d}\left\{\left|p''_{\lambda_{n2}}\left(\left|\alpha_l^0\right|\right)\right|, \alpha_l^0\neq 0\right\}$. Under the assumptions that $\lambda_{n1}\to 0$ and $\lambda_{n2}\to 0$, we have that $a_n=0$ and $c_n=0$. From the Taylor expansion and the Cauchy–Schwarz inequality, as $n\to\infty$, we further have that

$$
\begin{aligned}
&-\left\{n\sum_{s=2}^{p} p_{\lambda_{n1}}\left(|\widetilde{\beta}_s|\right)-n\sum_{s=2}^{p} p_{\lambda_{n1}}\left(\left|\beta_s^0\right|\right)\right\}\\
&\quad -\left\{n\sum_{l=d_1+1}^{d} p_{\lambda_{n2}}\left(|\widetilde{\alpha}_l|\right)-n\sum_{l=d_1+1}^{d} p_{\lambda_{n2}}\left(\left|\alpha_l^0\right|\right)\right\}\\
&\le -n\sum_{s=2}^{p_1}\left\{p_{\lambda_{n1}}\left(|\widetilde{\beta}_s|\right)-p_{\lambda_{n1}}\left(\left|\beta_s^0\right|\right)\right\}-n\sum_{l=d_1+1}^{d_1+d_{20}}\left\{p_{\lambda_{n2}}\left(|\widetilde{\alpha}_l|\right)-p_{\lambda_{n2}}\left(\left|\alpha_l^0\right|\right)\right\}\\
&\le n\left(n^{-1/2}\sqrt{p_1}a_n\left\|\mathbf{v}_{-1}\right\|_2+n^{-1}b_n\left\|\mathbf{v}_{-1}\right\|_2^2+n^{-1/2}\sqrt{d_{20}}c_n\|\mathbf{w}\|_2+n^{-1}d_n\|\mathbf{w}\|_2^2\right)\\
&\le C^2\left(b_n+d_n\right). \qquad (36)
\end{aligned}
$$

When $b_n\to 0$, $d_n\to 0$, and C is sufficiently large, the second term on the right-hand side of (35) dominates its first term and (36). Accordingly, for any give $\nu>0$, there exists a large constant $\tilde{C}$ such that,

$$
P\left\{\inf_{V_{12}}\mathcal{L}_n^*\left(\boldsymbol{\beta}^0+n^{-1/2}\mathbf{v},\alpha^0+n^{-1/2}\mathbf{w}\right)>\mathcal{L}_n^*\left(\boldsymbol{\beta}^0,\boldsymbol{\alpha}^0\right)\right\}\ge 1-\nu,
$$

as $n\to\infty$, where $V_{12}=\left\{\left(\mathbf{v}^{\mathrm{T}},\mathbf{w}^{\mathrm{T}}\right)^{\mathrm{T}}:\|\mathbf{v}\|=\tilde{C}\text{ and }\|\mathbf{w}\|=\tilde{C}\right\}$. Consequently, the rate of convergence of $\left\{\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{PQIF}}\right)^{\mathrm{T}},\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ is $O_p\left(n^{-1/2}\right)$.

Step II Demonstrate the sparsity of $\left\{\left(\widehat{\boldsymbol{\beta}}_{-1}^{\mathrm{PQIF}}\right)^{\mathrm{T}},\left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$. Assume that $\boldsymbol{\beta}_{(1)}=\left\{\beta_1,\left(\boldsymbol{\beta}_{(1),-1}\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$ and $\boldsymbol{\alpha}_{(1)}$ satisfy $\left\|\boldsymbol{\beta}_{(1)}-\boldsymbol{\beta}_{(1)}^0\right\|=O_p\left(n^{-1/2}\right)$ and $\left\|\boldsymbol{\alpha}_{(1)}-\boldsymbol{\alpha}_{(1)}^0\right\|=$

