Published in final edited form as:

Stat Med. 2018 March 15; 37(6): 1009–1030. doi:10.1002/sim.7573.

A partially linear additive model for clustered proportion data

Weihua Zhao^a, Heng Lian^b, and Dipankar Bandyopadhyay^{c,*}

^aSchool of Science, Nantong University, Nantong, P. R. China

^bDepartment of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong

^cDepartment of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Proportion data with support lying in the interval $[0,1]$ are a commonplace in various domains of medicine and public health. When this data are available as clusters, it is important to correctly incorporate the within-cluster correlation to improve the estimation efficiency while conducting regression-based risk evaluation. Furthermore, covariates may exhibit a non-linear relationship with the (proportion) responses while quantifying disease status. As an alternative to various existing classical methods for modeling proportion data (such as augmented Beta regression, etc) that utilizes maximum likelihood, or generalized estimating equations, we develop a partially linear additive model based on the quadratic inference function. Relying on quasi-likelihood estimation techniques and polynomial splines approximation for unknown nonparametric functions, we obtain the estimators for both parametric part and nonparametric part of our model, and study their large sample theoretical properties. We illustrate the advantages and usefulness of our proposition over other alternatives via extensive simulation studies, and application to a real dataset from a clinical periodontal study.

Keywords

clustered data; proportion data; quasi-likelihood; quadratic inference function; semiparametric

1. Introduction

Proportion data arises frequently in various biomedical disciplines and public health, where the response of interest (say, disease status) is usually expressed as a rate, proportion, fraction, or percentage (1; 2; 3) of the saturated (completely diseased) status. For example, consider the motivating data example derived from a clinical study (4) measuring periodontal disease (PD) status of enrolled Gullah African-American diabetics (henceforth, GAAD study), where the clinical attachment level (CAL), the most important bio-marker of PD, was measured at each of the six sites of a subject's tooth. Here, the response of interest is the proportion of diseased tooth-sites (with CAL value ≥ 3 mm) for each of the four tooth-types (incisors, canines, premolars and molars). Naturally, the responses lie in the closed interval $[0, 1]$, where 0 and 1 indicates 'completely disease-free' and 'completely diseased'

*Correspondence to: Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, One Capitol Square, 7th Floor, PO Box 980032, Richmond, VA 23298-0032, USA. Tel: +1 804 827 2058; Fax: +1 804 828 8900; dbandyop@vcu.edu.

tooth-types, respectively, while those in (0,1) indicates various level of the disease status. The underlying statistical question is cross-sectional evaluation of association of some important covariates (such as Age, Gender, etc) with these proportion responses. However, this boundedness feature of proportion response data, and the additional complexity induced due to the clustering by design (4 tooth-types per subject) renders standard regression models inapplicable.

At present, there exists a substantial body of literature on regression strategies for proportion data lying in (0, 1); for example, the logistic-normal model (5), simplex distribution based dispersion model (1), beta regression model (6) and their extensions (7; 8; 9; 10; 11; 12; 13). However, when the responses are in the interval [0, 1], i.e. the responses include the boundary values zeros and/or ones, the aforementioned methods are only appropriate after some ad hoc transformation for zeros and/or ones, such as the Lemon-squeezer (14; 15) transform. Intuitively, these ad hoc re-scaling might work for small proportions of zeros and ones, but may result into inaccurate parameter estimates and lack of reasonable interpretation for real data analysis when the proportions are large. In addition, presence of clustering (exhibited in cross-sectional, or longitudinal data) leads to an extra level of heterogeneity, and these transformations which are usually applied component-wise may not guarantee a tractable (multivariate) joint distribution (16). To alleviate this, several authors proposed to use zero- and/or one-augmented regression versions of various parametric models, such as beta, beta rectangular, simplex, etc, both under independent (17; 18) and clustered (16; 19) data settings. Although they seem to be computationally tractable, they are all based on standard parametric assumptions which are often not verifiable, and hence not generalizable.

