The Annals of Statistics
2014, Vol. 42, No. 2, 592–624
DOI: 10.1214/13-AOS1194

ESTIMATION AND MODEL SELECTION IN GENERALIZED ADDITIVE PARTIAL LINEAR MODELS FOR CORRELATED DATA WITH DIVERGING NUMBER OF COVARIATES[1]

BY LI WANG[2], LAN XUE[3], ANNIE QU[4] AND HUA LIANG[5]

University of Georgia, Oregon State University, University of Illinois at Urbana-Champaign and George Washington University

We propose generalized additive partial linear models for complex data which allow one to capture nonlinear patterns of some covariates, in the presence of linear components. The proposed method improves estimation efficiency and increases statistical power for correlated data through incorporating the correlation information. A unique feature of the proposed method is its capability of handling model selection in cases where it is difficult to specify the likelihood function. We derive the quadratic inference function-based estimators for the linear coefficients and the nonparametric functions when the dimension of covariates diverges, and establish asymptotic normality for the linear coefficient estimators and the rates of convergence for the nonparametric functions estimators for both finite and high-dimensional cases. The proposed method and theoretical development are quite challenging since the numbers of linear covariates and nonlinear components both increase as the sample size increases. We also propose a doubly penalized procedure for variable selection which can simultaneously identify nonzero linear and nonparametric components, and which has an asymptotic oracle property. Extensive Monte Carlo studies have been conducted and show that the proposed procedure works effectively even with moderate sample sizes. A pharmacokinetics study on renal cancer data is illustrated using the proposed method.

1. Introduction.

We encounter longitudinal data in many social and health studies where observations from clustered data are measured over time, and can often be discrete, such as binary or count data. Generalized additive partial linear models (GAPLM) are developed to model partial linear additive components while the remaining components are modeled nonparametrically [11] to combine the strengths of both the GPLM and the GAM for interpretability and flexibility.

Received June 2013; revised November 2013.

[1] The first two authors contribute equally to the paper.

[2] Supported in part by NSF Grants DMS-09-05730, DMS-11-06816 and DMS-13-09800.

[3] Supported in part by NSF Grant DMS-09-06739.

[4] Supported in part by NSF Grants DMS-09-06660 and DMS-13-08227.

[5] Supported in part by NSF Grant DMS-12-07444 and by Award Number 11228103, given by National Natural Science Foundation of China.

MSC2010 subject classifications. Primary 62G08; secondary 62G10, 62G20, 62J02, 62F12.

Key words and phrases. Additive model, group selection, model selection, oracle property, partial linear models, polynomial splines, quadratic inference function, SCAD, selection consistency.

Efficient estimation of linear and nonparametric function components is quite challenging even for cross-sectional data. To solve the "curse of dimensionality" problem in computing, [30] suggested a penalized regression splines approach to utilize the practical benefits of smoothing spline methods and the computational advantages of local scoring backfitting [2]. In addition, [25] applied polynomial splines to approximate the nonparametric components, and estimated coefficients through an efficient one-step procedure of maximizing the quasi-likelihood function. This can reduce computational costs significantly compared to the local scoring backfitting and marginal integration approaches. Another advantage of the polynomial spline approach is that it can formulate a penalized function for variable selection purposes, which cannot be easily implemented through other iterative methods.

However, [25]'s approach is valid only for independent data and the case with a fixed number of covariates for linear component model selection. In this paper, we develop a general framework for estimation and variable selection using the GAPLM. The proposed method can handle correlated categorical responses in addition to continuous ones, and allows both the number of covariates for linear and nonlinear terms to diverge as the sample size increases. Note that the theoretical development for model selection and estimation for diverging number of covariates in nonlinear components are completely different from the setting with finite dimension of covariates [33].

The GAPLM can be highly computationally intensive as it introduces high-dimensional nuisance parameters associated with nonparametric forms. Incorporating correlation structure brings additional challenges to modeling and estimation due to the additional correlation parameters involved. The extension of the GAPLM for correlated data imposes more challenges computationally and theoretically. However, it is well known that ignoring correlation could lead to inefficient estimation and diminish statistical power in hypothesis testing and the selection of correct models. Moreover, [28] and [36] indicate that in nonparametric settings ignoring the correlation could also result in biased estimation since the selection process is rather sensitive to small departures from the true correlation structure, and likely to cause overfitting of the nonparametric estimator to compensate for the overall bias. These problems could be more critical for the GAPLM since in contrast to the parametric setting, the true model here might be more difficult to verify. The proposed polynomial spline approach can efficiently take the within-cluster correlation into account because of its nonlocal behavior in longitudinal data [29]. This is substantially different from the kernel smoothing method, where only local data points are used in the estimation and, therefore, it cannot incorporate correlation structure efficiently.

We propose variable selection and estimation simultaneously based on the penalized quadratic inference function for correlated data when the dimension of covariates in GAPLM increases as the sample size. The quadratic inference function (QIF) [23] utilizes within-cluster correlation into account without specifying the likelihood function, and is less sensitive to the misspecification of working correlation matrices compared to the generalized estimating equation (GEE) method [19], in general. In addition, we perform variable selection for the marginal GAPLM to identify important variables, which is crucial to obtain efficient estimators for the nonzero components. We show that the proposed model selection for both parametric and nonparametric terms is consistent, the estimators of the nonzero linear coefficients are asymptotically normal, and the estimators of the nonzero nonparametric functions are L_2-norm consistent with the optimal rate of convergence if the dimension of nonparametric components is finite. However, the asymptotic properties on the rate of convergence are no longer the same as in [25] when the dimensions of covariates for parametric and nonparametric components both diverge as the sample size increases.

The semiparametric model containing both linear and nonparametric functions makes the estimation and model selection very different from the generalized additive model [33], which involves only nonparametric components. The establishment of the asymptotic normal distribution of the estimators for the parametric terms is quite challenging given that the number of covariates for both parametric and nonparametric terms diverge, and the convergence rate for the nonparametric component estimators is slower than $\sqrt{n}$. Another difficulty here is that the covariates in the parametric components and those in the nonparametric components could be dependent, in addition to dependent errors for repeated measurements, so traditional nonparametric tools such as the backfitting algorithm [2] cannot be applied here. In contrast, the proposed spline-based approach allows one to incorporate correlation effectively even when the number of covariates diverges.

In addition, the required techniques using the penalized quadratic distance function for the diverging numbers of linear and nonlinear covariates setting are very different from existing approaches such as the penalized least-squares approach for a finite dimension setting [20, 25, 31]; the generalized linear model selection approach for the parametric term only with diverging number of covariates [5]; or the GAPLM for a finite number of nonparametric functions [18], which does not perform model selection for the nonparametric term. This motivates us to develop new theoretical tools to derive large sample properties for linear and nonparametric components estimation and model selection to incorporate the dependent nature of the data for handling diverging numbers of covariates.

We organize the paper as follows. Section 2 presents the model framework, describes estimation procedures, and establishes asymptotic properties of the GAPLM for correlated data. Section 3 proposes a penalized QIF method for simultaneous estimation and variable selection when the dimension of covariates increases as the sample size. The theoretical properties on model selection consistency and rate of convergence for the nonparametric estimators are developed, in addition to algorithm implementation and tuning parameter selection. Sections 4 and 5 illustrate the performance of the proposed method through simulation studies and a pharmacokinetics study on renal cancer patients, respectively. We provide concluding remarks and discussion in Section 6. The proofs of the theorems along with technical lemmas are provided in the Appendix and supplementary material [27].

2. Estimation procedures and theoretical results.

2.1. *The GAPLM for correlated data.* For the clustered data, let Y_{it} be a response variable, $\mathbf{X}_{it} = (X_{it}^{(1)}, \dots, X_{it}^{(d_x)})^{\mathrm{T}}$ and $\mathbf{Z}_{it} = (1, Z_{it}^{(1)}, \dots, Z_{it}^{(d_z-1)})^{\mathrm{T}}$ be the d_x-vector and d_z-vector of covariates corresponding to the nonparametric and parametric components, respectively, where t is the tth ($t = 1, \dots, T_i$) observation for the ith ($i = 1, \dots, n$) cluster. Further denote $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})^{\mathrm{T}}$, $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT_i})^{\mathrm{T}}$, and $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{iT_i})^{\mathrm{T}}$. For presentation simplicity, we assume each cluster has the same size with $T_i = T < \infty$. The procedure for data with unequal cluster sizes can be adjusted following the same method of [33].

One of the advantages of marginal approaches is that we only need to specify the first two moments by $E(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_{it}) = \mu_{it}$, and $\mathrm{Var}(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_{it}) = \phi V(\mu_{it})$, where ϕ is a scale parameter and $V(\cdot)$ is a known variance function. Here, the marginal mean μ_{it} associates with the covariates through the known link function $g(\cdot)$ such that

$$\eta_{it} = g(\mu_{it}) = \sum_{l=1}^{d_x} \alpha_l(X_{it}^{(l)}) + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta}$ is d_z-vector of unknown parameters, and $\{\alpha_l(\cdot)\}_{l=1}^{d_x}$ are unknown smooth functions. Model (2.1) is called the generalized additive partial linear model (GAPLM), where $\mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta}$ are the parametric components, and $\sum_{l=1}^{d_x} \alpha_l(X_{it}^{(l)})$ are the nonparametric components. Here, the mean of Y_{it} depends only on the covariate vector for the tth observation [22], that is, $E(Y_{it}|\mathbf{X}_i, \mathbf{Z}_i) = E(Y_{it}|\mathbf{X}_{it}, \mathbf{Z}_{it})$. In addition, without loss of generality, we assume that each covariate $\{X^{(l)}\}_{l=1}^{d}$ can be rescaled into $[0, 1]$; and each $\alpha_l(\cdot)$ is centered with $\int_0^1 \alpha_l(x)\,dx = 0$ to make model (2.1) identifiable.

2.2. *Spline approximation.* We approximate smooth functions $\{\alpha_l(\cdot)\}_{l=1}^{d_x}$ in (2.1) by polynomial splines for their simplicity in computation, and they often provide a good approximation of smooth functions with a limited number of knots. For example, for each $1 \le l \le d_x$, let υ_l be a partition of $[0, 1]$, with N_n interior knots $\upsilon_l = \{0 = \upsilon_{l,0} < \upsilon_{l,1} < \cdots < \upsilon_{l,N_n} < \upsilon_{l,N_n+1} = 1\}$.

The polynomial splines of order $p + 1$ are functions with p-degree (or less) of polynomials on intervals $[\upsilon_{l,i}, \upsilon_{l,i+1})$, $i = 0, \dots, N_n - 1$, and $[\upsilon_{l,N_n}, \upsilon_{l,N_n+1}]$, and have $p - 1$ continuous derivatives globally. Let $\varphi_l = \varphi^p([0, 1], \upsilon_l)$ be the space of such polynomial splines, and $\varphi_l^0 = \{s \in \varphi_l : \int_0^1 s(x)\,dx = 0\}$. This ensures that the spline functions are centered.

Let $\{B_{lj}(\cdot)\}_{j=1}^{J_n}$ be a set of spline bases of φ_l^0 with the dimension of $J_n = N_n + p$. We approximate the nonparametric component $\alpha_l(\cdot)$ by a polynomial spline, that is $\alpha_l(\cdot) \approx s_l(\cdot) = \sum_{j=1}^{J_n} \gamma_{lj} B_{lj}(\cdot)$, with a set of coefficients $\boldsymbol{\gamma}_l = (\gamma_{l1}, \dots, \gamma_{lJ_n})^{\mathrm{T}}$. Accordingly, η_{it} is approximated by

$$\eta_{it}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{l=1}^{d_x} \sum_{j=1}^{J_n} \gamma_{lj} B_{lj}\bigl(X_{it}^{(l)}\bigr) + \mathbf{Z}_{it}^{\mathrm{T}} \boldsymbol{\beta},$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \dots, \boldsymbol{\gamma}_{d_x}^{\mathrm{T}})^{\mathrm{T}}$. Therefore, the mean function μ_{it} in (2.1) can be approximated by

$$\mu_{it} \approx \mu_{it}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = g^{-1}\left\{\sum_{l=1}^{d_x} \sum_{j=1}^{J_n} \gamma_{lj} B_{lj}\bigl(X_{it}^{(l)}\bigr) + \mathbf{Z}_{it}^{\mathrm{T}} \boldsymbol{\beta}\right\}.$$

We denote $\boldsymbol{\mu}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \{\mu_{i1}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \dots, \mu_{iT}(\boldsymbol{\beta}, \boldsymbol{\gamma})\}^{\mathrm{T}}$ in matrix notation. To incorporate the within-cluster correlation, we apply the QIF to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the parametric and nonparametric parts, respectively.

2.3. *Quadratic inference functions.* To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, one may use the GEE method [18], that is, using a working correlation matrix R which depends on fewer nuisance parameters. The estimates of regression parameters are consistent even when R is misspecified. However, one has to find a consistent estimator of R to obtain an efficient estimator of $\boldsymbol{\beta}$. The QIF approach [23] considers the approximation of $\mathbf{R}^{-1}$ with a linear combination of basis matrices of form $\mathbf{R}^{-1} \approx a_1\mathbf{M}_1 + \cdots + a_K\mathbf{M}_K$. For example, if $\mathbf{R}$ has an exchangeable structure with correlation ρ, then $\mathbf{R}^{-1}$ can be represented as $a_1\mathbf{I} + a_2\mathbf{M}_2$ with $\mathbf{I}$ being the identity matrix and $\mathbf{M}_2$ being a matrix with 0 on the diagonal and 1 off the diagonal. The corresponding coefficients are $a_1 = -\{(T-2)\rho + 1\}/k_1$, and $a_2 = \rho/k_1$, where $k_1 = (T-1)\rho^2 - (n-2)\rho - 1$ and T is the dimension of $\mathbf{R}$. The basis matrices are also available to approximate $\mathbf{R}^{-1}$ of other structure such as, AR-1 and the block diagonal correlation structures. If the candidate basis matrices represent a sufficiently rich class for the true structure, [35] show that the correlation structure can be selected consistently by minimizing the penalized difference between two estimating functions generated from the empirical correlation information and the model-based approximation, respectively. The penalization on the basis matrices ensures that an optimal number of basis matrices K will be selected to capture correlation information, yet not be burdened by too many moment conditions.

The quadratic inference function is established under the same principle as the generalized method of moments [10], and is shown to be the most efficient among estimators given the same class of estimating functions as the asymptotic variance reaches the minimum in the sense of Loewner ordering. This is especially useful under misspecified working correlation structures, since the true correlation structure is seldom known. For example, the QIF estimator is shown to be more efficient than the GEE estimator for diverging number of covariates under the generalized linear model framework [5]. Another advantage of the QIF is that the estimation of the linear coefficients a_i's is not required. In nonparametric modeling with diverging number of covariates, it is even more beneficial if we can avoid estimating the nuisance parameters associated with the correlations, since we are dealing with high-dimensional parameters involved in nonparametric components.

2.4. *Estimation procedure.* For any $\mathbf{x} \in R^{d_x}$, $\mathbf{z} \in R^{d_z}$, let $\mathbf{B}^{\mathrm{T}}(\mathbf{x}) = (B_{11}(x_1), \ldots, B_{1J_n}(x_1), \ldots, B_{d_x 1}(x_{d_x}), \ldots, B_{d_x J_n}(x_{d_x}))$, $\mathbf{D}^{\mathrm{T}}(\mathbf{x}, \mathbf{z}) = (\mathbf{z}^{\mathrm{T}}, \mathbf{B}^{\mathrm{T}}(\mathbf{x}))$ be vectors of dimensions $d_x J_n$ and $d_x J_n + d_z$, respectively. In addition, we denote matrices $\mathbf{B}_i = \{(\mathbf{B}(\mathbf{X}_{i1}), \ldots, \mathbf{B}(\mathbf{X}_{iT}))^{\mathrm{T}}\}_{T \times d_x J_n}$, $\mathbf{D}_i = \{(\mathbf{D}(\mathbf{X}_{i1}, \mathbf{Z}_{i1}), \ldots, \mathbf{D}(\mathbf{X}_{iT}, \mathbf{Z}_{iT}))^{\mathrm{T}}\}_{T \times (d_x J_n + d_z)}$.