$O_p\left(n^{-1/2}\right)$, respectively. We then show, with probability tending to 1, that

$$\mathcal{L}_n^*\left\{\begin{pmatrix}\boldsymbol{\beta}_{(1)}\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}\boldsymbol{\alpha}_{(1)}\\ \mathbf{0}\end{pmatrix}\right\}=\min_{\mathcal{C}}\mathcal{L}_n^*\left\{\begin{pmatrix}\boldsymbol{\beta}_{(1)}\\ \boldsymbol{\beta}_{(2)}\end{pmatrix},\begin{pmatrix}\boldsymbol{\alpha}_{(1)}\\ \boldsymbol{\alpha}_{(2)}\end{pmatrix}\right\}, \tag{37}$$

as $n\to\infty$, where $\mathcal{C}=\left\{(\boldsymbol{\beta}_{(2)}^{\mathrm{T}},\boldsymbol{\alpha}_{(2)}^{\mathrm{T}})^{\mathrm{T}}:\left\|\boldsymbol{\beta}_{(2)}\right\|\le C^*n^{-1/2} \text{ and } \left\|\boldsymbol{\alpha}_{(2)}\right\|\le C^*n^{-1/2}\right\}$ and C^* is a positive constant.

When $\beta_s\neq 0$, one has $\partial\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s=\partial Q_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s+np'_{\lambda_{n1}}(|\beta_s|)\operatorname{sgn}(\beta_s)$. By (34), it can be shown that $\partial Q_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s=O_p\left(n^{1/2}\right)$. Thus,

$$\partial\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s=n\lambda_{n1}\left\{\lambda_{n1}^{-1}n^{-1/2}+\lambda_{n1}^{-1}p'_{\lambda_{n1}}(|\beta_s|)\operatorname{sgn}(\beta_s)\right\}.$$

Using the fact that $\liminf_{n\to\infty}\liminf_{\beta_s\to 0^+}\lambda_{n1}^{-1}p'_{\lambda_{n1}}(|\beta_s|)>0$ and $n^{-1/2}\lambda_{n1}^{-1}\to 0$, we further obtain $\partial\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s>0$ for $\beta_s>0$ and $\partial\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\beta_s<0$ for $\beta_s<0$. Analogously, we can demonstrate that $\partial\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})/\partial\alpha_l>0$ for $\alpha_l>0$ and $\partial\mathcal{L}_n^*(\boldsymbol{\beta},\ \boldsymbol{\alpha})/\partial\alpha_l<0$ for $\alpha_l<0$. Consequently, the minimum of $\mathcal{L}_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})$ is attained at $\boldsymbol{\beta}_{(2)}=0$ and $\boldsymbol{\alpha}_{(2)}=0$, which proves (37). This, together with the result of Step I, implies that, with probability tending to 1, $\widehat{\boldsymbol{\beta}}_{(2)}^{\mathrm{PQIF}}=0$ and $\widehat{\boldsymbol{\alpha}}_{(2)}^{\mathrm{PQIF}}=0$, as $n\to\infty$. This completes the proof of part (i) in Theorem 3.

Step III Demonstrate the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{(1),-1}^{\mathrm{PQIF}}$ and $\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}$. Define

$$\begin{aligned}
R_{\lambda_{n1}}&=\left\{p'_{\lambda_{n1}}\left(\left|\beta_2^0\right|\right)\operatorname{sgn}\left(\beta_2^0\right),\ldots,p'_{\lambda_{n1}}\left(\left|\beta_{p_1}^0\right|\right)\operatorname{sgn}\left(\beta_{p_1}^0\right)\right\}^{\mathrm{T}},\\
\Sigma_{\lambda_{n1}}&=\operatorname{diag}\left\{p''_{\lambda_{n1}}\left(\left|\beta_2^0\right|\right),\ldots,p''_{\lambda_{n1}}\left(\left|\beta_{p_1}^0\right|\right)\right\},\\
R_{\lambda_{n2}}&=\left\{p'_{\lambda_{n2}}\left(\left|\alpha_{d_1+1}^0\right|\right)\operatorname{sgn}\left(\alpha_{d_1+1}^0\right),\ldots,p'_{\lambda_{n2}}\left(\left|\alpha_{d_1+d_{20}}^0\right|\right)\operatorname{sgn}\left(\alpha_{d_1+d_{20}}^0\right)\right\}^{\mathrm{T}},\ \text{and}\\
\Sigma_{\lambda_{n2}}&=\operatorname{diag}\left\{p''_{\lambda_{n2}}\left(\left|\alpha_{d_1+1}^0\right|\right),\ldots,p''_{\lambda_{n2}}\left(\left|\alpha_{d_1+d_{20}}^0\right|\right)\right\}.
\end{aligned}\tag{38}$$