Under a classical (frequentist) setting, two main (parametric) strategies were developed to account for the induced correlation (by design) within each subject; the beta generalized estimating equation (GEE) framework (20), and the simplex mixed-effects framework (8). Furthermore, in the mixed-effects formulation, the random effects specification would require producing inference involving complicated multivariate integrals (21). As an alternative, a Bernoulli quasi-likelihood (QL) method (22) was proposed, and extended to include non-linear (23) and multivariate (24) structures. In that vein, a QL formulation of the partially linear single-index model with a polynomial-spline approximation for the nonparametric function was recently proposed (25). The advantages of the QL formulation are two-fold; it only requires conditional moment specifications of the response, and is a robust estimation technique that avoids ad hoc transformation of the boundary values. As pointed out by a referee, there exists some related work on semi-parametric partially linear modeling of clustered data using GEE (26; 27), where the former uses kernel smoothing while the later uses spline smoothing techniques for the nonparametric function estimation. However, none of them considers modeling proportion data, although it is relatively straightforward to adapt the method of He *et al.* (27) to our case.

In this paper, we improve upon the aforementioned formulations, and propose a semi-parametric partially linear additive model (S-PLAM) for modeling correlated proportion data based on QL and quadratic inference function (QIF). Note that the QIF technique (28) is often considered a more appealing alternative to the GEE (29), and is more efficient than

the GEE (30) when the working correlation is misspecified. In tune to its predecessors, our setup avoids any underlying parametric specifications, deals with boundary values avoiding ad hoc data transformations, yet, can seamlessly integrate within subject correlation. The rest of the paper is organized as follows. In Section 2, we present the estimation approach of our S-PLAM based on QL and QIFs, where the nonparametric functions are approximated by polynomial splines. The large sample properties of the proposed estimators for both the parametric and nonparametric parts are outlined in Section 3. Next, Section 4 presents simulation studies comparing the finite sample performance of our proposition to other alternatives, while Section 5 illustrates the method via application to the GAAD dataset. Finally, Section 6 presents some concluding remarks. Detailed technical proofs of theorems relevant to Section 3, associated lemmas and additional tables from the simulation studies are relegated to the Appendix (Sections A and B).

2. Modeling framework

2.1. Partially linear additive model

Denote $(\mathbf{x}_{ij}, \mathbf{z}_{ij}, y_{ij})$ as the j th observation in the i th cluster, $i = 1, \dots, n, j = 1, \dots, m_i$, where responses $y_{ij} \in \{0, 1\}$ are associated with the general set of covariates: $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ (continuous) and $\mathbf{z}_{ij} = (z_{ij1}, \dots, z_{ijq})^T$ (discrete). We assume the (proportion) responses and covariates are related via the first conditional moment and logistic link function

$$\begin{cases} E(y_{ij} | \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \mu_{ij}, & i=1, \dots, n, j=1, \dots, m_i, \\ \eta_{ij} \triangleq \log \frac{\mu_{ij}}{1-\mu_{ij}} = \sum_{l=1}^p g_l(x_{ijl}) + \mathbf{z}_{ij}^T \boldsymbol{\beta}, \end{cases} \quad (1)$$

where $g_l(\cdot)$, $l = 1, \dots, p$ are unknown univariate functions, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T \in \mathbb{R}^q$ is a q -dimensional regression coefficient vector associated with $\mathbf{z}_{ij}$. Here, the logistic link function ensures that the predicted values of y_{ij} lie in the unit interval, and equation (1) is well-defined, even if y_{ij} take values 0 and/or 1 with positive probabilities. Certainly, one can utilize other link functions, such as the probit. Without loss of generality, we further assume that each covariate $\{x_{ijl}\}_{l=1}^p$ can be rescaled into $[0, 1]$, and each $g_l(\cdot)$ is centered with $\int_0^1 g_l(x) dx = 0$. This preserves the identifiability of the model. In addition, assuming $z_{ij1} \equiv 1$ yields a model with an intercept term. This model defined in (1) is our PLAM for clustered proportion data. Note that our current development only considers the first conditional moment, bypassing further assumptions on higher order conditional moments like conditional variance, etc. This assumption leads to a flexible and computationally scalable framework with an eye towards our motivating application dataset, and leads to robust estimates.