For $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, we define $K(d_x J_n + d_z)$-dim extended scores to incorporate correlation for correlated data as follows:

$$\mathbf{g}_i(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Delta}_i \mathbf{A}_i^{-1/2} \mathbf{M}_1 \mathbf{A}_i^{-1/2} \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})\} \\ \vdots \\ \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Delta}_i \mathbf{A}_i^{-1/2} \mathbf{M}_K \mathbf{A}_i^{-1/2} \{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})\} \end{pmatrix}, \tag{2.2}$$

where $\boldsymbol{\Delta}_i = \operatorname{diag}\{\dot{\mu}_{i1}, \ldots, \dot{\mu}_i\}$ and $\dot{\mu}_{it}$ is the first order derivative of g^{-1} evaluated at $\mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it})\boldsymbol{\gamma} + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta}$; and $\mathbf{A}_i = \operatorname{diag}\{V(\mu_{i1}), \ldots, V(\mu_i)\}$. We define the sample mean and sample variance of the moment conditions as

$$\mathbf{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta}), \qquad \mathbf{C}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta})\mathbf{g}_i^{\mathrm{T}}(\boldsymbol{\theta}). \tag{2.3}$$

If we set $\mathbf{G}_n(\boldsymbol{\theta}) = 0$ as our estimating equations, there are more equations than the number of unknown parameters, and the parameters are over-identified. The QIF approach estimates $\alpha_l(\cdot)$ and $\boldsymbol{\beta}$ by making $\mathbf{G}_n$ as close to zero as possible, in the sense of minimizing the QIF $Q_n(\boldsymbol{\theta})$, that is,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} &= ((\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}})^{\mathrm{T}}, (\widehat{\boldsymbol{\gamma}}^{\mathrm{QIF}})^{\mathrm{T}})^{\mathrm{T}} \\ &= \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \{n \mathbf{G}_n^{\mathrm{T}}(\boldsymbol{\theta}) \mathbf{C}_n^{-1}(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta})\}. \end{aligned} \tag{2.4}$$

Consequently, for any $\mathbf{x} \in [0, 1]^{d_x}$ and $l = 1, \ldots, d_x$, the estimators of the nonparametric components in (2.1) are provided as

$$\widehat{\alpha}_l^{\mathrm{QIF}}(x^{(l)}) = \sum_{j=1}^{J_n} \widehat{\gamma}_{lj}^{\mathrm{QIF}} B_{lj}(x^{(l)}) \quad \text{and} \quad \widehat{\alpha}^{\mathrm{QIF}}(\mathbf{x}) = \sum_{l=1}^{d_x} \widehat{\alpha}_l^{\mathrm{QIF}}(x^{(l)}). \tag{2.5}$$

The advantages of the spline basis approach lie not only in its computation efficiency, but also in the ease of implementation. Using the spline basis approximation, we can easily convert a problem with infinite-dimensional parameters to one with a finite number of parameters [17]. In the following Theorem 1, we also show that the proposed estimators of the nonparametric components using polynomial spline achieve the optimal rate of convergence. This result is useful for providing an initial consistent estimator for later development in simultaneous variable selection and estimation for both parametric and nonparametric functions.

2.5. *Asymptotic properties.* We establish the asymptotic properties of the QIF estimators, summarize the main results in the following theorems and provide detailed proofs in the Appendices. Note that the asymptotic results still hold for unequal cluster size data.

In the following, denote the true nonparametric components by $\alpha_{0,l}$, $1 \le l \le d_x$ and the true parameters for the parametric components by $\boldsymbol{\beta}_0$. Let $\mu_{0,it}$ be the true marginal means. In addition, let $\boldsymbol{\mu}_{0,i} = (\mu_{0,i1}, \dots, \mu_{0,iT})^{\mathrm{T}}$ and $\mathbf{e}_i = \mathbf{Y}_i - \boldsymbol{\mu}_{0,i}$. Let $\boldsymbol{\Gamma}_{0,i}^{(k)} = \boldsymbol{\Delta}_{0,i}\mathbf{V}_{0,i}^{(k)}\boldsymbol{\Delta}_{0,i}$, where $\mathbf{V}_{0,i}^{(k)} = \mathbf{A}_{0,i}^{-1/2}\mathbf{M}_k\mathbf{A}_{0,i}^{-1/2}$ and $\boldsymbol{\Delta}_{0,i}$, $\mathbf{A}_{0,i}$ are evaluated at $\boldsymbol{\mu}_{0,i}$. Similarly, define $\boldsymbol{\mu}_0$, $\mathbf{e}$, $\boldsymbol{\Gamma}_0^{(k)}$ as the generic versions of $\boldsymbol{\mu}_{0,i}$, $\mathbf{e}_i$ and $\boldsymbol{\Gamma}_{0,i}^{(k)}$, respectively, for $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$. Let $d_n = d_x J_n + d_z$, and $\rho_n = ((1-\delta)/2)^{(d_x-1)/2}$, for some constant $\delta \in (0,1)$. Further, we denote $a \asymp b$, if there exist constants $c \ge c^* > 0$ such that $c^* b \le a \le cb$.

THEOREM 1. *Under conditions* (C1)–(C3), (C5)–(C8) *in Appendix* A.2, *if* $d_x/\log(n) \to 0$, $n^{-1/4}d_z \to 0$, $J_n \asymp n^b$, *for some* $1/(4r) \le b < 1/4$ *with the smoothing parameter* $r > 1$ *defined in condition* (C1), *the estimators* $\widehat{\alpha}_l^{\mathrm{QIF}}(x^{(l)})$, $1 \le l \le d_x$, *defined in* (2.5) *satisfy*

$$\frac{1}{n}\sum_{l=1}^{d_x}\sum_{i=1}^{n}\sum_{t=1}^{T}\{\widehat{\alpha}_l^{\mathrm{QIF}}(x_{it}^{(l)}) - \alpha_{0,l}(x_{it}^{(l)})\}^2 = O_P(n^{-1}d_n + J_n^{-2r}d_x),$$

where r determines the smoothness of the nonparametric functions. In particular, if $J_n \asymp n^{1/(2r+1)}$ *and* $d_z = O(J_n d_x)$, *then*

$$\frac{1}{n}\sum_{l=1}^{d_x}\sum_{i=1}^{n}\sum_{t=1}^{T}\{\widehat{\alpha}_l^{\mathrm{QIF}}(x_{it}^{(l)}) - \alpha_{0,l}(x_{it}^{(l)})\}^2 = O_P(n^{-2r/(2r+1)}d_x).$$

REMARK 1. Note that $d_n = d_x J_n + d_z$, so if the number of nonparametric functions, d_x, is finite, and $J_n \asymp n^{1/(2r+1)}$, then we obtain an optimal convergence rate $n^{-2r/(2r+1)}$. In addition, for a cluster size of 1, this reduces to a special case where the responses are independent, and is the same as in [14] and [31] for independent data.

Next, we establish the asymptotic normal distribution for the parametric estimator. We denote $\mathbf{g}_{0,i} = (\mathbf{g}_{0,i1}^{\mathrm{T}}, \ldots, \mathbf{g}_{0,iK}^{\mathrm{T}})^{\mathrm{T}}$ with $\mathbf{g}_{0,ik} = \mathbf{D}_i^{\mathrm{T}}\boldsymbol{\Delta}_{0,i}\mathbf{V}_{0,i}^{(k)}\mathbf{e}_i$, the value of $\mathbf{g}_i$ in (2.2) at $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{0,i}$. Similarly, let

$$\mathbf{G}_n^0 = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_{0,i}, \qquad \mathbf{C}_n^0 = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_{0,i}\mathbf{g}_{0,i}^{\mathrm{T}}, \qquad \mathbf{Q}_n^0 = n(\mathbf{G}_n^0)^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\mathbf{G}_n^0 \tag{2.6}$$

be the corresponding values of $\mathbf{G}_n$, $\mathbf{C}_n$ and $\mathbf{Q}_n$ defined in (2.3) and (2.4) at $\boldsymbol{\mu}_i = \boldsymbol{\mu}_{0,i}$. Next, denote $\widehat{\mathbf{Z}}_i = \mathbf{Z}_i - \mathrm{Proj}_{\Gamma_n}\mathbf{Z}_i$, where Proj_{Γ_n} is the projection onto the empirically centered additive spline space. See (S.17) for the exact formula of $\widehat{\mathbf{Z}}_i$. Further denote

$$\widehat{\mathbf{J}}_{\mathrm{DZ}}^{(k)} = \frac{1}{n}\sum_{i=1}^n \mathbf{D}_i^{\mathrm{T}}\boldsymbol{\Gamma}_{0,i}^{(k)}\widehat{\mathbf{Z}}_i, \qquad \widehat{\mathbf{J}}_{\mathrm{DZ}} = \{(\widehat{\mathbf{J}}_{\mathrm{DZ}}^{(1)})^{\mathrm{T}}, \ldots, (\widehat{\mathbf{J}}_{\mathrm{DZ}}^{(K)})^{\mathrm{T}}\}^{\mathrm{T}}, \tag{2.7}$$

$$\mathbf{W}_i^{(k)} = \mathbf{D}_i^{\mathrm{T}}\boldsymbol{\Delta}_{0,i}\mathbf{V}_{0,i}^{(k)}\boldsymbol{\Sigma}_i^{1/2}, \qquad \mathbf{W}_i = \{(\mathbf{W}_i^{(1)})^{\mathrm{T}}, \ldots, (\mathbf{W}_i^{(K)})^{\mathrm{T}}\}^{\mathrm{T}}. \tag{2.8}$$

In what follows, $\mathbf{A}^{\otimes 2}$ and $\mathbf{A}_{\mathbf{B}}^{\otimes 2}$ stand for $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}\mathbf{B}\mathbf{A}^{\mathrm{T}}$ for any matrix/vector $\mathbf{A}$ and square matrix $\mathbf{B}$, respectively.

THEOREM 2. *Assume that conditions* (C1)–(C3), (C5)–(C9) *in Appendix* A.2 *are satisfied, if* $d_x/\log(n) \to 0$, $n^{-1/5}d_z \to 0$, *and* $J_n \asymp n^b$, *for some* $1/(2r+1) \leq b < 1/5$, *where the smoothing parameter* $r > 2$, *then the estimator* $\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}}$ *of* $\boldsymbol{\beta}_0$ *is consistent and* $\sqrt{n}\mathbf{A}_n\boldsymbol{\Sigma}_n^{-1/2}(\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}} - \boldsymbol{\beta}_0) \to^D N(0, \boldsymbol{\Sigma}_A)$, *where* $\mathbf{A}_n$ *is any* $q \times d_z$ *matrix with a finite* q *such that* $\mathbf{A}_n^{\otimes 2}$ *converges to a* $q \times q$ *nonnegative symmetric* $\boldsymbol{\Sigma}_A$, *and* $\boldsymbol{\Sigma}_n = \widehat{\boldsymbol{\Psi}}_n^{-1}\widehat{\boldsymbol{\Omega}}_n\widehat{\boldsymbol{\Psi}}_n^{-1}$ *with*

$$\widehat{\boldsymbol{\Psi}}_n = \widehat{\mathbf{J}}_{\mathrm{DZ}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\widehat{\mathbf{J}}_{\mathrm{DZ}} \quad \textit{and} \quad \widehat{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{i=1}^n \{\widehat{\mathbf{J}}_{\mathrm{DZ}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\mathbf{W}_i\}^{\otimes 2}. \tag{2.9}$$

To establish the asymptotic properties of the QIF estimators for diverging number of covariates, a crucial step is to obtain the upper and lower bounds of the eigenvalues of the matrix $\mathbf{C}_n^{-1}(\boldsymbol{\theta})$ in (2.3) and (2.4). Note that $\mathbf{C}_n(\boldsymbol{\theta})$ is a random matrix with increasing dimension of linear and nonlinear components as n increases. The derivation of its bounds relies heavily on Lemma 1 of [24]; see [25, 33]. When d_x is finite, the term $\rho_n = ((1-\delta)/2)^{(d_x-1)/2}$ in Lemma 1 of [24] is a constant, which makes the derivation of the bounds relatively easy. However, this is no longer true in the diverging case since ρ_n goes to zero as d_x goes to infinity, and it requires special techniques for asymptotic derivations. Another major difficulty in the derivation of Theorem 2 is to resolve the dependence between $\mathbf{X}$ and $\mathbf{Z}$ in addition to establishing the convergence results for the first- and second-order partial derivatives of the quadratic inference function, which could be infinite-dimensional.

3. Penalized QIF for marginal GAPLM. In this section, we define predictor variables X_l and Z_k as redundant in model (2.1), if and only if $\alpha_l(X_l) = 0$ and $\beta_k = 0$. Suppose there is only an unknown subset of predictor variables which is relevant in model (2.1) with nonzero components, we are interested in identifying such subsets of relevant predictors consistently while estimating the nonzero parameters and functions in (2.1) simultaneously.

3.1. *Model selection.* To perform model selection for the GAPLM, we propose the penalized quadratic inference function in (2.4) which shrinks small components of estimated functions to zero. Through consistent model selection, we are able to improve the efficiency of estimators for the nonzero components since the correlation within clusters is taken into account. We define the penalized QIF (PQIF) estimator as

$$\begin{aligned}&((\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}})^{\mathrm{T}}, (\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF}})^{\mathrm{T}})^{\mathrm{T}} \\ &\quad = \operatorname*{arg\,min}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left\{ Q_n(\boldsymbol{\beta}, \boldsymbol{\gamma}) + n \sum_{l=1}^{d_x} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) + n \sum_{l=1}^{d_z} p_{\lambda_{2,n}}(|\beta_l|) \right\},\end{aligned}$$

where $p_{\lambda_{\bullet,n}}(\cdot)$ are given penalty functions of tuning parameters $\lambda_{\bullet,n}$, and $\|\boldsymbol{\gamma}_l\|^2_{\mathbf{K}_l} = \boldsymbol{\gamma}_l^{\mathrm{T}} \mathbf{K}_l \boldsymbol{\gamma}_l$, in which $\mathbf{K}_l = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T} \sum_{t=1}^{T} \mathbf{B}_l(\mathbf{X}_{it}^{(l)}) \mathbf{B}_l^{\mathrm{T}}(\mathbf{X}_{it}^{(l)})$, and $\mathbf{B}_l(\cdot) = (\mathbf{B}_{l1}(\cdot), \ldots, \mathbf{B}_{lJ_n}(\cdot))^{\mathrm{T}}$. The empirical norm of the spline function s_l is

$$\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T} \sum_{t=1}^{T} s_l^2(\mathbf{X}_{it}^{(l)}) \right\}^{1/2} = \|s_l\|_n.$$

The advantage of choosing the penalization using $\|s_l\|_n$ is that it no longer relies on a particular choice of spline bases. This type of penalization ensures that the coefficients within the same nonparametric component are treated as an entire group in model selection and, therefore, it achieves the same effect as the group-wise model selection approach [34].

The penalty function $p_{\lambda_n}(\cdot)$ can be the L_1-penalty with $p_{\lambda_n}(|\cdot|) = \lambda_n |\cdot|$ which provides a LASSO estimator, or the L_2 penalty $p_{\lambda_n}(|\cdot|) = \lambda_n |\cdot|^2$ which produces a ridge-type estimator. However, we do not apply the L_0 penalty here as it is highly computationally intensive and unstable. The smoothly clipped absolute deviation (SCAD) [7] penalty is considered here, where the derivative is defined as

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

here the constant a is chosen to be 3.7 as in [7], and $\lambda_n > 0$ is a tuning parameter, whose selection is described in Section 3.3. The SCAD penalty has several advantages such as unbiasedness, sparsity and continuity.