By (37), with probability tending to 1, $\widehat{\boldsymbol{\beta}}_{(1),-1}^{\mathrm{PQIF}}$ and $\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}$ are obtained by minimizing

$$\mathcal{L}_n^*\left(\boldsymbol{\beta}_{(1)},\boldsymbol{\alpha}_{(1)}\right)=\frac{1}{2}Q_n^*\left(\boldsymbol{\beta}_{(1)},\boldsymbol{\alpha}_{(1)}\right)+n\sum\nolimits_{s=2}^{p_1}p_{\lambda_{n1}}(|\beta_s|)+n\sum\nolimits_{l=d_1+1}^{d_1+d_{20}}p_{\lambda_{n2}}(|\alpha_l|),$$

where $Q_n^*\left(\boldsymbol{\beta}_{(1)},\alpha_{(1)}\right)$ is defined similar to $Q_n^*(\boldsymbol{\beta},\boldsymbol{\alpha})$ using their nonzero components. We then have

$$\begin{aligned}
\mathbf{0}&=\begin{pmatrix}\partial\mathcal{L}_n^*\left(\widehat{\boldsymbol{\beta}}_{(1)}^{\mathrm{PQIF}},\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}\right)/\partial\boldsymbol{\beta}_{(1),-1}\\ \partial\mathcal{L}_n^*\left(\widehat{\boldsymbol{\beta}}_{(1)}^{\mathrm{PQIF}},\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}\right)/\partial\boldsymbol{\alpha}_{(1)}\end{pmatrix}\\
&=\frac{1}{2}\dot{Q}_n^*\left(\widehat{\boldsymbol{\beta}}_{(1)}^{\mathrm{PQIF}},\widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}\right)+n\begin{pmatrix}R_{\lambda_{n1}}\\ R_{\lambda_{n2}}\end{pmatrix}+n\Sigma_\lambda\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{(1),-1}^{\mathrm{PQIF}}-\boldsymbol{\beta}_{(1),-1}^0\\ \widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}-\boldsymbol{\alpha}_{(1)}^0\end{pmatrix}+O_p(1),
\end{aligned}$$

where $\Sigma_\lambda = \begin{pmatrix} \Sigma_{\lambda_{n1}} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & \Sigma_{\lambda_{n2}} \end{pmatrix}$. Subsequently, applying similar techniques to those used the proof of Theorem 1, we obtain that

$$\sqrt{n}\left(\Sigma_{n1}^{(1)}\right)^{-1/2}\left(\Sigma_{n1}^{(1)}+\Sigma_\lambda\right)\left\{\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{(1),-1}^{\mathrm{PQIF}}-\boldsymbol{\beta}_{(1),-1}^{0} \\ \widehat{\boldsymbol{\alpha}}_{(1)}^{\mathrm{PQIF}}-\boldsymbol{\alpha}_{(1)}^{0} \end{pmatrix}+\left(\Sigma_{n1}^{(1)}+\Sigma_\lambda\right)^{-1}\begin{pmatrix} R_{\lambda_{n1}} \\ R_{\lambda_{n2}} \end{pmatrix}\right\}$$
$$\to N\left(\mathbf{0}, \mathbf{I}_{(p_1+d_{20})}\right).$$

Finally, under the assumptions that $\lambda_{n1} \to 0$ and $\lambda_{n2} \to 0$, and the fact that $\sqrt{n}\Sigma_\lambda = \sqrt{n}R_{\lambda_{n1}} = \sqrt{n}R_{\lambda_{n2}} = \mathbf{0}$, we complete the proof of part (ii) in Theorem 3. □

Proof of Theorem 4

Assume that the true parameters $(\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0)$ in model (1) are known. Then, the resulting penalized estimator of $\boldsymbol{\gamma}$, $\widetilde{\boldsymbol{\gamma}}^{\mathrm{PQIF}} = \left\{\left(\widetilde{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}}\right)^{\mathrm{T}}, 0 \le l \le d_1+1\right\}^{\mathrm{T}}$, is obtained by minimizing the following penalized QIF:

$$\mathcal{L}_n(\boldsymbol{\gamma}) = \frac{1}{2}Q_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) + n\sum\nolimits_{l=1}^{d_1} p_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|\right).$$