Next, we approximate the smooth functions $\{g_l(\cdot)\}_{l=1}^p$ via polynomial splines. For each $1 \leq l \leq p$, let $\xi_{-l,0} = \dots = 0 = \xi_{l,0} < \xi_{l,1} < \dots < \xi_{l,K_l} < 1 = \xi_{l,K_l+1} = \dots = \xi_{l,K_l+d}$ be a partition of $[0, 1]$ into subintervals $[\xi_{l,k}, \xi_{l,k+1})$, $k = 0, \dots, K_l$ with K_l interior knots. A polynomial spline of order d is a polynomial function of degree $d-1$ in each subinterval, and $d-2$ times continuously differentiable globally on $[0, 1]$. Denote the d -th order B-spline basis as $\mathbf{B}_d(x) =$

$(B_{J_l}(x), \dots, B_{\tilde{J}_l}(x))^T$, and $\tilde{J}_l = K_l + d$. Due to the centering constraint $\int g_l = 0$, $l = 1, \dots, p$, we focus on the subspace of spline functions

$$\mathcal{F}_l^0 := \left\{ \phi_l : \phi_l(x) = \sum_{k=1}^{\tilde{J}_l} B_{lk}(x) \gamma_{lk}, \int \phi_l(x) dx = 0 \right\}$$

with normalized basis

$$\left\{ \sqrt{J_l} (B_{lk}(x) - \int B_{lk}(x) dx), k=1, \dots, J_l = \tilde{J}_l - 1 \right\}.$$

With an abuse of notation, the normalized basis is still denoted by $\mathbf{B}_l(x)$. This subspace has dimension $J_l = \tilde{J}_l - 1$ due to the zero-integral constraint. Using spline approximation, each unknown function $g_l(\cdot)$ can now be approximated by a linear combination of the normalized basis, such that

$$g_l(x) \approx s_l(x) = \sum_{k=1}^{J_l} B_{lk}(x) \gamma_{lk}, \quad l=1, \dots, p.$$

In our numerical studies, the knots are placed at quantiles of the observed covariate value, and the number of knots is selected by the BIC criterion presented in Subsection 2.3. Letting

$\mathbf{x} = (x_1, \dots, x_p)^T$, $\mathbf{B}(\mathbf{x}) = (\mathbf{B}_1^T(x_1), \dots, \mathbf{B}_p^T(x_p))^T$, and $\boldsymbol{\pi}_{ij} = \mathbf{B}(\mathbf{x}_{ij})$, the polynomial approximation yields the conditional expectation μ_{ij} defined as

$$\mu_{ij} = \mu_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \exp(\boldsymbol{\pi}_{ij}^T \boldsymbol{\gamma} + \mathbf{z}_{ij}^T \boldsymbol{\beta}) / (1 + \exp(\boldsymbol{\pi}_{ij}^T \boldsymbol{\gamma} + \mathbf{z}_{ij}^T \boldsymbol{\beta})). \quad (2)$$

2.2. Estimation based on QIF

To alleviate the distributional assumptions on the response variable, we now develop a robust QL estimation procedure (22). Our Bernoulli log-likelihood function is given by

$$l_{ij}(\boldsymbol{\theta}) = y_{ij} \log(\mu_{ij}(\boldsymbol{\theta})) + (1 - y_{ij}) \log(1 - \mu_{ij}(\boldsymbol{\theta})), \quad (3)$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ leading to estimators with good theoretical properties. Ignoring clustering for the present, the QL estimator $\hat{\boldsymbol{\theta}}_{\text{QL}}$ is given by:

$$\hat{\boldsymbol{\theta}}_{\text{QL}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^n \sum_{j=1}^{m_i} l_{ij}(\boldsymbol{\theta}).$$