The penalized estimator $\widehat{\boldsymbol{\gamma}}^{\mathrm{PQIF}}$ is obtained by minimizing the penalized objective function in (3.1). Then for any $\mathbf{x} \in [0, 1]^{d_x}$, the estimator of the nonparametric functions in (2.1) is calculated by

$$\widehat{\alpha}_l^{\mathrm{PQIF}}(x^{(l)}) = \sum_{j=1}^{J_n} \widehat{\gamma}_{lj}^{\mathrm{PQIF}} B_{lj}(x^{(l)}), \qquad l = 1, \dots, d_x.$$

We establish the asymptotic properties of the penalized parametric and nonparametric components estimators for the marginal GAPLM in the following theorems. We assume that in the true model only the first s_z $(0 \le s_z \le d_z)$ linear components and the first s_x $(0 \le s_x \le d_x)$ nonlinear components are nonzero, and the remaining components are all zeros. Let $\alpha_0(\mathbf{x}_{it}) = \sum_{l=1}^{d_x} \alpha_{0,l}(x_{it}^{(l)}) = \sum_{l=1}^{s_x} \alpha_{0,l}(x_{it}^{(l)}) + \sum_{l=s_x+1}^{d_x} \alpha_{0,l}(x_{it}^{(l)})$, with $\alpha_{0,l} = 0$ almost surely for $l = s_x + 1, \dots, d_x$, where s_x is the number of nonzero nonlinear components. Similarly, let s_z be the number of nonzero components of $\boldsymbol{\beta}_0$. Let $\boldsymbol{\beta}_0 = (\beta_{0,1}, \dots, \beta_{0,d_z})^{\mathrm{T}} = (\boldsymbol{\beta}_{\mathcal{S}0}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathcal{N}0}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_{\mathcal{S}0}$ consists of all s_z nonzero components of $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_{\mathcal{N}0} = \mathbf{0}$ without loss of generality. In a similar fashion to $\boldsymbol{\beta}_0$, denote $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}} = \{(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PQIF}})^{\mathrm{T}}, (\widehat{\boldsymbol{\beta}}_{\mathcal{N}}^{\mathrm{PQIF}})^{\mathrm{T}}\}^{\mathrm{T}}$.

We first derive the convergence rate of the penalized QIF estimators $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}$ and $\{\widehat{\alpha}_l^{\mathrm{PQIF}}\}_{l=1}^{d_x}$. In particular, if d_x is finite, we show that this convergence rate is the same as the rate of convergence for the unpenalized estimators $\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}}$ and $\{\widehat{\alpha}_l^{\mathrm{QIF}}\}_{l=1}^{d_x}$ in Theorem 3. Furthermore, we prove that the penalized estimators $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}$, $\{\widehat{\alpha}^{\mathrm{PQIF}}\}_{l=1}^{d_x}$ possess the sparsity property as in Theorem 4. That is, $\widehat{\alpha}_l^{\mathrm{PQIF}} = 0$ almost surely for $l = s_x + 1, \dots, d_x$, and $\widehat{\boldsymbol{\beta}}_{\mathcal{N}}^{\mathrm{QIF}} = 0$. The sparsity property implies that the model selection procedure is consistent, that is, the selected model converges to the corrected model with probability tending to one. We define

$$\begin{aligned} a_n &= \max_{1 \le l \le d_z} \{|p'_{\lambda_{2,l}}(|\beta_{0,l}|)|, \beta_{0,l} \neq 0\}, \\ b_n &= \max_{1 \le l \le d_z} \{|p''_{\lambda_{2,l}}(|\beta_{0,l}|)|, \beta_{l0} \neq 0\}. \end{aligned} \tag{3.1}$$

THEOREM 3. *Under conditions* (C1)–(C9) *and* (P2) *in Appendix* A.2, *if* $d_x/\log(n) \to 0$, $n^{-1/4} d_z \to 0$, $J_n \asymp n^b$, *for some* $1/(4r) \le b < 1/4$ *with smoothing parameter* $r > 1$ *defined in condition* (C1), *and the tuning parameters* $\lambda_{jn} \to 0$, $j = 1, 2$, $n \to \infty$, *then there exists a local solution* $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}$ *in* (3.1) *such that its rate of convergence is* $O_P\{\rho_n^{-3} d_n^{1/2}(n^{-1/2} + a_n)\}$, *and there exists a local minimizer of* (3.1) *such that*

$$\frac{1}{n} \sum_{l=1}^{d_x} \sum_{i=1}^{n} \sum_{t=1}^{T} \{\widehat{\alpha}_l^{\mathrm{PQIF}}(x_{it}^{(l)}) - \alpha_{0,l}(x_{it}^{(l)})\}^2 = O_P\{\rho_n^{-6} d_n (n^{-1/2} + a_n)^2\}.$$

REMARK 2. If the number of nonparametric functions, d_x, is finite, then ρ_n is a fixed constant. Further if $J_n \asymp n^{1/(2r+1)}$ and $a_n = O(n^{-1/2})$, we obtain the optimal nonparametric convergence rate $n^{-2r/(2r+1)}$ as in [33]. For the parametric terms, if d_x is finite, and $n^{1/(4r)} \ll J_n \ll d_z \ll n^{1/4}$, then we obtain the same parametric convergence rate as in [5].

THEOREM 4. *Assume that conditions* (C1)–(C9), (P1)–(P2) *in Appendix* A.2 *hold. If* $d_x/\log(n) \to 0$, $n^{-1/5} d_z \to 0$, $J_n \asymp n^b$, *for some* $1/(4r) \le b < 1/5$ *with smoothing parameter* $r > 1$ *defined in condition* (C1), *and the tuning parameters* $\lambda_{jn} \to 0$, *and* $\rho_n^{-1} d_n^{-1/2} n^{1/2} \lambda_{jn} \to \infty$, $j = 1, 2$, $n \to \infty$, *then with probability approaching* 1, $\widehat{\alpha}_l = 0$ *almost surely for* $l = s_x + 1, \dots, d_x$, *and the estimator* $\widehat{\boldsymbol{\beta}}^{\mathrm{PQIF}}$ *has the sparsity property, that is,* $P(\widehat{\boldsymbol{\beta}}_{\mathcal{N}}^{\mathrm{PQIF}} = 0) \to 1$ *as* $n \to \infty$.

Theorem 4 indicates that the proposed selection method possesses model selection consistency. Theorems 3 and 4 provide similar results for the nonparametric components as those for the penalized generalized additive models in [33] when d_x is finite. However, the theoretical proof is very different from the penalized generalized additive model approach and is much more challenging, due to the involvement of both parametric and nonparametric components, where two sets of covariates could be dependent, and the dimensions of linear and nonlinear terms increase along with the sample size.

We also investigate the asymptotic distribution of the estimators for the parametric term. Define a vector $\boldsymbol{\kappa}_{\mathcal{S}} = \{p'_{\lambda_{2,n}}(|\beta_{0,1}|)\operatorname{sgn}(\beta_{0,1}), \dots, p'_{\lambda_{2,n}}(|\beta_{0,s_z}|) \times \operatorname{sgn}(\beta_{0,s_z})\}^{\mathrm{T}}$ and a diagonal matrix $\boldsymbol{\Lambda}_{\mathcal{S}} = \operatorname{diag}\{p''_{\lambda_{2,n}}(|\beta_{0,1}|), \dots, p''_{\lambda_{2n}}(|\beta_{0,s_z}|)\}$. In a similar fashion to $\boldsymbol{\beta}$, we write the collections of all components, $\mathbf{X}_i = (\mathbf{X}_{\mathcal{S}i}^{\mathrm{T}}, \mathbf{X}_{\mathcal{N}i}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{Z}_i = (\mathbf{Z}_{\mathcal{S}i}^{\mathrm{T}}, \mathbf{Z}_{\mathcal{N}i}^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\mathbf{Z}}_i = (\widehat{\mathbf{Z}}_{\mathcal{S}i}^{\mathrm{T}}, \widehat{\mathbf{Z}}_{\mathcal{N}i}^{\mathrm{T}})^{\mathrm{T}}$. Further denote $\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}} = \{(\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(1)})^{\mathrm{T}}, \dots, (\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(K)})^{\mathrm{T}}\}^{\mathrm{T}}$, where $\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{D}_i^{\mathrm{T}} \boldsymbol{\Gamma}_{0,i}^{(k)} \widehat{\mathbf{Z}}_{\mathcal{S}i}$. Next, let $\widehat{\boldsymbol{\Psi}}_{\mathcal{S},n} = \widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}} (\mathbf{C}_n^0)^{-1} \widehat{\mathbf{J}}_{\mathrm{DZ_S}}$, $\widehat{\boldsymbol{\Omega}}_{\mathcal{S},n} = \frac{1}{n}\sum_{i=1}^{n} \{\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}} (\mathbf{C}_n^0)^{-1} \mathbf{W}_i\}^{\otimes 2}$ with $\mathbf{W}_i$ in (2.8).

THEOREM 5. *Assume conditions* (C1)–(C9), (P1)–(P2) *in Appendix* A.2 *hold. If* $d_x/\log(n) \to 0$, $n^{-1/5} d_z \to 0$, $J_n \asymp n^b$, *for some* $1/(2r+1) \le b < 1/5$ *with smoothing parameter* $r > 2$ *in condition* (C1), *and the tuning parameters* $\lambda_{jn} \to 0$, $\rho_n^{-1} d_n^{-1/2} n^{1/2} \lambda_{jn} \to +\infty$, $j = 1, 2$, *as* $n \to \infty$, *then*

$$\sqrt{n}\mathbf{A}_n \boldsymbol{\Sigma}_{\mathcal{S},n}^{-1/2} (\widehat{\boldsymbol{\Psi}}_{\mathcal{S},n} + \boldsymbol{\Sigma}_{\mathcal{S}}) \{ (\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \boldsymbol{\beta}_{\mathcal{S}0}) + (\widehat{\boldsymbol{\Psi}}_{\mathcal{S},n} + \boldsymbol{\Lambda}_{\mathcal{S}})^{-1} \boldsymbol{\kappa}_{\mathcal{S}} \} \xrightarrow{D} N(0, \boldsymbol{\Sigma}_A),$$

where $\mathbf{A}_n$ *is any* $q \times d_z$ *matrix with a finite* q *such that* $\boldsymbol{\Sigma}_A = \lim_{n\to\infty} \mathbf{A}_n^{\otimes 2}$, *and* $\boldsymbol{\Sigma}_{\mathcal{S},n} = \widehat{\boldsymbol{\Psi}}_{\mathcal{S},n}^{-1} \widehat{\boldsymbol{\Omega}}_{\mathcal{S},n} \widehat{\boldsymbol{\Psi}}_{\mathcal{S},n}^{-1}$.

3.2. *An algorithm.* To minimize the PQIF in (3.1), we develop an algorithm based on the local quadratic approximation [7]. To obtain an initial estimator $(\boldsymbol{\beta}^0, \boldsymbol{\gamma}^0)$ which is sufficiently close to the true minimizer of (3.1), we could choose the unpenalized QIF estimator $\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} = \{(\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}})^{\mathrm{T}}, (\widehat{\boldsymbol{\gamma}}^{\mathrm{QIF}})^{\mathrm{T}}\}^{\mathrm{T}}$ as the initial value. Let $\boldsymbol{\beta}^k = (\beta_1^k, \dots, \beta_{d_z}^k)^{\mathrm{T}}$ and $\boldsymbol{\gamma}^k = (\boldsymbol{\gamma}_1^{k\mathrm{T}}, \dots, \boldsymbol{\gamma}_{d_x}^{k\mathrm{T}})^{\mathrm{T}}$ be the values at the kth iteration. If β_l^k (or $\boldsymbol{\gamma}_{l'}^k$) is close to zero, such that $|\beta_l^k| \le \epsilon$ (or $\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}} \le \epsilon$) with some small threshold value ϵ, then β_l^{k+1} (or $\boldsymbol{\gamma}_{l'}^{k+1}$) is set to $\mathbf{0}$. We consider $\epsilon = 10^{-6}$ in our numerical examples.

Suppose $\beta_l^{k+1} = 0$, for $l = b_k + 1, \dots, d_z$, and $\boldsymbol{\gamma}_l^{k+1} = 0$, for $l = b_k' + 1, \dots, d_x$, and $\boldsymbol{\theta}^{k+1} = (\beta_1^{k+1}, \dots, \beta_{b_k}^{k+1}, \beta_{b_k+1}^{k+1}, \dots, \beta_{dz}^{k+1}, (\boldsymbol{\gamma}_1^{k+1})^{\mathrm{T}}, \dots, (\boldsymbol{\gamma}_{b_k'}^{k+1})^{\mathrm{T}}, (\boldsymbol{\gamma}_{b_k'+1}^{k+1})^{\mathrm{T}}, \dots, (\boldsymbol{\gamma}_{d_x}^{k+1})^{\mathrm{T}})^{\mathrm{T}} = \{(\boldsymbol{\beta}_{\mathcal{S}}^{k+1})^{\mathrm{T}}, (\boldsymbol{\beta}_{\mathcal{N}}^{k+1})^{\mathrm{T}}, (\boldsymbol{\gamma}_{\mathcal{S}}^{k+1})^{\mathrm{T}}, (\boldsymbol{\gamma}_{\mathcal{N}}^{k+1})^{\mathrm{T}}\}^{\mathrm{T}}$, in which $\boldsymbol{\beta}_{\mathcal{N}}^{k+1} = \mathbf{0}$, $\boldsymbol{\gamma}_{\mathcal{N}}^{k+1} = \mathbf{0}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\beta}_{\mathcal{N}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathcal{N}}^{\mathrm{T}})^{\mathrm{T}}$ be the partition of any $\boldsymbol{\theta}$.

The local quadratic approximation is implemented for obtaining the nonzero components $\boldsymbol{\theta}_{\mathcal{S}}^{k+1} = \{(\boldsymbol{\beta}_{\mathcal{S}}^{k+1})^{\mathrm{T}}, (\boldsymbol{\gamma}_{\mathcal{S}}^{k+1})^{\mathrm{T}}\}^{\mathrm{T}}$. Specifically, for $|\beta_l^k| > \epsilon$, the penalty for the parametric term is approximated by

$$\begin{aligned} p_{\lambda_n}(|\beta_l|) &\approx p_{\lambda_n}(|\beta_l^k|) + p'_{\lambda_n}(|\beta_l^k|)(|\beta_l| - |\beta_l^k|) \\ &\approx p_{\lambda_n}(|\beta_l^k|) + \tfrac{1}{2} p'_{\lambda_n}(|\beta_l^k|)|\beta_l^k|^{-1}\{\beta_l^2 - (\beta_l^k)^2\}. \end{aligned}$$

For $\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}} > \epsilon$, the penalty function for the nonparametric part is approximated by

$$\begin{aligned} & p_{\lambda_n}(\|\boldsymbol{\gamma}_{l'}\|_{\mathbf{K}_{l'}}) \\ &\quad \approx p_{\lambda_n}(\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}}) + p'_{\lambda_n}(\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}})\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}}^{-1} \boldsymbol{\gamma}_{l'}^{kT}\mathbf{K}_{l'}(\boldsymbol{\gamma}_{l'} - \boldsymbol{\gamma}_{l'}^k) \\ &\quad \approx p_{\lambda_n}(\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}}) + \tfrac{1}{2} p'_{\lambda_n}(\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}})\|\boldsymbol{\gamma}_{l'}^k\|_{\mathbf{K}_{l'}}^{-1} (\boldsymbol{\gamma}_{l'}^{\mathrm{T}}\mathbf{K}_{l'}\boldsymbol{\gamma}_{l'} - \boldsymbol{\gamma}_{l'}^{kT}\mathbf{K}_{l'}\boldsymbol{\gamma}_{l'}^k), \end{aligned}$$

where p'_{λ_n} is the first-order derivative of p_{λ_n}.