Define $\widetilde{g}^{\mathrm{PQIF}}(u) = B_1(u)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_0^{\mathrm{PQIF}}$ and $\widetilde{\alpha}_l^{\mathrm{PQIF}}(t) = B_2(t)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}}$. In the following, we will show the convergence rate for $\widetilde{g}^{\mathrm{PQIF}}(\cdot)$ and $\widetilde{\alpha}_l^{\mathrm{PQIF}}(\cdot)$ as well as demonstrate the sparsity of $\widetilde{\boldsymbol{\gamma}}^{\mathrm{PQIF}}$.

Let $\widetilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^0 + \tilde{\varrho}_n\mathbf{v} = \left\{(\widetilde{\boldsymbol{\gamma}}_l)^{\mathrm{T}}, 0 \le l \le d_1+1\right\}^{\mathrm{T}}$, where $\mathbf{v} = \left\{(\mathbf{v}_l)^{\mathrm{T}}, 0 \le l \le d_1+1\right\}^{\mathrm{T}}$, $\mathbf{v}_l = \left(v_{1,l}, \ldots, v_{N+q,l}\right)^{\mathrm{T}}$, and $\|\mathbf{v}\| = C$ for some positive constant C. In addition, let $\dot{Q}_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \partial Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial\boldsymbol{\gamma}$ and $\ddot{Q}_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = \partial^2 Q_n^*(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}$. Then, we obtain

$$\begin{aligned} &Q_n\left(\widetilde{\boldsymbol{\gamma}}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) - Q_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) \\ &\quad = \left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right)^{\mathrm{T}}\dot{Q}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) + \frac{1}{2}\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right)^{\mathrm{T}}\ddot{Q}_n\left(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right), \end{aligned} \tag{39}$$

where $\boldsymbol{\gamma}^*$ lies between $\widetilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}^0$. By (26) and (28), with probability approaching 1,

$$\begin{aligned} \left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right)^{\mathrm{T}}\ddot{Q}_n\left(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^0, \alpha^0\right)\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right) &\asymp 2\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right)^{\mathrm{T}}\Omega_n^{\mathrm{T}}\Xi_n^{-1}\Omega_n\left(\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\right) \\ &\quad + O\left(nC^3\tilde{\varrho}_n^3\right) \\ &\asymp C^2\tilde{\varrho}_n^2\left(nN^{-1}\right) + nC^3\tilde{\varrho}_n^3. \end{aligned}$$

Furthermore, by the weak law of large numbers and (24), there exist constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$ such that

$$\left(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right)^{\mathrm{T}} \dot{Q}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) \leq C\tilde{\varrho}_n \left\| \dot{Q}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\right\| \\ \leq 2CC_1 n\tilde{\varrho}_n \left\|\phi_n^0\right\| \leq 2CC_1C_2 n\tilde{\varrho}_n \left(n^{-1/2} + N^{-r-1/2}\right).$$

Accordingly, by the assumption $N \asymp n^{1/(2r+1)}$ and taking $\tilde{\varrho}_n = \sqrt{N} n^{-r/(2r+1)}$, we obtain, with probability approaching 1, that

$$\left(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right)^{\mathrm{T}} \ddot{Q}_n\left(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)\left(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\right) \asymp C^2 N$$

and $(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0)^{\mathrm{T}} \dot{Q}_n\left(\boldsymbol{\gamma}^0, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right) = O\left(CN\right)$.