The corresponding QL estimating equation is $\Psi(\hat{\theta}_{QL}) = 0$, where

$$\Psi(\theta) = \sum_{i=1}^n \mathbf{U}_i^T (\mathbf{y}_i - \boldsymbol{\mu}_i(\theta))$$

with $\mathbf{U}_i = (\boldsymbol{\pi}_i^T \mathbf{z}_i^T)^T$, $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{im_p})^T$, $\mathbf{z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{im_p})^T$, $\mathbf{y}_i = (y_{i1}, \dots, y_{im_p})^T$, $\boldsymbol{\mu}_i(\theta) = (\mu_{i1}, \dots, \mu_{im_p})^T$ with μ_{ij} defined in (2). Now, following the GEE (30) that accounts for the within-cluster correlation, a more efficient estimating equation takes the form

$$\sum_{i=1}^n \mathbf{U}_i^T \mathbf{A}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\theta)) = 0, \quad (4)$$

where $\mathbf{A}_i$ is the correlation matrix of $\mathbf{y}_i - \boldsymbol{\mu}_i(\theta)$, which may depend on some nuisance parameters. However, it may be difficult to estimate these nuisance parameters. To circumvent this problem, we use the QIF approach by approximating $\mathbf{A}_i^{-1}$ with a linear combination of basis matrices as $\mathbf{A}_i^{-1} = \sum_{k=1}^v a_k \mathbf{M}_{ki}$, where $\mathbf{M}_{ki}$'s are known symmetric matrices and $a_1, \dots, a_v$ are unknown constants. As stated in (28), the most commonly used correlation structures, such as compound symmetric and AR(1) correlation matrices, can be represented or approximated by several simple basis matrices. The QIF approach can bypass explicit estimation of the correlation parameters, thus improving estimation efficiency significantly with simultaneous saving of computational cost.

Define the QIF estimating equations

$$\mathbf{S}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_i(\theta), \quad (5)$$

where

$$\mathbf{S}_i(\theta) = \begin{pmatrix} \mathbf{U}_i^T \mathbf{M}_{1i} (\mathbf{y}_i - \boldsymbol{\mu}_i(\theta)) \\ \vdots \\ \mathbf{U}_i^T \mathbf{M}_{vi} (\mathbf{y}_i - \boldsymbol{\mu}_i(\theta)) \end{pmatrix}.$$

Since there are more estimation equations than the number of unknown parameters in (5), θ can be estimated as the minimizer of the following quadratic function

$$\hat{\theta} = \operatorname{argmin}_{\theta} \mathbf{Q}_n(\theta), \quad (6)$$

where

$$\mathbf{Q}_n(\boldsymbol{\theta}) = n\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_n^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta}) \quad \text{and} \quad \mathbf{C}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_i(\boldsymbol{\theta})\mathbf{S}_i^T(\boldsymbol{\theta}).$$

The estimators of the unknown functions and parametric part are defined respectively as $\hat{g}_l(u) = \mathbf{B}_l^T(u)\hat{\boldsymbol{\gamma}}_l$, $l = 1, \dots, p$ and $\hat{\boldsymbol{\beta}}$ respectively. Estimation can proceed using the Newton-Raphson iterative procedure, with the initial values obtained via the QL method, ignoring correlations. The standard error estimates of the parameters are obtained from the expression of the asymptotic covariance, given in Theorem 2. In practice, this iterative procedure unlike the GEE avoids estimation of the correlation matrix and computing its inverse at each iteration, leading to quick convergence.