This leads to the local approximation of the objective function in (3.1) by a quadratic function:

$$\begin{aligned} & Q_n(\boldsymbol{\theta}^k) + \dot{Q}_n(\boldsymbol{\theta}^k)^{\mathrm{T}} \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^k \\ \boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^k \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^k \\ \boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^k \end{pmatrix}^{\mathrm{T}} \ddot{Q}_n(\boldsymbol{\theta}^k) \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^k \\ \boldsymbol{\gamma}_{\mathcal{S}} - \boldsymbol{\gamma}_{\mathcal{S}}^k \end{pmatrix} \\ &\quad + \frac{1}{2} n \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{S}} \\ \boldsymbol{\gamma}_{\mathcal{S}} \end{pmatrix}^{\mathrm{T}} \Lambda(\boldsymbol{\theta}^k) \begin{pmatrix} \boldsymbol{\beta}_{\mathcal{S}} \\ \boldsymbol{\gamma}_{\mathcal{S}} \end{pmatrix}, \end{aligned}$$

where $\dot{Q}_n(\boldsymbol{\theta}^k) = \frac{\partial Q_n(\boldsymbol{\theta}^k)}{\partial \boldsymbol{\theta}_{\mathcal{S}}}$, $\ddot{Q}_n(\boldsymbol{\beta}^k) = \frac{\partial^2 Q_n(\boldsymbol{\theta}^k)}{\partial \boldsymbol{\theta}_{\mathcal{S}}\, \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}}$ with $\boldsymbol{\theta}_{\mathcal{S}} = (\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathcal{S}}^{\mathrm{T}})^{\mathrm{T}}$, and

$$\begin{aligned} \Lambda(\boldsymbol{\theta}^k) = \operatorname{diag}\{ & |\beta_1^k|^{-1} p'_{\lambda_n}(|\beta_1^k|), \dots, |\beta_{b_k}^k|^{-1} p'_{\lambda_n}(|\beta_{b_k}^k|), \\ & \|\boldsymbol{\gamma}_1^k\|_{\mathbf{K}1}^{-1} p'_{\lambda_n}(\|\boldsymbol{\gamma}_1^k\|_{\mathbf{K}_1})\mathbf{K}_1, \dots, \|\boldsymbol{\gamma}_{b_k'}^k\|_{\mathbf{K}_{b_k'}}^{-1} p'_{\lambda_n}(\|\boldsymbol{\gamma}_{b_k'}^k\|_{\mathbf{K}_{b_k'}})\mathbf{K}_{b_k'}\}. \end{aligned} \tag{3.2}$$

We minimize the above quadratic function to get $\boldsymbol{\theta}_S^{k+1}$. The corresponding Newton–Raphson algorithm provides

$$\boldsymbol{\theta}_S^{k+1} = \boldsymbol{\theta}_S^k - \{\ddot{Q}_n(\boldsymbol{\theta}^k) + n\Lambda(\boldsymbol{\theta}^k)\}^{-1}\{\dot{Q}_n(\boldsymbol{\theta}^k) + n\Lambda(\boldsymbol{\theta}^k)\boldsymbol{\theta}_S^k\}.$$

The above iteration process is repeated until convergence is reached, where the convergence criterion is based on $\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\| \leq 10^{-6}$. The proposed algorithm is quite stable and converges quickly. However, in general, the computational time increases as the dimension of covariates increases.

3.3. *Tuning parameter and knots selection.* Tuning parameter and knots selections play important roles in the performance of model selection. The spline approximation for the nonparametric components requires an appropriate selection of the knot sequences $\{\upsilon_l\}_{l=1}^{d_x}$ in Section 2.2. For the penalized QIF method in Section 3.1, in addition to knots selection, we also need to address how to choose tuning parameters $\lambda_{1,n}$ and $\lambda_{2,n}$ in the SCAD penalty function. To reduce computational complexity, we consider $\lambda_{1,n} = \lambda_{2,n} = \lambda_n$ and select only λ_n. This is justified by Theorems 3, 4 and 5 in Section 3.

Although selecting the number and position of spline knots is important in curve smoothing, in our simulation study we found that knot selection seems to be less critical for the estimation of the parametric coefficients and model selection than for the estimation of the nonparametric components. For convenience, we choose equally spaced knots and the number of interior knots is selected as the integer part of $N_n = n^{1/(2p+3)}$, where n is the sample size and p is the order of the polynomial spline. This approach is also adopted in [16, 33] and [32]. Furthermore, we use the same knot sequences selected in the unpenalized procedure for the penalized QIF estimation. Therefore, we only need to determine the tuning parameter for the penalization part. For any given tuning parameter λ_n, the estimator minimizing (3.1) is denoted as $\widehat{\boldsymbol{\theta}}_{\lambda_n}$. We propose to use the extended Bayesian Information Criterion (EBIC) to select the optimal tuning parameters based on [3] and [13]. Because the QIF Q_n is analog to minus twice the log-likelihood function [23], we define the EBIC in the PQIF procedure as

$$\begin{aligned}\mathrm{EBIC}(\lambda_n) = {} & Q_n(\widehat{\boldsymbol{\theta}}_{\lambda_n}) + \log(n)\hat{d}_z(\lambda_n) + \log(\nu_z(\lambda_n)) \\ & + \log(n)N_n\hat{d}_x(\lambda_n) + N_n\log(\nu_x(\lambda_n)),\end{aligned} \tag{3.3}$$

where $\hat{d}_z(\lambda_n)$ and $\hat{d}_x(\lambda_n)$ are the nonzero parametric and nonparametric terms in $\widehat{\boldsymbol{\theta}}_{\lambda_n}$, respectively, and $\nu_z(\lambda_n) = \binom{d_z}{\hat{d}_z(\lambda_n)}$, which is a combination operator and represents the number of choices for selecting $\hat{d}_z(\lambda_n)$ terms out of d_z parametric terms. Similarly, define $\nu_x(\lambda_n) = \binom{d_x}{\hat{d}_x(\lambda_n)}$. See [3] for details. However, when the full likelihood is available, it is more accurate to use minus twice the log-likelihood function instead of Q_n as the first term in (3.3). That is,

$$\begin{aligned}\mathrm{EBIC}(\lambda_n) &= -2\log L(\widehat{\boldsymbol{\theta}}_{\lambda_n}) + \log(n)\hat{d}_z(\lambda_n) + \log(\nu_z(\lambda_n)) \\ &\quad + \log(n)N_n\hat{d}_x(\lambda_n) + N_n\log(\nu_x(\lambda_n)),\end{aligned}$$

where $L(\cdot)$ is the full likelihood function. As indicated in [26], the one using the full likelihood, if it is available, has better finite sample performance when the sample size is small. The optimal λ_n is chosen such that the EBIC value reaches the minimum, or equivalently, $\widehat{\lambda}_n = \arg\min_{\lambda_n} \mathrm{EBIC}(\lambda_n)$.

4. Simulation studies.

In this section, we assess the numerical performance of the proposed methods through simulation studies. To assess estimation accuracy and efficiency, define the model error (ME) as

$$\frac{1}{n^*T}\sum_{i=1}^{n^*}\sum_{t=1}^{T}\left\{g^{-1}\left(\sum_{l=1}^{d_x}\widehat{\alpha}_l(x_{it}^{(l)}) + \mathbf{z}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}\right) - g^{-1}\left(\sum_{l=1}^{d_x}\alpha_l(x_{it}^{(l)}) + \mathbf{z}_{it}^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}^2,$$

where $(\mathbf{x}_{it}, \mathbf{z}_{it})_{i=1,t=1}^{n^*,T}$ are independently generated test data and follow the same distribution as the training data. In our simulations, we take $n^* = 1000$. Furthermore, g^{-1} is the identity link function for continuous outcomes and the logit link function for binary outcomes. The model error measures the prediction performance of different methods. Denote the index sets of the selected and true models by $\hat{\mathcal{S}}$ and $\mathcal{S}_0$, respectively. If $\hat{\mathcal{S}} = \mathcal{S}_0$, then $\hat{\mathcal{S}}$ is a correct selection; if $\mathcal{S}_0 \subset \hat{\mathcal{S}}$ and $\mathcal{S}_0 \neq \hat{\mathcal{S}}$, then we call $\hat{\mathcal{S}}$ over selection; otherwise, if $\mathcal{S}_0 \not\subset \hat{\mathcal{S}}$, then $\hat{\mathcal{S}}$ under selection. The number of replications is 500 in the following simulation studies.

4.1. *Example* 1: *Continuous response.*

The continuous responses $\{Y_{it}\}$ are generated from

$$Y_{it} = \sum_{l=1}^{d_x}\alpha_l(X_{it}^{(l)}) + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_{it}, \qquad i = 1, \ldots, n, t = 1, \ldots, 5, \tag{4.1}$$

where $n = 100, 200$, or 500, and $d_x = d_z = 2n^{1/4}$ which is rounded to the nearest integer and takes values of 6, 8 and 10, respectively, for $n = 100, 200$ and 500. We take $\alpha_1(x) = \sin(2\pi x)$, $\alpha_2(x) = 8x(1-x) - 4/3$, and $\alpha_l(x) = 0$ for $l = 3, \ldots, d_x$, and $\beta_1 = 1$, $\beta_2 = 2$, and $\beta_l = 0$ for $l = 3, \ldots, d_z$. Therefore, only the first two variables in $\mathbf{X}_{it}$ and $\mathbf{Z}_{it}$ are relevant and the rest are null variables. The covariates $\mathbf{X}_{it} = (X_{it}^{(1)}, \ldots, X_{it}^{(d_x)})^{\mathrm{T}}$ are generated by $X_{it}^{(l)} = (2W_{it}^{(l)} + U_{it})/3$, where $\mathbf{W}_{it} = (W_{it}^{(1)}, \ldots, W_{it}^{(d_x)})$ and U_{it} are independently generated from Uniform$([0, 1]^{d_x})$ and Uniform([0, 1]), respectively. Therefore, the covariates $\mathbf{X}_{it}$ have an exchangeable correlation structure. In addition, $\mathbf{Z}_{it} = (Z_{it}^{(1)}, \ldots, Z_{it}^{(d_z)})^{\mathrm{T}}$ are generated with $Z_{it}^{(1)} = 1$ and $(Z_{it}^{(2)}, \ldots, Z_{it}^{(d_z)})$ being generated from a zero mean multivariate normal distribution with a marginal variance of 1 and an AR-1 correlation with parameter $\rho = 0.7$. The errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i5})^{\mathrm{T}}$ follows a zero mean multivariate

TABLE 1

Example 1: *The simulation results using the SCAD penalty with exchangeable* (*EC*), *AR*-1 *or independent* (*IND*) *working correlation and linear or cubic splines*. *The columns of C*, *O and U provide the percentage of correct selection*, *over selection and under selection*, *and MME provides the averaged model errors from* 500 *replications*

		Linear spline				**Cubic spline**			
Method	***n***	**C**	**O**	**U**	**MME**	**C**	**O**	**U**	**MME**
EC	100	0.936	0.050	0.014	0.0461	0.842	0.006	0.098	0.1337
	200	0.992	0.008	0.000	0.0258	1.000	0.000	0.000	0.0175
	500	1.000	0.000	0.000	0.0089	1.000	0.000	0.000	0.0054
AR-1	100	0.930	0.012	0.058	0.0660	0.766	0.012	0.222	0.2312
	200	0.994	0.00	0.006	0.0268	1.000	0.000	0.000	0.0206
	500	1.000	0.000	0.000	0.0097	1.000	0.000	0.000	0.0062
IND	100	0.836	0.008	0.156	0.1046	0.648	0.002	0.350	0.2707
	200	0.986	0.012	0.002	0.0322	0.994	0.006	0.000	0.0259
	500	1.000	0.000	0.000	0.0128	1.000	0.000	0.000	0.0091

normal with a marginal variance of $\sigma^2 = 1.5$ and an exchangeable correlation with correlation $\rho = 0.7$.

The proposed penalized QIF method with the SCAD penalty is considered. In spline approximation, we use both the linear splines and cubic splines. Furthermore, we consider basis matrices from three different working correlation structures: exchangeable (EC), AR-1 and independent (IND), and compare their estimation efficiencies to illustrate the effect on efficiency gain of incorporating within-cluster correlation.

Table 1 presents the variable selection and estimation results. It summarizes the percentages of correct selection (C), over selection (O) and under selection (U). It also gives the mean model errors (MME) from 500 replications. Table 1 indicates that the probability of recovering the correct model increases to 1 quickly and the MME decreases as the sample size increases. This confirms the consistency theorems of variable selection and estimation provided in Section 3.1. It also shows that the procedures with a correct EC working correlation always have the smallest MMEs and, therefore, the estimators are more efficient than their counterparts with IND structure, which ignore within-cluster correlation. The method with a misspecified AR-1 correlation is less efficient than the one using the true EC structure, but is still more efficient than assuming independent structure. Furthermore, it also shows that the percentage of correct model-fitting using EC structure is higher than the one using IND when the sample size is small ($n = 100$).

4.2. *Example* 2: *Continuous response with randomly generated correlation structure*. To assess our method in a more challenging scenario, we consider a model similar to (4.1), but with randomly generated correlation structures. In particular, we assume that the dimensions of $\mathbf{X}$ and $\mathbf{Z}$ are $d_x = 9$, $d_z = 5$, respectively. As in (4.1), only $X_{it}^{(1)}$, $X_{it}^{(2)}$, $Z_{it}^{(1)}$ and $Z_{it}^{(2)}$ are relevant and take the same forms as in Example 1. Furthermore, we consider the number of clusters $n = 25$ or 250, and cluster size 3. The set-up of $n = 25$ mimics the real data analyzed in Section 5. The errors $\{\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{i3})^{\mathrm{T}}\}_{i=1}^{25}$ independently follow a multivariate normal distribution as in Example 1, but with a randomly generated correlation matrix $\boldsymbol{\Gamma}_r$ for each replication r. Let $\boldsymbol{\Sigma}_1$ be a matrix with diagonals being 1 and all the off-diagonals with value 0.5, and $\boldsymbol{\Sigma}_{r2} = \mathbf{Q}_r \boldsymbol{\Lambda}_r \mathbf{Q}_r^T$ with $\mathbf{Q}_r$ being a randomly generated orthogonal matrix and $\boldsymbol{\Lambda}_r = \operatorname{diag}(\lambda_{r1}, \lambda_{r2}, \lambda_{r3})$ with $\{\lambda_{rj}\}_{j=1}^{3}$ being randomly generated from Uniform[0.2, 2]. Let $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_{r2}$ and $\sigma_{r1}, \dots, \sigma_{r3}$ be the diagonal elements of $\boldsymbol{\Sigma}_r$. Let $\boldsymbol{\Delta}_r = \operatorname{diag}\{\sigma_{r1}^{-1/2}, \dots, \sigma_{r3}^{-1/2}\}$. Then the randomly generated correlation structure for the rth replication is $\boldsymbol{\Gamma}_r = \boldsymbol{\Delta}_r \boldsymbol{\Sigma}_r \boldsymbol{\Delta}_r$. We use this example to investigate the performance of the QIF method in approximating the randomly generated correlation structures.

We estimate the model using the proposed penalized QIF method with linear spline and SCAD penalty, and assume IND, EC or AR-1 working correlation structure. We also consider linear spline QIF estimations of a full model (FULL) and an oracle model (ORACLE), where the full model contains all 14 variables while the oracle one has only the four nonzero variables. The oracle model is not available in real data analysis where the underlying data-generating process is unknown.