Next, let $\tilde{a}_n = \max_{1\leq l\leq d_1}\left\{\left|p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l^0\|\right)\right|, \boldsymbol{\gamma}_l^0 \neq \mathbf{0}\right\}$ and $\tilde{b}_n = \max_{1\leq l\leq d_1}$ $\left\{\left|p''_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l^0\|\right)\right|, \boldsymbol{\gamma}_l^0 \neq \mathbf{0}\right\}$. Under the assumptions that $\lambda_{n3} \to 0$, we have that $\tilde{a}_n = 0$. By the Taylor expansion and the Cauchy–Schwarz inequality, as $n \to \infty$, we further have that

$$\begin{aligned} &-\left\{n\sum\nolimits_{l=1}^{d_1} p_{\lambda_{n3}}\left(\|\widetilde{\boldsymbol{\gamma}}_l\|_{\mathbf{W}n}\right) - n\sum\nolimits_{l=1}^{d_1} p_{\lambda_{n3}}\left(\left\|\boldsymbol{\gamma}_l^0\right\|_{\mathbf{W}n}\right)\right\} \\ &\leq -n\sum\nolimits_{l=1}^{d_{10}}\left\{p_{\lambda_{n3}}\left(\|\widetilde{\boldsymbol{\gamma}}_l\|_{\mathbf{W}n}\right) - p_{\lambda_{n3}}\left(\left\|\boldsymbol{\gamma}_l^0\right\|_{\mathbf{W}n}\right)\right\} \\ &\leq n\tilde{b}_n\sum\nolimits_{l=1}^{d_{10}}\left\|\widetilde{\boldsymbol{\gamma}}_l - \boldsymbol{\gamma}_l^0\right\|_{\mathbf{W}n}^2 \asymp n\tilde{b}_n C^2\tilde{\varrho}_n^2 N^{-1} = C^2\tilde{b}_n N. \end{aligned} \tag{40}$$

When $\tilde{b}_n \to 0$ and C is sufficiently large, the second term on the right-hand side of (39) dominates its first term and (40). Thus, for any given $\nu > 0$, there exists a large constant C such that,

$$P\left\{\inf_V \mathcal{L}_n\left(\boldsymbol{\gamma}^0 + \tilde{\varrho}_n\mathbf{v}\right) > \mathcal{L}_n\left(\boldsymbol{\gamma}^0\right)\right\} \geq 1 - \nu,$$

as $n \to \infty$, where $V = \{\mathbf{v} : \|\mathbf{v}\| = C\}$. As a result, $\left\|\widetilde{\boldsymbol{\gamma}}^{\mathrm{PQIF}} - \boldsymbol{\gamma}^0\right\| = O_p\left(\tilde{\varrho}_n\right)$, which leads to

$$\left\|\widetilde{g}^{\mathrm{PQIF}}\left(\cdot\right) - g\left(\cdot\right)\right\| \asymp N^{-1/2}\left\|\widetilde{\boldsymbol{\gamma}}_0^{\mathrm{PQIF}} - \boldsymbol{\gamma}_0^0\right\| = O_p(N^{-r/(2r+1)})$$

and $\left\|\widetilde{\alpha}_l^{\mathrm{PQIF}}\left(\cdot\right) - \alpha_l\left(\cdot\right)\right\| \asymp N^{-1/2}\left\|\widetilde{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}} - \boldsymbol{\gamma}_l^0\right\| = O_p\left(N^{-r/(2r+1)}\right)$.

Finally, let $\boldsymbol{\gamma} = \left\{\left(\boldsymbol{\gamma}_{(1)}^{\mathrm{T}}\right)_{(d_{10}+1)\times 1}, \left(\boldsymbol{\gamma}_{(2)}^{\mathrm{T}}\right)_{(d_1-d_{10})\times 1}\right\}^{\mathrm{T}}$. We then show that, with probability tending to 1,

$$\mathcal{L}_n\left\{\left(\boldsymbol{\gamma}_{(1)}^{\mathrm{T}}, \boldsymbol{\gamma}_{(2)}^{\mathrm{T}}\right)^{\mathrm{T}}\right\} = \min_{\mathcal{C}} \mathcal{L}_n\left\{\left(\boldsymbol{\gamma}_{(1)}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}}\right)^{\mathrm{T}}\right\},$$

as $n \to \infty$, where $\mathcal{C} = \left\{\| \boldsymbol{\gamma}_{(2)} \| \leq C^* \varrho_n\right\}$ and C^* is a positive constant. When $\|\boldsymbol{\gamma}_l\| \neq 0$, there exists a constant $0 < c < \infty$ such that, with probability approaching 1,

$$\begin{aligned}
\partial \mathcal{L}_n(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}_l &= \partial Q_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)/\partial \boldsymbol{\gamma}_l + n p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}\right) \|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}^{-1} \mathbf{W}_n \boldsymbol{\gamma}_l \\
&\asymp \partial Q_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)/\partial \boldsymbol{\gamma}_l + c N^{-1} n p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}\right) \boldsymbol{\gamma}_l \\
&= \partial Q_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)/\partial \boldsymbol{\gamma}_l + c n^{2r/(2r+1)} p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}\right) \boldsymbol{\gamma}_l.
\end{aligned}$$