Table 2 summarizes variable selection performance on correct, over and under selection percentages of the SCAD approach with IND, EC and AR-1 working correlations and reports the mean model error (MME) for FULL, ORACLE and SCAD when the sample size $n = 25$ and 250, respectively. Table 2 clearly indicates that, for a randomly generated correlation, SCAD with an EC working correlation still performs better than the one with IND working structure. Furthermore, when

TABLE 2

Example 2: *Continuous response with randomly generated correlation. The percentage of correct selection* (*C*), *over selection* (*O*) *and under selection* (*U*) *are provided using linear spline with the SCAD penalty for three working correlation*: *exchangeable* (*EC*), *AR*-1 *or independent* (*IND*). *The columns of SCAD*, *ORACLE and FULL report the mean model error* (*MME*) *of the SCAD approach and a standard linear spline estimation of the oracle model* (*ORACLE*), *and the full model* (*FULL*) *from* 500 *replications*

n	**Method**	**C**	**O**	**U**	**SCAD**	**ORACLE**	**FULL**
250	EC	0.998	0.002	0.000	0.0242	0.0223	0.0656
	AR1	0.990	0.002	0.008	0.0245	0.0233	0.0704
	IND	0.986	0.006	0.008	0.0256	0.0250	0.0713
25	EC	0.616	0.194	0.190	0.5081	0.3858	1.6886
	AR1	0.566	0.212	0.222	0.5281	0.3723	1.7528
	IND	0.536	0.256	0.208	0.5546	0.3518	0.7729

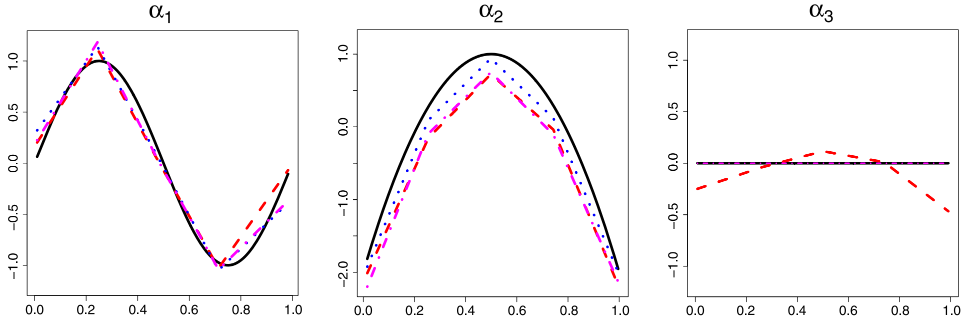

FIG. 1. *Example* 2: *Plots of the first three estimated functions from SCAD* (*dot–dash*), *Oracle* (*dotted*) *and Full* (*dashed*) *approaches with the true functions* (*line*). *For* α_3, *both SCAD and Oracle give exactly zero estimates. The cluster size is* $n = 250$.

the sample size is large ($n = 250$), the estimation using EC always yield a smaller MME than the one with IND working structure. It indicates that although EC is a misspecified correlation structure, it can still improve estimation and inference performances by incorporating some correlation in the data into the estimation. When the sample size is small ($n = 25$), the estimation using EC or AR1 working correlations of FULL and ORACLE is worse due to the extra noise in modeling within-cluster correlation. However, the SCAD with EC or AR1 working correlations still give smaller MMEs than SCAD with IND correlation, due to their better performances in recovering the correct model. Finally, Table 2 also shows that the penalized procedure dramatically improves estimation accuracy compared to the un-penalized approach, with MMEs from the SCAD being very close to the MMEs from the ORACLE model, and much smaller than the FULL model.

From one selected data set, Figure 1 plots the first three estimated functional components from the SCAD, FULL and ORACLE models using linear spline and exchangeable working correlation for cluster size $n = 250$. Note that for the third variable, both the true and estimated functions from SCAD are zero. It shows that the proposed SCAD method estimates unknown functions reasonably well.

4.3. *Example* 3: *Binary response.* A random sample of 250 clusters is generated in each simulation run. Within each cluster, binary responses $\{Y_{it}\}_{t=1}^{20}$ are generated from a marginal logit model

$$\operatorname{logit} P(Y_{it} = 1 | \mathbf{X}_{it} = \mathbf{x}_{it}, \mathbf{Z}_{it} = \mathbf{z}_{it}) = \sum_{l=1}^{5} \alpha_l(x_{it}^{(l)}) + \mathbf{z}_{it}^{\mathrm{T}} \boldsymbol{\beta},$$

where $\alpha_1(x) = \cos(2\pi x)/4$, $\alpha_l(x) = 0$, for $l = 2, \dots, 5$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{10})^{\mathrm{T}}$ with $\beta_1 = 1$ and $\beta_l = 0$ for $l = 2, \dots, 10$. The covariates $\mathbf{X}_{it} = \{X_{it}^{(l)}\}_{l=1}^{5}$ and $\mathbf{Z}_{it} = \{Z_{it}^{(l)}\}_{l=1}^{10}$ are generated in the same way as in Example 1. The covariates $\mathbf{X}_{it}$ have an exchangeable correlation structure, and $\mathbf{Z}_{it}$ have an AR-1 correlation

TABLE 3

Example 3: *Binary response. The percentages of correct selection* (*C*), *over selection* (*O*) *and under selection* (*U*) *are provided using linear spline with the SCAD penalty for three working correlation*: *exchangeable* (*EC*), *AR*-1 *or independent* (*IND*). *The columns of SCAD*, *ORACLE and FULL provide the mean model error* ($\times 10^3$) *of the SCAD approach and a standard linear spline estimation of the oracle model* (*ORACLE*), *and the full model* (*FULL*). *The number of replications is* 500

	C	**O**	**U**	**SCAD**	**ORACLE**	**FULL**
EC	0.868	0.084	0.048	0.1102	0.0547	1.6820
AR1	0.692	0.012	0.292	0.1264	0.0586	1.8489
IND	0.684	0.006	0.208	0.1484	0.0616	1.9057

structure with $\rho = 0.7$. We use the algorithm described in [21] to generate the correlated binary data. It has an exchangeable correlation structure with a correlation coefficient of 0.3.

We conduct variable selection using the proposed penalization method with linear spline (SCAD). We also consider estimation of the full (FULL) and oracle (ORACLE) models using the unpenalized QIF with linear spline. We minimize (2.4) and (3.1) using AR-1 and independent working structures, in addition to the true exchangeable correlation structure.

Table 3 summarizes the MMEs for the SCAD, ORACLE and FULL with three different working correlations. Table 4 also reports the sample means and sample standard deviations (SD) of the estimators of the nonzero regression coefficient $\widehat{\beta}_1$ from 500 replications. It again shows that estimation based on correctly specified exchangeable correlation structure is the most efficient, having the smallest MMEs and SDs. Estimation with a misspecified AR-1 correlation results in some efficiency loss compared to using the true structure, but it is still much more efficient than assuming independent structure. However, for GEE, estimation using a misspecified AR-1 correlation structure could be less efficient than assuming independence, since the GEE requires the estimation of the correlation ρ for misspecified AR-1, and the estimator of ρ may not be valid.

TABLE 4

Example 3: *Binary response. The sample mean and standard deviation* (*in parenthesis*) *of* $\widehat{\beta}$ *from the SCAD*, *ORACLE and FULL model approaches*

	SCAD	**ORACLE**	**FULL**
EC	1.0258 (0.0461)	1.0115 (0.0436)	1.0945 (0.0598)
AR1	0.9969 (0.0558)	1.0177 (0.0537)	1.0748 (0.0738)
IND	0.9932 (0.0792)	1.0543 (0.0758)	1.0801 (0.0893)

Furthermore, similar to the previous study, MMEs calculated based on the SCAD approach are very close to the ones from ORACLE, and much smaller than the MMEs from the FULL model. The MMEs of the FULL model are close to 4 times the MMEs of SCAD. This shows that the SCAD penalization improves estimation accuracy significantly by effectively removing the redundant variables. Table 3 also gives the frequency of correct, over and under selection for the SCAD approach. Overall, the SCAD procedure works reasonably well, and the SCAD with a correct EC working correlation structure provides noticeably better variable selection results than the SCAD with IND working structure.

5. Real data analysis. In this section, the proposed methods are applied to analyze a pharmacokinetics study for investigating CCI-779 effects on renal cancer patients [1]. CCI-779 is an anticancer agent with demonstrated inhibitory effects on tumor growth. In this study, patients with advanced renal cell carcinoma received CCI-779 treatment weekly until demonstrated evidence of disease progression. One goal of the study is to identify transcripts in peripheral blood mononuclear cells (PBMCs) which are useful for predicting the temporal pharmacogenomic profile of CCI-799, after initiation of CCI-779 therapy. The data consists of expression levels of 12,626 genes from 33 patients on three scheduled visits: baseline, week 8 and week 16. However, not all patients have measurements at all three visits. We have unbalanced data with a total of only 54 observations. To account for the cumulative-dose drug exposure, CCI-779 cumulative AUC was used to quantify the pharmacogenomic measure of CCI-799 for each patient at each visit. The AUC is of popular use in estimating bioavailability of drugs in pharmacology. Since the response variable CCI-779 cumulative AUC is continuous, we consider our model (2.1) with an identity link function.

With a total of 12,626 genes as covariates and only 54 observations, we first apply the nonparametric independence screening method (NIS) described in [6] to reduce the dimensionality to a moderate size. We ranked the genes according to their empirical marginal function norms, and kept only the first 205 genes with marginal function norms larger than the 99th% quantile of the empirical norms of randomly permuted data. After variable screening, we then applied the penalized polynomial splines [13, 32] for high-dimensional additive model selection. We used the linear spline with a LASSO penalty function and selected the tuning parameters with a five-fold cross-validation procedure. This procedure further reduced dimensionality and selected only 14 genes. Out of the selected 14 genes, we then applied our proposed methods for more refined variable selection and estimation.

We first considered a generalized additive model (GAM), which is a special case of a GAPLM model with Z_{it} in (2.1) consisting of an intercept term only. We applied the linear spline QIF method to estimate the function components. The plots of the estimated functions in Figure 2 suggested that the function forms of the five variables (1198_at, 290_s_at, 32463_at, 33344_ at, 34100_at) are almost linear.

TABLE 5

Real data. The mean squared estimation error (MSEE), EBIC values and averaged mean squared prediction error (AMSPE) of different methods

Method	MSEE	EBIC	Model size	AMSPE
GAM	0.0127	0.1902	14	0.4618
GAPLM	0.0162	−0.6275	14	0.3496
GAM-SCAD	0.0132	0.2285	14	0.2949
GAPLM-SCAD	0.0205	−0.8969	11	0.2398
GAPLM-Linear	0.0191	−0.7774	11	0.4069
GLM	0.2801	0.2772	14	0.6760
GLM-LASSO	0.0989	0.9716	31	0.8530

Therefore, we further considered a GAPLM model with these five terms as linear terms, and the rest as additive terms. For both models, we applied our proposed penalized QIF method for more refined variable selection. For the GAPLM, we also considered the variable selection method of [25]. However, it can only select linear terms and keeps all additive terms. We refer to this method as GAPLM-Linear. Finally, as a benchmark, we also considered two linear models; one contains only the 14 genes selected in the high-dimensional additive model and is referred as GLM, the other one begins with 205 genes, and variable selection in this high-dimensional linear model is then conducted using LASSO, which is referred as GLM-LASSO.

For the GAM, we kept all 14 variables, while both GAPLM and GAPLM-Linear selected 11 variables. In Table 5, we report their mean squared estimation errors (MSEE) and EBIC values. With the response being continuous, let $\hat{Y}_{it} = \sum_{l=1}^{d_x} \hat{\alpha}_l(x_{it}^{(l)}) + \mathbf{z}_{it}^t \hat{\boldsymbol{\beta}}$ be the estimator of Y_{it} from any method. Then define $\text{MSEE} = \frac{1}{N_t} \sum_{i=1}^{n} \sum_{t=1}^{T_i} (Y_{it} - \hat{Y}_{it})^2$, with N_t being the total number of observations and T_i being the size of cluster i. Equation (3.4) with a Gaussian likelihood was used to compute the EBIC, since the response variable is continuous and a working independent structure is used here. It is not surprising that the GAM gave the smallest MSEE since it has the most complicated model; while the GAPLM-SCAD gives the most parsimonious model with the smallest EBIC value. This suggests that with a simpler model, one may be able to make more efficient estimation and inference. For the two linear models, their much larger MSEEs suggest that the data contains nonlinear dynamics which cannot be fully incorporated by linear models.

Furthermore, as suggested by one referee, we also compared the above methods by their prediction performances. We randomly selected 28 patients for estimation and left the remaining 5 patients for prediction. We calculated the mean squared prediction errors (MSPE) for each method for 100 replications. Table 5 reports the averaged MSPEs from 100 replications. It shows that the GAPLM-SCAD gives

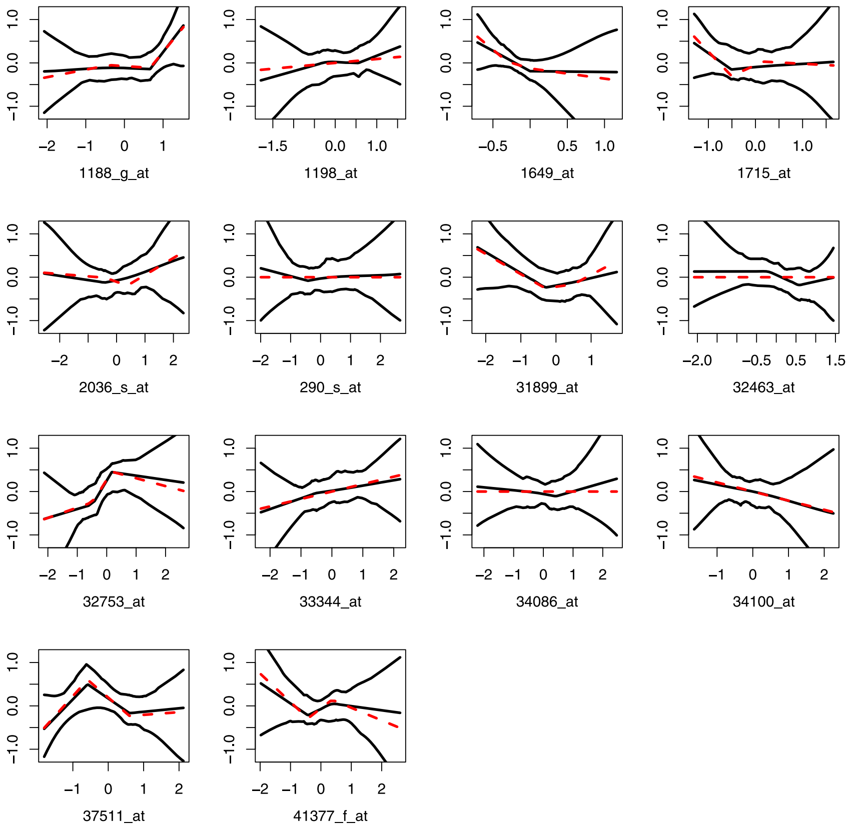

FIG. 2. *Real data: Plots of the estimated components of GAM (line) and GAPLM-SCAD (dashed) with* 95% *point-wise bootstrap confidence intervals from the GAM.*

the smallest prediction error, and all non or semiparametric methods give smaller prediction errors than the linear models. It again suggests that the data contains a nonlinear structure. Those findings are consistent with the ones observed from EBICs. In the above, we have used an independent correlation structure in all procedures. Using other types of correlation structure (e.g., exchangeable, AR-1) in the estimation of GAM, GAPLM and GLM, which are not reported here, always gives larger MSEEs due to the extra noise in modeling within-cluster correlation when the sample size is rather small.

6. Discussion. In this paper, we provide new statistical theory for model selection and estimation with diverging numbers of linear covariates and nonlinear components for generalized partial linear additive models. Our work differs from existing works in three major aspects. First, we consider model selection for both the parametric and nonparametric parts simultaneously, while most of the literature focuses on selection for either the parametric or the nonparametric part. Second, we allow the numbers of linear covariates and nonlinear components to increase with the sample size. Theoretical development for model selection and estimation for diverging number of covariates in nonparametric components is completely different from finite dimension settings. Third, we allow dependence between the covariates in the nonparametric and parametric part, and also dependence between the longitudinal responses. All of these impose significant challenges in developing asymptotic theory and oracle properties.

Note that the growing dimensions of the nonparametric part are smaller than the parametric part, since the nonparametric components involve many more parameters than the parametric part. The order of the parametric dimension is comparable to that in the existing literature for parametric model selection with diverging number of covariates [5, 9, 18]. To establish the asymptotic properties of the QIF estimators, a crucial step is to obtain the upper and lower bounds of the eigenvalues of the matrix $\mathbf{C}_n$ in the QIF equation. These bounds are assumed for the parametric models [8] or can be derived for independent observations [31] using Lemma 1 of [24]. However, neither of these are valid in our setting. Instead, we develop an alternative strategy through proving Lemma S.4, which is essential in establishing bounds for the eigenvalues of a large random matrix. The result in Lemma S.4 can also be used for verifying the second-order KKT condition on demand of bounds of random matrix with diverging dimension.

It is worth noting that the GEE estimator under the generalized partial linear additive model framework is semiparametric efficient under the correct correlation structure [4]. Since the GEE and QIF are asymptotically equivalent when the correlation structure is correctly specified, the proposed QIF estimator for the generalized partial linear additive model is also semiparametric efficient under the correct correlation structure.

APPENDIX: ASSUMPTIONS AND PROOFS

A.1. Notation and definitions. For any functions $s_1, s_2 \in \mathcal{L}_2([0,1])$, define

$$\langle s_1, s_2\rangle = \int_0^1 s_1(x)s_2(x)\,dx \quad\text{and}\quad \|s\|_2^2 = \int_0^1 s^2(x)\,dx. \tag{A.1}$$

Let $\mathcal{H}_0$ be the space of constant functions on $[0,1]$, and let $\mathcal{H}_0^{\perp} = \{s : \langle s, 1\rangle = 0, s \in \mathcal{L}_2\}$ and 1 is the constant function on $[0,1]$. Define the additive model space $\mathcal{M}$ and the space of additive polynomial spline functions $\mathcal{M}_n$ as

$$\mathcal{M} = \left\{ s(\mathbf{x}) = \sum_{l=1}^{d_x} s_l(x_l); s_l \in \mathcal{H}_0^{\perp} \right\}, \qquad \mathcal{M}_n = \left\{ s(\mathbf{x}) = \sum_{l=1}^{d_x} s_l(x_l); s_l \in \varphi_l^{0,n} \right\},$$

where $\varphi_l^{0,n} = \{s_l(\cdot) : s_l \in \varphi_l, \langle s_l, 1\rangle = 0\}$ is the centered polynomial spline space. Let $\mathbf{s}(\mathbf{X}) = (s(\mathbf{X}_1), \ldots, s(\mathbf{X}_T))^{\mathrm{T}}$, for any $s \in \mathcal{M}$ and $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \ldots, \mathbf{X}_T^{\mathrm{T}})^{\mathrm{T}}$. We define the theoretical and empirical norms of $\mathbf{s}$: $\|\mathbf{s}\|^2 = E\{\mathbf{s}^{\mathrm{T}}(\mathbf{X})\mathbf{s}(\mathbf{X})\}$ and $\|\mathbf{s}\|_n^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{s}^{\mathrm{T}}(\mathbf{X}_i)\mathbf{s}(\mathbf{X}_i)$.

For $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$, denote a d_n vector and a $d_n \times d_n$ matrix

$$\mathbf{S}_n(\boldsymbol{\theta}) = \dot{\mathbf{G}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}), \qquad \mathbf{H}_n(\boldsymbol{\theta}) = \dot{\mathbf{G}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\dot{\mathbf{G}}_n(\boldsymbol{\theta}), \tag{A.2}$$

where the $(Kd_n) \times d_n$ matrix

$$\dot{\mathbf{G}}_n(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{G}_n(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{G}_n(\boldsymbol{\theta}), \frac{\partial}{\partial \boldsymbol{\gamma}}\mathbf{G}_n(\boldsymbol{\theta})\right) \equiv \left(\dot{\mathbf{G}}_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \dot{\mathbf{G}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta})\right). \tag{A.3}$$

By [23] and Lemma S.4, the estimating equation for $\boldsymbol{\theta}$ is

$$n^{-1}\dot{Q}_n(\boldsymbol{\theta}) \equiv n^{-1}\frac{\partial}{\partial \boldsymbol{\theta}}Q_n(\boldsymbol{\theta}) = 2\mathbf{S}_n(\boldsymbol{\theta}) + O_P\left(\rho_n^{-1}n^{-1}d_n\right) = \mathbf{0}, \tag{A.4}$$

and the second derivative of $Q_n(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$

$$n^{-1}\ddot{Q}_n(\boldsymbol{\theta}) \equiv n^{-1}\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}}Q_n(\boldsymbol{\theta}) = 2\mathbf{H}_n(\boldsymbol{\theta}) + o_P(1). \tag{A.5}$$

To facilitate technical arguments in the following proofs, we write

$$\mathbf{S}_n(\boldsymbol{\theta}) = \begin{pmatrix} \dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}) \\ \dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}) \end{pmatrix}, \tag{A.6}$$

$$\mathbf{H}_n(\boldsymbol{\theta}) = \begin{pmatrix} \dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\dot{\mathbf{G}}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) & \dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\dot{\mathbf{G}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \\ \dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\dot{\mathbf{G}}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) & \dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\dot{\mathbf{G}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \end{pmatrix}. \tag{A.7}$$

A.2. Assumptions. We denote $(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ which are i.i.d. samples from population $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ with $\mathbf{Y} = (Y_1, \ldots, Y_T)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)^{\mathrm{T}}$, and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_T)^{\mathrm{T}}$ for correlated data with cluster size T. Denote $C^{(r)}([0, 1]) = \{f: f$ has continuous derivatives up to order r on $[0, 1]\}$ as the space of the rth order smooth functions on $[0, 1]$. For any vector $\mathbf{a}$, let $\|\mathbf{a}\|$ be the usual Euclidean norm. For any matrix $\mathbf{A}$, let $\|\mathbf{A}\|$ be the modulus of the largest singular value of $\mathbf{A}$. We provide the regularity conditions to obtain Theorems 1–5.

(C1) For some $r \geq 2$, $\alpha_{0,l} \in C^{(r)}([0, 1])$ $l = 1, \ldots, d$.

(C2) *The covariance matrix* $\boldsymbol{\Sigma} = E\mathbf{e}\mathbf{e}^{\mathrm{T}}$ *is positive definite, and* $E\|\mathbf{e}\|^{2+\delta} < +\infty$ *for some* $\delta > 0$.

(C3) *For each* $\mathbf{X}_t$, $t = 1, \ldots, T$, its density function $f_t(\mathbf{x})$ *is absolutely continuous and bounded away from zero and* ∞ *on a compact support* $\chi = [0, 1]^{d_x}$.

(C4) *The number of nonzero components in the nonparametric part* s_x *is fixed; there exists* $c_\alpha > 0$ *such that* $\min_{1 \leq l \leq s_x} \|\alpha_{0,l}\| > c_\alpha$. *The nonzero coefficients in the linear part satisfy that* $\min_{1 \leq k \leq s_z} \|\beta_{0k}\|/\lambda_{2n} \to \infty$.

(C5) *The eigenvalues of* $E(\boldsymbol{\Gamma}_0^{(k)})$ *are bounded away from* 0 *and* ∞, *uniformly in* $k = 1, \ldots, K$, *for sufficiently large* n.

(C6) *The second derivative of* $g^{-1}(\cdot)$ *exists and is bounded; function* $V(\cdot)$ *has a bounded second derivative, and is bounded away from* 0 *and* ∞.

(C7) *The modular of the singular value of* $\mathbf{M} = (\mathbf{M}_1^{\mathrm{T}}, \dots, \mathbf{M}_K^{\mathrm{T}})^{\mathrm{T}}$ *is bounded away from* 0 *and* ∞.

(C8) *The eigenvalues of* $E(\mathbf{X}_t \mathbf{X}_t^{\mathrm{T}} | \mathbf{Z}_t)$ *are bounded away from* 0 *and* ∞, *uniformly in* $1 \le t \le T$.

(C9) *There is a large enough open subset* $\widetilde{\Theta}_n \in R^{d_n}$ *which contains* $\widetilde{\boldsymbol{\theta}}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, *for* $\widetilde{\boldsymbol{\gamma}}$ *in Section* A.3, *such that* $\sup_{\boldsymbol{\theta} \in \widetilde{\Theta}_n} |n^{-1} \frac{\partial^3 Q_n(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k \partial \theta_l}| = O_P(\rho_n^{-1})$.

(P1) $\liminf_{n \to \infty} \liminf_{\theta \to 0+} p_{\lambda_{j,n}}(\theta)/\lambda_{j,n} > 0$, $j = 1, 2$.

(P2) $a_n = o(1/\sqrt{n d_n})$, $b_n = o(d_n^{-1/2})$, *where* a_n *and* b_n *are defined in (*3.1*)*.

Conditions (C1)–(C3) are quite standard in the spline smoothing literature. Assumptions similar to (C1)–(C3) can be found in [14, 15, 31] and [33]. The smoothness condition in (C1) controls the rate of convergence of the spline estimators $\widehat{\alpha}_l$, $l = 1, \dots, d_x$, and $\widehat{\alpha}$. Conditions (C5) and (C6) are similar to assumptions (A3) and (A4) in [12], which can be verified for other distributions as well. The boundedness condition in condition (C7) is essentially a requirement that the matrix $\mathbf{C}_n$ in (2.5) is asymptotically positive definite. This assumption is clearly satisfied if the basis matrices are exchangeable or AR-1 correlation structures as discussed previously. The condition on eigenvalues in (C8) is to ensure that we do not have a multicolinear problem. Condition (C9) controls the magnitude of the third-order derivative of the quadratic inference function. Similar conditions have been assumed in [5] and [9]. Here, we require a slightly stronger condition. Instead of assuming boundedness, we require it be of the order $O_P(\rho_n^{-1})$, where $\rho_n = ((1-\delta)/2)^{(d_x - 1)/2}$ to facilitate the technical derivation for the nonparametric components in a GAPLM model, while both [5] and [9] consider pure parametric models.

A.3. Proof of Theorem 1. According to Lemma A.7 of [33], for any function $\alpha \in \mathcal{M}$ with $\alpha_l \in C^{(r)}([0, 1])$, $l = 1, \dots, d_x$, there exists an additive spline function $\widetilde{\alpha} = \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{B} \in \mathcal{M}_n$ and a constant C such that

$$\|\widetilde{\alpha} - \alpha\|_\infty \le C d_x J_n^{-r}. \tag{A.8}$$

From the results of Lemma S.10 in the online supplementary material [27] and Lemma A.6 in the online supplement of [33], we have

$$\begin{aligned} \|\mathbf{B}^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\gamma}})\|_n^2 &= \frac{1}{n} \sum_{l=1}^{d_x} \sum_{i=1}^{n} \sum_{t=1}^{T} [\mathbf{B}_l^{\mathrm{T}}(x_{it}^{(l)})(\widehat{\boldsymbol{\gamma}}_l^{\mathrm{QIF}} - \widetilde{\boldsymbol{\gamma}}_l)]^2 \\ &= O_P(n^{-1} d_n). \end{aligned} \tag{A.9}$$

The triangular inequality implies that, for each $l = 1, \dots, d_x$,

$$\begin{aligned} \frac{1}{n} \sum_{l=1}^{d_x} \sum_{i=1}^{n} \sum_{t=1}^{T} [\widehat{\alpha}_l(x_{it}^{(l)}) - \alpha_{0,l}(x_{it}^{(l)})]^2 &\le \frac{2}{n} \sum_{l=1}^{d_x} \sum_{i=1}^{n} \sum_{t=1}^{T} [\mathbf{B}_l^{\mathrm{T}}(x_{it}^{(l)})(\widehat{\boldsymbol{\gamma}}_l^{\mathrm{QIF}} - \widetilde{\boldsymbol{\gamma}}_l)]^2 \\ &\quad + C d_x J_n^{-2r}. \end{aligned}$$

This completes the proof.

A.4. Proof of Theorem 2. To study the asymptotic properties of $\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}}$, we consider the case that α_0 in (2.1) can be estimated at reasonable accuracy, for example, we can approximate α_0 by the spline smoother $\widetilde{\alpha}$ in (A.8). We begin our proof by replacing α_0 with $\widetilde{\alpha}$ and defining an *intermediate* QIF estimator for $\boldsymbol{\beta}_0$.

For any fixed $\boldsymbol{\beta}$ and $i = 1, \ldots, n$, we denote $\widetilde{\boldsymbol{\eta}}_i(\boldsymbol{\beta}) = \widetilde{\alpha}(\mathbf{X}_i) + \mathbf{Z}_{it}^{\mathrm{T}}\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\mu}}_i(\boldsymbol{\beta}) = g^{-1}\{\widetilde{\boldsymbol{\eta}}_i(\boldsymbol{\beta})\}$. Let $\dot{\widetilde{\mu}}_{it}(\boldsymbol{\beta})$ be the first-order derivative of $g^{-1}(\eta)$ evaluated at $\eta = \widetilde{\boldsymbol{\eta}}_i(\boldsymbol{\beta})$. Define $\widetilde{\boldsymbol{\Delta}}_i(\boldsymbol{\beta}) = \mathrm{diag}\{\dot{\widetilde{\mu}}_{i1}(\boldsymbol{\beta}), \ldots, \dot{\widetilde{\mu}}_{iT}(\boldsymbol{\beta})\}$ and $\widetilde{\mathbf{A}}_i = \mathrm{diag}\{V(\widetilde{\mu}_{i1}), \ldots, V(\widetilde{\mu}_{iT})\}$. Let

$$\text{(A.10)} \qquad \widetilde{\mathbf{g}}_i(\boldsymbol{\beta}) = \mathbf{g}_i(\boldsymbol{\beta}, \widetilde{\boldsymbol{\gamma}}) = \begin{pmatrix} \mathbf{D}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Delta}}_i\widetilde{\mathbf{A}}_i^{-1/2}\mathbf{M}_1\widetilde{\mathbf{A}}_i^{-1/2}(\mathbf{Y}_i - \widetilde{\boldsymbol{\mu}}_i(\boldsymbol{\beta})) \\ \vdots \\ \mathbf{D}_i^{\mathrm{T}}\widetilde{\boldsymbol{\Delta}}_i\widetilde{\mathbf{A}}_i^{-1/2}\mathbf{M}_K\widetilde{\mathbf{A}}_i^{-1/2}(\mathbf{Y}_i - \widetilde{\boldsymbol{\mu}}_i(\boldsymbol{\beta})) \end{pmatrix}.$$

Define $\widetilde{\mathbf{G}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{g}}_i(\boldsymbol{\beta})$. In a similar way, we define $\widetilde{\mathbf{C}}_n(\boldsymbol{\beta})$, and $\widetilde{Q}_n(\boldsymbol{\beta})$. Let $\widetilde{\boldsymbol{\beta}}_{\mathrm{QIF}} = \arg\min_{\boldsymbol{\beta}} n^{-1}\widetilde{Q}_n(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}}\{\widetilde{\mathbf{G}}_n^{\mathrm{T}}(\boldsymbol{\beta})\widetilde{\mathbf{C}}_n^{-1}(\boldsymbol{\beta})\widetilde{\mathbf{G}}_n(\boldsymbol{\beta})\}$. The asymptotic properties of $\widetilde{\boldsymbol{\beta}}_{\mathrm{QIF}}$ are given in the supplementary material [27]. Let $\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{QIF}}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_{\mathrm{QIF}}^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\boldsymbol{\theta}}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}}^{\mathrm{QIF}} = (\widetilde{\boldsymbol{\beta}}_{\mathrm{QIF}}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$.