By (26), it can be shown that $\partial Q_n\left(\boldsymbol{\gamma}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0\right)/\partial \boldsymbol{\gamma}_l = O_p\left(n^{r/(2r+1)}\right)$. As a result,

$$\partial \mathcal{L}_n(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}_l = n^{2r/(2r+1)} \lambda_{n3}\left\{\lambda_{n3}^{-1} n^{-r/(2r+1)} + \lambda_{n3}^{-1} p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}\right) \boldsymbol{\gamma}_l\right\}.$$

Using the fact that $\lambda_{n3}^{-1} n^{-r/(2r+1)} \to 0$ and $\liminf_{n\to\infty} \liminf_{\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n} \to 0^+} \lambda_{n3}^{-1} p'_{\lambda_{n3}}\left(\|\boldsymbol{\gamma}_l\|_{\mathbf{W}n}\right) > 0$, we further obtain $\partial \mathcal{L}_n(\boldsymbol{\gamma})/\partial \gamma_{J,l} > 0$ for $\gamma_{J,l} > 0$ and $\partial \mathcal{L}_n(\boldsymbol{\gamma})/\partial \gamma_{J,l} < 0$ for $\gamma_{J,l} < 0$. Consequently, the minimum of $\partial \mathcal{L}_n(\boldsymbol{\gamma})$ is attained at $\boldsymbol{\gamma}_l = \mathbf{0}$ for $(d_{10}+1) \leq l \leq d_1$. This implies, with probability tending to 1, $\widetilde{\boldsymbol{\gamma}}_l^{\mathrm{PQIF}} = \mathbf{0}$ for $(d_{10}+1) \leq l \leq d_1$. Subsequently, using the fact that $\left\|\left\{\left(\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}\right)^{\mathrm{T}}, \left(\widehat{\boldsymbol{\alpha}}^{\mathrm{PQIF}}\right)^{\mathrm{T}}\right\}^{\mathrm{T}} - \left(\boldsymbol{\beta}^{0\mathrm{T}}, \boldsymbol{\alpha}^{0\mathrm{T}}\right)^{\mathrm{T}}\right\| = O_p\left(n^{-1/2}\right)$ and those assumptions given in Theorem 4, the above results of convergence rate and sparsity can be applied to the penalized estimators $\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF}}$, $\widehat{g}^{\mathrm{PQIF}}(\cdot)$, and $\widehat{\alpha}_l^{\mathrm{PQIF}}(\cdot)$. This completes the proof. □

Acknowledgments The authors thank the Editor, the Associate and two anonymous referees for their helpful comments that led to significant improvements in the article.

References

Barnhart, H. X., Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics, 54*, 720–729.

Boente, G., He, X., Zhou, J. (2006). Robust estimates in generalized partially linear models. *The Annals of Statistics, 34*, 2856–2878.

Bosq, D. (1998). *Nonparametric statistics for stochastic processes*. New York: Springer.

Brillinger, D. R. (1983). A generalized linear model with “Gaussian” regression variables. In P. J. Bickel, K. A. Doksum, J. L. Hodges (Eds.), *A Festschrift for Erich L. Lehmann in honor of his sixty-fifth birthday* (pp. 97–114). California: Wadsworth.

Cai, Z., Fan, J., Li, R. (2000a). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, *95*, 888–902.

Cai, Z., Fan, J., Yao, Q. (2000b). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*, 941–956.

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

Chen, R., Tsay, R. S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, *88*, 289–308.