PROOF OF THEOREM 2. By Taylor expansion,

$$\begin{aligned} &\dot{Q}_n(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}}) - \dot{Q}_n(\widetilde{\boldsymbol{\theta}}_0) \\ &\qquad = \ddot{Q}_n(\widetilde{\boldsymbol{\theta}}_0)(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)^{\mathrm{T}} \frac{\partial \dot{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0), \end{aligned}$$

where $\boldsymbol{\theta}^* = t\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} + (1-t)\widetilde{\boldsymbol{\theta}}_0$, for some $t \in [0, 1]$. Since $\dot{Q}_n(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}}) = 0$,

$$-\dot{Q}_n(\widetilde{\boldsymbol{\theta}}_0) = \ddot{Q}_n(\widetilde{\boldsymbol{\theta}}_0)(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0) + \frac{1}{2}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)^{\mathrm{T}} \frac{\partial \dot{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0).$$

According to the Cauchy–Schwarz inequality, one has

$$\left\|\frac{1}{n}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)^{\mathrm{T}}\frac{\partial \dot{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)\right\|^2 \le \|\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0\|^4 \frac{1}{n}\sum_{j,k,l=1}^{d_n}\left\{\frac{\partial^3 Q_n(\boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k\,\partial\theta_l}\right\}^2.$$

Lemma S.10 and condition (C9) implies that

$$\begin{aligned} \left\|\frac{1}{n}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)^{\mathrm{T}}\frac{\partial \dot{Q}_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0)\right\|^2 &\le \|\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0\|^4 \times O_P(\rho_n^{-1}d_n^3) \\ &= O_P(n^{-2}d_n^2) \times O_P(\rho_n^{-1}d_n^3) = o_P(n^{-1}). \end{aligned}$$

Next by (A.4) and (A.5), we have

$$\begin{aligned} -\{2\mathbf{S}_n(\widetilde{\boldsymbol{\theta}}_0) + O_P(\rho_n^{-1}n^{-1}d_n)\} &= \{2\mathbf{H}_n(\widetilde{\boldsymbol{\theta}}_0) + o_P(1)\}(\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\theta}}_0) \\ &\qquad + o_P(n^{-1/2}), \end{aligned}$$

where $\mathbf{S}_n(\boldsymbol{\theta})$ and $\mathbf{H}_n(\boldsymbol{\theta})$ are defined in (A.2). Thus,

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}^{\mathrm{QIF}} - \boldsymbol{\beta}_0 \\ \widehat{\boldsymbol{\gamma}}^{\mathrm{QIF}} - \widetilde{\boldsymbol{\gamma}} \end{pmatrix} = -\left[2 \begin{pmatrix} \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}} \\ \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{pmatrix} + o_P(1) \right]^{-1}$$
$$\times \left[2\mathbf{S}_n(\widetilde{\boldsymbol{\theta}}_0) + O_P(\rho_n^{-1} n^{-1} d_n)\right] + o_P(n^{-1/2}),$$

which leads to

$$\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}} - \boldsymbol{\beta}_0 = \{\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}} \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}\}^{-1} (\mathbf{I}, \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}} \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}) \mathbf{S}_n(\widetilde{\boldsymbol{\theta}}_0)$$
$$+ O_P(\rho_n^{-1} n^{-1} d_n) + o_P(n^{-1/2}).$$

According to (A.6),

$$\begin{aligned} &(\mathbf{I}, \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\theta}) \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\boldsymbol{\theta})) \mathbf{S}_n(\boldsymbol{\theta}) \\ &\quad = (\mathbf{I}, \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\theta}) \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\boldsymbol{\theta})) (\dot{\mathbf{G}}_{\boldsymbol{\beta}}(\boldsymbol{\theta}), \dot{\mathbf{G}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}))^{\mathrm{T}} \mathbf{C}_n^{-1}(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta}) \\ &\quad = \{\dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\theta}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\theta}) \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\boldsymbol{\theta}) \dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\boldsymbol{\theta})\} \mathbf{C}_n^{-1}(\boldsymbol{\theta}) \mathbf{G}_n(\boldsymbol{\theta}). \end{aligned}$$

Hence, the asymptotic distribution of $\sqrt{n}\mathbf{A}_n \boldsymbol{\Sigma}_n^{-1/2}(\widehat{\boldsymbol{\beta}}^{\mathrm{QIF}} - \boldsymbol{\beta}_0)$ is the same as that of

$$\sqrt{n}\mathbf{A}_n \boldsymbol{\Sigma}_n^{-1/2} \{\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_0) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_0) \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_0) \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_0)\}^{-1}$$
$$\times \{(\dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_0) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}} \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} \dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_0)) \mathbf{C}_n^{-1}(\widetilde{\boldsymbol{\theta}}_0) \mathbf{G}_n(\widetilde{\boldsymbol{\theta}}_0)\}.$$

The desired result follows from Lemmas S.11 and S.12. □

A.5. Proof of Theorem 3. In the following, let $L_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta}) + n\sum_{l=1}^{d_x} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) + \sum_{l=1}^{d_z} p_{\lambda_{2,n}}(|\boldsymbol{\beta}_l|)$ be the object function in (3.1). Let $\Theta_{\mathcal{A}} = \{\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}} : \beta_{s_z+1} = \cdots = \beta_{d_z} = 0, \boldsymbol{\gamma}_{s_x+1} = \cdots = \boldsymbol{\gamma}_{d_x} = \mathbf{0}\}$ and define $\widehat{\boldsymbol{\theta}}_{\mathcal{A}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_{\mathcal{A}}^{\mathrm{T}})^{\mathrm{T}} = \arg\min_{\boldsymbol{\theta}\in\Theta_{\mathcal{A}}} Q_n(\boldsymbol{\theta})$, which leads to the spline QIF estimator of the nonzero components, given that the rest terms are zero. Note that $\|\mathbf{B}^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_{\mathcal{A}} - \widetilde{\boldsymbol{\gamma}})\|_n = O_P(n^{-1/2} d_n^{1/2})$ and $\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} d_n^{1/2})$ from the results of Theorems 1 and 2. It is sufficient to show that for large n and any $\epsilon > 0$, there exists a sufficient large constant C such that

$$P\Big\{ \inf_{\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}\| = C\rho_n^{-3} d_n^{1/2}(n^{-1/2} + a_n)} L_n(\boldsymbol{\theta}) > L_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \Big\} \geq 1 - \epsilon. \tag{A.11}$$

Equation (A.11) implies that $L_n(\cdot)$ has a local minimum in the set $\Theta^*(C) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}\| \leq C\rho_n^{-3} d_n^{1/2}(n^{-1/2} + a_n)\}$. Thus, one has $\|\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}\| = O_P\{\rho_n^{-3} d_n^{1/2}(n^{-1/2} + a_n)\}$. Further, the triangular inequality yields that $\|\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \boldsymbol{\theta}_0\| \leq \|\widehat{\boldsymbol{\theta}}^{\mathrm{QIF}} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}\| + \|\widehat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_0\| = O_P\{\rho_n^{-3} d_n^{1/2}(n^{-1/2} + a_n)\}$. The theorem follows from condition (C4).

In the following, we show that (A.11) holds. Observing that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) \geq 0$, one has

$$L_n(\boldsymbol{\theta}) - L_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \geq Q_n(\boldsymbol{\theta}) - Q_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) + \sum_{l=1}^{s_x} n\{p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) - p_{\lambda_{1,n}}(\|\widehat{\boldsymbol{\gamma}}_{\mathcal{A}}\|_{\mathbf{K}_l})\} + \sum_{l=1}^{s_z} n\{p_{\lambda_{2n}}(|\beta_l|) - p_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|)\}.$$

Note that

$$Q_n(\boldsymbol{\theta}) - Q_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \ddot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}) + R_n^*,$$

where

$$|R_n^*| = \frac{C^3}{6}\left| \sum_{i,j,k=1}^{d_n} \frac{\partial Q_n(\boldsymbol{\theta}^*)}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k}(\theta_i - \widehat{\theta}_{\mathcal{A},i})(\theta_j - \widehat{\theta}_{\mathcal{A},j})(\theta_k - \widehat{\theta}_{\mathcal{A},k})\right|,$$

$\boldsymbol{\theta}^* = t\widehat{\boldsymbol{\theta}}_{\mathcal{A}} + (1-t)\boldsymbol{\theta}$ for some $t \in [0, 1]$.

Following [23] and Lemma S.4, for any $\boldsymbol{\theta} \in \Theta^*(C)$, one has

$$\begin{aligned}
&(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \\
&\quad = n(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{\mathbf{G}}_n^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \mathbf{C}_n^{-1}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \mathbf{G}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})\{1 + o_P(1)\} \\
&\quad \leq C\rho_n^{-4} n^{1/2} d_n (n^{-1/2} + a_n), \\
&(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \ddot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \\
&\quad = n(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{\mathbf{G}}_n^{\mathrm{T}}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \mathbf{C}_n^{-1}(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \dot{\mathbf{G}}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})\{1 + o_P(1)\} \\
&\quad \geq C^2 \rho_n^{-4} n d_n (n^{-1/2} + a_n)^2.
\end{aligned}$$

By the Cauchy–Schwarz inequality, $|R_n^*| \leq \frac{C^3}{6}\{\sum_{i,j,k=1}^{d_n}(\frac{\partial^3 Q_n(\boldsymbol{\theta}^*)}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k})^2\}^{1/2} \times \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}}\|^3$. According to assumption $n^{-1}d_n^4 = o(1)$, one has

$$\begin{aligned}
&Q_n(\boldsymbol{\theta}) - Q_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) \\
&\quad = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}}) + \tfrac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \ddot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})\{1 + o_P(1)\}.
\end{aligned} \tag{A.12}$$

Thus, for sufficiently large C, the first term $(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \dot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})$ is dominated by the second term $\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})^{\mathrm{T}} \ddot{Q}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{A}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{\mathcal{A}})$.

Following the proof of Theorem 2 in [31], if $\lambda_{1n} \to 0$, then for any $\boldsymbol{\gamma}$ with $\|\mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{\mathcal{A}})\|_n = C_2\rho_n^{-3} n^{-1/2} d_n^{1/2}$, one has $\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l} \geq a\lambda_{1,n}$, and $\|\widehat{\boldsymbol{\gamma}}_{\mathcal{A},l}\|_{\mathbf{K}_l} \geq a\lambda_{1,n}$ for each $l = 1, \ldots, s_x$, when n is large enough. By the definition of the

SCAD penalty, $\sum_{l=1}^{s_x}\{p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) - p_{\lambda_{1,n}}(\|\widehat{\boldsymbol{\gamma}}_{\mathcal{A},l}\|_{\mathbf{K}_l})\} = 0$ for large n. Furthermore, for any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\mathcal{A}}\| \le C\rho_n^{-3}d_n^{1/2}(n^{-1/2} + a_n)$,

$$\begin{aligned}
&\sum_{l=1}^{s_z} n\{p_{\lambda_{2,n}}(|\beta_l|) - p_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|)\}\\
&\qquad = \sum_{l=1}^{s_z} np'_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|)(\beta_l - \widehat{\beta}_{\mathcal{A},l})\operatorname{sgn}(\widehat{\beta}_{\mathcal{A},l})\\
&\qquad\quad + \sum_{l=1}^{s_z} np''_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|)(\beta_l - \widehat{\beta}_{\mathcal{A},l})^2\operatorname{sgn}(\widehat{\beta}_{\mathcal{A},l})\{1 + o(1)\},
\end{aligned}$$

$p'_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|) - p'_{\lambda_{2n}}(|\beta_{0,l}|) = p''_{\lambda_{2,n}}(|\beta_{0,l}|)(\widehat{\beta}_{\mathcal{A},l} - \beta_{0,l})\operatorname{sgn}(\beta_{0,l})\{1 + o(1)\}$. Thus,

$$\begin{aligned}
&\sum_{l=1}^{s_z} p'_{\lambda_{2,n}}(|\widehat{\beta}_{0,l}|)(\beta_l - \widehat{\beta}_{\mathcal{A},l})\operatorname{sgn}(\widehat{\beta}_{\mathcal{A},l})\\
&\qquad = \sum_{l=1}^{s_z}\{p'_{\lambda_{2n}}(|\beta_{0,l}|)\}(\beta_l - \widehat{\beta}_{\mathcal{A},l})\operatorname{sgn}(\widehat{\beta}_{\mathcal{A},l})\\
&\qquad \le C\rho_n^{-3}s_z^{1/2}a_n d_n^{1/2}(n^{-1/2} + a_n) \le C\rho_n^{-3}d_n(n^{-1/2} + a_n)^2.
\end{aligned}$$

Meanwhile,

$$\sum_{l=1}^{s_z} np''_{\lambda_{2,n}}(|\widehat{\beta}_{0,l}|)(\beta_l - \widehat{\beta}_{\mathcal{A},l})^2\operatorname{sgn}(\widehat{\beta}_{\mathcal{A},l}) \le C^2\rho_n^{-6}nb_n d_n(n^{-1/2} + a_n)^2.$$

Hence, $\sum_{l=1}^{s_z} n\{p_{\lambda_{2,n}}(|\beta_l|) - p_{\lambda_{2,n}}(|\widehat{\beta}_{\mathcal{A},l}|)\}$ is also dominated by the second term of (A.12). Hence, by choosing a sufficiently large C, (A.11) holds for large n. The proof of Theorem 3 is completed.

A.6. Proof of Theorem 4. Let $\varrho_{n,d} = \rho_n^{-3}n^{-1/2}d_n^{1/2}$, and define $\Theta_1 = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta_{\mathcal{A}}, \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(\varrho_{n,d}), \|\mathbf{B}^{\mathrm{T}}(\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}})\|_n = O_P(\varrho_{n,d})\}$, $\Theta_l = \{(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}, \boldsymbol{\gamma} = (\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{\gamma}_l^{\mathrm{T}}, \mathbf{0}, \ldots, \mathbf{0})^{\mathrm{T}}, \|\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}\|_n = O_P(\varrho_{n,d})\}$ for $l = s_x + 1, \ldots, d_x$ and $\Theta_l = \{(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}} : \boldsymbol{\beta}_2 = \mathbf{0}, \boldsymbol{\gamma} = \mathbf{0}, \|\boldsymbol{\beta}_1\| = O_P(\varrho_{n,d})\}$ for $l = d_x + 1$. It suffices to show that uniformly for any $\boldsymbol{\theta} \in \Theta_1$ and $\boldsymbol{\theta}_l^* \in \Theta_l$, $L_n(\boldsymbol{\theta}) \le L_n(\boldsymbol{\theta} + \boldsymbol{\theta}_l^*)$, with probability 1 as $n \to \infty$ for any $s_x + 1 \le l \le d_x + 1$. Observe that, for $l = s_x + 1, \ldots, d_x$,

$$\begin{aligned}
&L_n(\boldsymbol{\theta} + \boldsymbol{\theta}_l^*) - L_n(\boldsymbol{\theta})\\
&\qquad = Q_n(\boldsymbol{\theta} + \boldsymbol{\theta}_l^*) - Q_n(\boldsymbol{\theta}) + np'_{\lambda_{1,n}}(w_l)(\|\boldsymbol{\gamma}_l^*\|_{\mathbf{K}_l})\\
&\qquad = \boldsymbol{\gamma}_l^{*\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\gamma}_l}Q_n(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\gamma}_l^{*\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\gamma}_l\,\partial\boldsymbol{\gamma}_l^{\mathrm{T}}}Q_n(\boldsymbol{\theta})\boldsymbol{\gamma}_l^*\{1 + o_P(1)\}
\end{aligned}$$

$$+ np'_{\lambda_{1,n}}(w_l)(\|\boldsymbol{\gamma}_l^*\|_{\mathbf{K}_l})$$
$$= n\lambda_{1,n}\|\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_l\|_n \left\{\frac{R_n}{\lambda_{1,n}} + \frac{p'_{\lambda_{1n}}(w_l)}{\lambda_{1,n}}\right\}\{1 + o_P(1)\},$$

where w_l is a value between 0 and $\|\boldsymbol{\gamma}_l^*\|_{\mathbf{K}_l}$ and

$$R_n = n^{-1}\|\mathbf{B}^{\mathrm{T}}\boldsymbol{\gamma}_l\|_n^{-1}\left\{\boldsymbol{\gamma}_l^{*\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\gamma}_l}Q_n(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\gamma}_l^{*\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\gamma}_l\,\partial\boldsymbol{\gamma}_l^{\mathrm{T}}}Q_n(\boldsymbol{\theta})\boldsymbol{\gamma}_l^*\{1 + o_P(1)\}\right\}$$
$$= O_P(\rho_n^{-4}n^{-1/2}d_n^{1/2}).$$

Noting that $R_n/\lambda_{1,n} = o_P(1)$, and $\liminf_{n\to\infty}\liminf_{w\to 0^+} p'_{\lambda_{1n}}(w)/\lambda_{1,n} = 1$, thus, uniformly for any $\boldsymbol{\theta} \in \Theta_1$ and $\boldsymbol{\theta}_l^* \in \Theta_l$, $L_n(\boldsymbol{\theta}) \le L_n(\boldsymbol{\theta} + \boldsymbol{\theta}_l^*)$, with probability tending to 1 as $n \to \infty$ for any $l = s_x + 1, \dots, d_x$. On the other hand, for $l = d_x + 1$,

$$L_n(\boldsymbol{\theta} + \boldsymbol{\theta}_l^*) - L_n(\boldsymbol{\theta})$$
$$= \boldsymbol{\beta}_1^{*\mathrm{T}}\frac{\partial}{\partial\boldsymbol{\beta}_1}Q_n(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\beta}_1^{*\mathrm{T}}\frac{\partial^2}{\partial\boldsymbol{\beta}_1\,\partial\boldsymbol{\beta}_1^{\mathrm{T}}}Q(\boldsymbol{\theta})\boldsymbol{\beta}_1^*\{1 + o_P(1)\}$$
$$+ n\sum_{q=1}^{s_z} p'_{\lambda_{2,n}}(|w_{l,q}|)\beta_q^*\operatorname{sgn}(\beta_q^*).$$

Similar arguments show that uniformly for any $\boldsymbol{\theta} \in \Theta_1$ and $\boldsymbol{\theta}_{d_x+1}^* \in \Theta_{d_x+1}$, $L_n(\boldsymbol{\theta}) \le L_n(\boldsymbol{\theta} + \boldsymbol{\theta}_{d_x+1}^*)$, with probability tending to 1 as $n \to \infty$. This establishes the desired result.

A.7. Proof of Theorem 5. Let $\boldsymbol{\beta}_{\mathcal{S}} = (\beta_1, \dots, \beta_{s_z})^{\mathrm{T}}$, $\boldsymbol{\gamma}_{\mathcal{S}} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \dots, \boldsymbol{\gamma}_{s_x}^{\mathrm{T}})^{\mathrm{T}}$. Denote $\boldsymbol{\theta}_{\mathcal{S}} = (\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathcal{S}}^{\mathrm{T}})^{\mathrm{T}}$. In a similar way, define $\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0} = (\boldsymbol{\beta}_{\mathcal{S}0}^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}_{\mathcal{S}}^{\mathrm{T}})^{\mathrm{T}}$, in which $\boldsymbol{\beta}_{\mathcal{S}0} = (\beta_{10}, \dots, \beta_{s_z0})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\gamma}}_{\mathcal{S}} = (\widetilde{\boldsymbol{\gamma}}_1^{\mathrm{T}}, \dots, \widetilde{\boldsymbol{\gamma}}_{s_x}^{\mathrm{T}})^{\mathrm{T}}$. It can be shown easily that there exist $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$ and $\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}$ minimizing $L_n((\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}, (\boldsymbol{\gamma}_{\mathcal{S}}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}})$, that is,

$$\frac{\partial}{\partial\boldsymbol{\theta}_{\mathcal{S}}}L_n(\boldsymbol{\theta})\Big|_{\boldsymbol{\beta}=(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}},\boldsymbol{\gamma}=(\widehat{\boldsymbol{\gamma}}_{\mathcal{S}}^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}}} = \mathbf{0}.$$

In the following, we consider $L_n(\cdot)$ as a function of $\boldsymbol{\theta}_{\mathcal{S}}$, and denote $\dot{L}_n$ and $\ddot{L}_n$ the gradient vector and Hessian matrix of $L_n(\cdot)$ with respect to $\boldsymbol{\theta}_{\mathcal{S}}$. The rest of the proof follows similarly as that of Theorem 2. Using Taylor expansion, one has

$$\dot{L}_n(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}}) - \dot{L}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) = \ddot{L}_n(\boldsymbol{\theta}_{\mathcal{S}}^*)(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})$$
$$+ \frac{1}{2}(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})^{\mathrm{T}}\frac{\partial\ddot{L}_n(\boldsymbol{\theta}_{\mathcal{S}})}{\partial\boldsymbol{\theta}_{\mathcal{S}}\,\partial\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}}\Big|_{\boldsymbol{\theta}_{\mathcal{S}}=\boldsymbol{\theta}_{\mathcal{S}}^*}(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}),$$

where $\boldsymbol{\theta}_{\mathcal{S}}^{*} = t\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} + (1-t)\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}$, for some $t \in [0, 1]$. Thus, we have

$$\begin{aligned}
&-n^{-1}\dot{Q}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \boldsymbol{\kappa}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) \\
&\quad = n^{-1}\{\ddot{Q}_n(\boldsymbol{\theta}_{\mathcal{S}0}) + \boldsymbol{\Lambda}(\boldsymbol{\theta}_{\mathcal{S}0})\}(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) \\
&\qquad + \frac{1}{2}n^{-1}(\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})^{\mathrm{T}} \frac{\partial \dot{Q}_n(\boldsymbol{\theta}_{\mathcal{S}})}{\partial \boldsymbol{\theta}_{\mathcal{S}}\, \partial \boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}}\bigg|_{\boldsymbol{\theta}_{\mathcal{S}} = \boldsymbol{\theta}_{\mathcal{S}}^{*}} (\widehat{\boldsymbol{\theta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}),
\end{aligned}$$

where $\boldsymbol{\kappa}_n^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) = (\{p'_{\lambda_{2,n}}(|\beta_{0,k}|)\operatorname{sgn}(\beta_{0,k})\}_{k=1}^{s_z}, \{\frac{\partial}{\partial \boldsymbol{\gamma}_l} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})|_{\boldsymbol{\gamma}_l = \widetilde{\boldsymbol{\gamma}}_l}\}_{l=1}^{s_x})$ and $\boldsymbol{\Lambda}(\boldsymbol{\theta}_{\mathcal{S}0}) = \operatorname{diag}(\{p''_{\lambda_{2,n}}(|\beta_{0,k}|)\}_{k=1}^{s_z}, \{\frac{\partial^2}{\partial \boldsymbol{\gamma}_l\, \partial \boldsymbol{\gamma}_l^{\mathrm{T}}} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})|_{\boldsymbol{\gamma}_l = \widetilde{\boldsymbol{\gamma}}_l}\}_{l=1}^{s_x})$. Note that

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\gamma}_l} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) = p'_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}^{-1}\mathbf{K}_l\boldsymbol{\gamma}_l, \\
&\frac{\partial^2}{\partial \boldsymbol{\gamma}_l\, \partial \boldsymbol{\gamma}_l^{\mathrm{T}}} p_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}) \\
&\quad = p'_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}^{-1}\mathbf{K}_l \\
&\qquad + \{p''_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}^{-2} - p'_{\lambda_{1,n}}(\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l})\|\boldsymbol{\gamma}_l\|_{\mathbf{K}_l}^{-3}\}\mathbf{K}_l\boldsymbol{\gamma}_l\boldsymbol{\gamma}_l^{\mathrm{T}}\mathbf{K}_l,
\end{aligned}$$

$J_n \to \infty$ and $\lambda_{1,n} \to 0$, as $n \to \infty$, so $\|\widetilde{\boldsymbol{\gamma}}_l\|_{\mathbf{K}_l} \geq a\lambda_{1,n}$ for n large enough and for each $l = 1, \dots, s_x$. Thus, $p'_{\lambda_{1,n}}(\|\widetilde{\boldsymbol{\gamma}}_l\|_{\mathbf{K}_l}) = 0$ and $p''_{\lambda_{1,n}}(\|\widetilde{\boldsymbol{\gamma}}_l\|_{\mathbf{K}_l}) = 0$.

Similar to the proof of Theorem 2, one has

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \boldsymbol{\beta}_{\mathcal{S}0} &= \{(\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}_{\mathcal{S}0}) + \boldsymbol{\Lambda}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\mathcal{S}0})(\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\mathcal{S}0}))^{-1}\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\theta}_{\mathcal{S}0})\}^{-1} \\
&\quad \times (\mathbf{I}, \mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\theta}_{\mathcal{S}}^{*})(\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\mathcal{S}}^{*}))^{-1})\{\mathbf{S}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\kappa}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\} \\
&\quad + O_P(\rho_n^{-1}n^{-1}d_n) + o_P(n^{-1/2}).
\end{aligned}$$

Note that

$$\begin{aligned}
&(\mathbf{I}, \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}))\{\mathbf{S}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\kappa}_n(\widetilde{\boldsymbol{\theta}}_0)\} \\
&\quad = \{\dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\}\mathbf{C}_n^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{G}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\kappa}_{\mathcal{S}}.
\end{aligned}$$

The asymptotic distribution of $\sqrt{n}\mathbf{A}_n\boldsymbol{\Sigma}_{\mathcal{S},n}^{-1/2}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{PQIF}} - \boldsymbol{\beta}_{\mathcal{S}0})$ is the same as that of

$$\begin{aligned}
&\sqrt{n}\mathbf{A}_n\boldsymbol{\Sigma}_{\mathcal{S},n}^{-1/2}\{\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\Lambda}_{\mathcal{S}0}\}^{-1} \\
&\quad \times \{(\dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}))\mathbf{C}_n^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{G}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) \\
&\qquad + \boldsymbol{\kappa}_{\mathcal{S}}\}.
\end{aligned}$$

Next, write $\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}} = \{(\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(1)})^{\mathrm{T}}, \dots, (\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(K)})^{\mathrm{T}}\}^{\mathrm{T}}$, where $\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i^{\mathrm{T}}\boldsymbol{\Gamma}_{0,i}^{(k)}\widehat{\mathbf{Z}}_{\mathcal{S}i}$ and $\widehat{\mathbf{Z}}_{\mathcal{S}i} = \mathbf{Z}_{\mathcal{S}i} - \mathbf{B}_{\mathcal{S}i}\{\mathbf{J}_{\mathrm{DB_S}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\mathbf{J}_{\mathrm{DB_S}}\}^{-1}\mathbf{J}_{\mathrm{DB_S}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\mathbf{J}_{\mathrm{DZ_S}}$. Then we can express

$$\mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) = \widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}}\{1 + o_P(1)\}.$$

Using similar arguments as given in Lemma S.11, we know

$$
\begin{aligned}
&\mathbf{A}_n \boldsymbol{\Sigma}_{\mathcal{S},n}^{-1/2}\{\mathbf{H}_{\boldsymbol{\beta\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\Lambda}_{\mathcal{S}0}\}^{-1} \\
&\quad \times (\dot{\mathbf{G}}_{\boldsymbol{\beta}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\dot{\mathbf{G}}_{\boldsymbol{\gamma}}^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}))\mathbf{C}_n^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{G}_n(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) \\
&\qquad = \mathbf{A}_n \boldsymbol{\Sigma}_{\mathcal{S},n}^{-1/2}\{\mathbf{H}_{\boldsymbol{\beta\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) - \mathbf{H}_{\boldsymbol{\beta\gamma}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma\gamma}}^{-1}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0})\mathbf{H}_{\boldsymbol{\gamma\beta}}(\widetilde{\boldsymbol{\theta}}_{\mathcal{S}0}) + \boldsymbol{\Lambda}_{\mathcal{S}0}\}^{-1} \\
&\qquad\quad \times \widehat{\mathbf{J}}_{\mathrm{DZ_S}}^{\mathrm{T}}(\mathbf{C}_n^0)^{-1}\mathbf{G}_n^0 + o_P(n^{-1/2}).
\end{aligned}
$$

Thus, the desired result follows.

Acknowledgements. The authors thank the Co-Editor, an Associate Editor and three referees for their constructive comments that have substantially improved an earlier version of this paper. L. Wang and L. Xue contribute equally for the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates" (DOI: 10.1214/13-AOS1194SUPP; .pdf). The supplementary material provides a number of technical lemmas and their proofs. The technical lemmas are used in the proofs of Theorems 1–5 in the paper.

REFERENCES

[1] Boni, J. P., Leister, C., Bender, G., Fitzpatrick, V., Twine, N., Stover, J., Dorner, A., Immermann, F. and Burczynski, M. E. (2005). Population pharmacokinetics of CCI-779: Correlations to safety and pharmacogenomic responses in patients with advanced renal cancer. *Clinical Pharmacology & Therapeutics* **77** 76–89.

[2] Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249

[3] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

[4] Cheng, G., Zhou, L. and Huang, J. Z. (2014). Efficient semiparametric estimation in generalized partially linear additive models for longitudinal/clustered data. *Bernoulli* **20** 141–163. MR3160576

[5] Cho, H. and Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Statist. Sinica* **23** 901–927. MR3086662

[6] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[8] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[9] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

[10] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

[11] HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparametric Models*. Springer, New York. MR2061786

[12] HE, X., FUNG, W. K. and ZHU, Z. (2005). Robust estimation in generalized partial linear models for clustered data. *J. Amer. Statist. Assoc.* **100** 1176–1184. MR2236433

[13] HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

[14] HUANG, J. Z. (1998). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67** 49–71. MR1659096

[15] HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827

[16] HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

[17] HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Stat.* **34** 451–477. MR2368793

[18] LIAN, H., LIANG, H. and WANG, L. (2014). Generalized additive partial linear models for clustered data with diverging number of covariates using GEE. *Statist. Sinica* **24** 173–196.

[19] LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[20] MA, S., SONG, Q. and WANG, L. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli* **19** 252–274. MR3019494

[21] MACKE, J. H., BERENS, P., ECKER, A. S., TOLIAS, A. S. and BETHGE, M. (2009). Generating spike trains with specified correlation coefficients. *Neural Comput.* **21** 397–423. MR2477865

[22] PEPE, M. S. and ANDERSON, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simulation Comput.* **23** 939–951.

[23] QU, A., LINDSAY, B. G. and LI, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87** 823–836. MR1813977

[24] STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

[25] WANG, L., LIU, X., LIANG, H. and CARROLL, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.* **39** 1827–1851. MR2893854

[26] WANG, L. and QU, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 177–190. MR2655529

[27] WANG, L., XUE, L., QU, A. and LIANG, H. (2014). Supplement to "Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates." DOI:10.1214/13-AOS1194SUPP.

[28] WANG, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* **90** 43–52. MR1966549

[29] WELSH, A. H., LIN, X. and CARROLL, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *J. Amer. Statist. Assoc.* **97** 482–493. MR1941465

[30] WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99** 673–686. MR2090902

[31] XUE, L. (2009). Consistent variable selection in additive models. *Statist. Sinica* **19** 1281–1296. MR2536156

[32] XUE, L. and QU, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.* **13** 1973–1998. MR2956349

[33] XUE, L., QU, A. and ZHOU, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *J. Amer. Statist. Assoc.* **105** 1518–1530. MR2796568

[34] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[35] ZHOU, J. and QU, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107** 701–710. MR2980078

[36] ZHU, Z., FUNG, W. K. and HE, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95** 907–917. MR2461219

L. WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602
USA
E-MAIL: lilywang@uga.edu

L. XUE
DEPARTMENT OF STATISTICS
OREGON STATE UNIVERSITY
CORVALLIS, OREGON 97331
USA
E-MAIL: xuel@stat.oregonstate.edu

A. QU
DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
CHAMPAIGN, ILLINOIS 61820
USA
E-MAIL: anniequ@illinois.edu

H. LIANG
DEPARTMENT OF STATISTICS
GEORGE WASHINGTON UNIVERSITY
WASHINGTON, D.C. 20052
USA
E-MAIL: hliang@gwu.edu