Cleveland, W. S., Grosse, E., Shyu, W. M. (1991). Local regression models. In J. M. Chambers, T. J. Hastie (Eds.), *Statistical Models in S* (pp. 309–376). Wadsworth & Brooks: California.
Cui, X., Härdle, W. K., Zhu, L. (2011). The EFM approach for single-index models. *The Annals of Statistics*, *39*, 1658–1688.
Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York: Springer.
de Boor, C. (2001). *A practical guide to splines. Applied mathematical sciences* (revised ed., Vol. 27). New York: Springer.
DeVore, R. A., Lorentz, G. G. (1993). *Constructive approximation*. Berlin: Springer.
Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.
Emrich, L. J., Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, *45*, 302–304.
Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.
Fan, J., Li, R. (2004). New estimation and model selection procedures for semiparametric modelling in longitudinal data analysis. *Journal of the American Statistical Association*, *99*, 710–723.
Fan, J., Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, *11*, 1031–1057.
Fan, J., Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface*, *1*, 179–195.
Fan, J., Lin, H., Zhou, Y. (2006). Local partial-likelihood estimation for lifetime data. *The Annals of Statistics*, *34*, 290–325.
Goldstein, M., Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society, Series B, 36*, 284–291.
Härdle, W., Liang, H., Gao, J. (2000). *Partially linear models*. Heidelberg: Physica.
Hastie, T. J., Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B, 55*, 757–796.
Horowitz, J. L. (1998). *Semiparametric methods in econometrics*. New York: Springer.
Huang, J. Z., Wu, C. O., Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika, 89*, 111–128.
Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D., DeMets, D. L. (1984). The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Opthalmology, 102*, 520–526.
Li, R., Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics, 36*, 261–286.
Liang, H., Liu, X., Li, R., Tsai, C. L. (2010). Estimation and testing for partially linear single-index models. *The Annals of Statistics, 38*, 3811–3836.
Lian, H., Liang, H., Wang, L. (2013). Generalized additive partial linear models for clustered data with diverging number of covariates using GEE. *Statistica Sinica, 24*, 173–196.
Liang, K. Y., Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.
Lin, D. Y., Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association, 96*, 103–126.
Lin, X., Carroll, R. J. (2006). Semiparametric estimation in general repeated measure problems. *Journal of the Royal Statistical Society, Series B, 68*, 68–88.
Ma, S. (2012). Two-step spline estimating equations for generalized additive partially linear models with large cluster sizes. *The Annals of Statistics, 40*, 2943–2972.
Ma, S., Song, P. X. K. (2014). Varying index coefficient models. *Journal of the American Statistical Association*. doi:10.1080/01621459.2014.903185.
Ma, Y., Zhu, L. (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society, Series B, 75*, 305–322.
McCullagh, P., Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall.
Qu, A., Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics, 62*, 379–391.
Qu, A., Lindsay, B. G., Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika, 87*, 823–836.

Speckman, P. (1998). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B, 50*, 413–436.
Stute, W., Zhu, L.-X. (2009). Nonparametric checks for single-index models. *The Annals of Statistics, 33*, 1048–1083.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.
Wang, H., Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association, 104*, 747–757.
Wang, H., Li, R., Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika, 94*, 553–568.
Wang, L., Li, H., Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association, 103*, 1556–1569.
Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika, 90*, 43–52.
Xia, Y. C., Härdle, W. (2006). Semi-parametric estimation of partially single-index models. *Journal of Multivariate Analysis, 97*, 1162–1184.
Xie, H., Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics, 37*, 673–696.
Xu, P., Zhu, L. (2012). Estimation for a marginal generalized single-index longitudinal model. *Journal of Multivariate Analysis, 105*, 285–299.
Xue, L., Qu, A., Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association, 105*, 1518–1530.
Yan, Y., Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association, 97*, 1042–1054.
Zeng, P., He, T., Zhu, Y. (2012). A lasso-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics, 21*, 92–109.
Zhang, W., Lee, S. Y., Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis, 82*, 166–188.
Zhang, Y., Li, R., Tsai, C. L. (2010). Regularization parameter selection via generalized information criterion. *Journal of the American Statistical Association, 105*, 312–323.
Zhou, J., Qu, A. (2012). In semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association, 96*, 103–126.
Zhou, S., Shen, X., Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics, 26*, 1760–1782.