2.3. Knots selection

To implement our proposed QIF procedure, one needs to select the order and the number of knots in the spline approximation. In all simulation studies and real data analysis, for simplicity, the nonparametric functions are approximated by cubic splines and the same number of interior knots, i.e. $K := K_1 = \dots = K_p$. The number of interior knots is chosen from the interval $[n^{1/(2d+1)}, 2n^{1/(2d+1)}]$ by minimizing the following BIC criterion $\text{BIC}(K) = \mathbf{Q}_n(\hat{\boldsymbol{\theta}}) + p(K+d)\log(n)$. The above proposed range for K satisfies the order assumptions in Theorem 1 (see Section 3) on the number of interior knots K , and our simulation studies and real data analysis demonstrate that this knots selection strategy is adequate.

3. Theoretical properties

In this section, we study the rate of convergence for $\hat{g}(x)$ and $\hat{\boldsymbol{\beta}}$ as the number of subject n increases. Throughout this paper, denote $\|f\|_\infty = \sup_{x \in [0,1]} |f(x)|$ for any real-valued function f on $[0,1]$. Let $\|\cdot\|_2$ be the L_2 norm for functions and l_2 norm for vectors. For any matrix $\mathbf{A}$, define $\|\mathbf{A}\| = (\text{Tr}(\mathbf{A}\mathbf{A}^T))^{1/2}$ as its Frobenius norm. Also, denote $\|\cdot\|_n$ as the Euclidean norm of a vector.

Let $\mathcal{H}_r$ be the collection of all functions on the support $[0, 1]$ whose s -th order derivative satisfies the Hölder condition of the order ϑ with $r = s + \vartheta$. Then, for each $h \in \mathcal{H}_r$, there exists a positive constant C_0 such that $|h^{(s)}(x_1) - h^{(s)}(x_2)| \leq C_0|x_1 - x_2|^\vartheta$, $\forall x_1, x_2 \in [0, 1]$. According to the results in De Boor (31) (see page 149), there exists a constant C such that

$$\sup_{x \in [0,1]} |g_l(x) - \mathbf{B}_l^T(x)\boldsymbol{\gamma}_l^0| \leq C J_l^{-r}, \quad (7)$$

$g_l \in \mathcal{H}_r$ where $\boldsymbol{\gamma}_l^0 = (\gamma_{l1}, \dots, \gamma_{lj_l})^T$ can be viewed as the best approximation coefficient vector for $g_l(x)$. The following assumptions are needed to begin with.

(C1) The cluster sizes m_i are uniformly bounded for all $i = 1, \dots, n$.

- (C2) $g(\cdot) \in \mathcal{H}_r$, $l = 1, \dots, p$, for some $r > 1/2$.
- (C3) The knots sequences ξ_{l,K_l} for $l = 1, \dots, p$, are quasi-uniform. That is, there exists $c > 0$, such that $\max_l (\max_k h_{l,k} / \min_k h_{l,k}) \leq c$, where $h_{l,k} = \xi_{l,k+1} - \xi_{l,k}$ ($0 \leq k < K_l$) are the distances between neighboring knots. Furthermore, the number of interior knots $K_l \asymp n^{1/(2r+1)}$, where $\asymp$ means that both sides have the same order.
- (C4) There exist two positive constants c_1 and c_2 such that

$$c_1 \leq \lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{U}_i^T \mathbf{U}_i \right) \leq \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{U}_i^T \mathbf{U}_i \right) \leq c_2,$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denotes the smallest and largest eigenvalues of a matrix.

- (C5) The eigenvalues for each $\mathbf{M}_{ki}$ are bounded away from 0 and ∞ .

These conditions are common in the literature (32; 33; 34; 35; 36). Condition (C1) is a standard assumption for clustered data. Condition (C2) is a smoothness assumption on the nonparametric functions commonly used in additive models. Condition (C3) is a mild assumption on the choice of knots. The assumption in (C4) implies the eigenvalues of

$\frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i^T \mathbf{Z}_i$ and $\frac{1}{n} \sum_{i=1}^n \mathbf{\Pi}_i^T \mathbf{\Pi}_i$ are bounded away from zero and infinity. Finally, Condition (C5) is satisfied for commonly used basis matrices.

To explore consistency of the resulting estimator, we first list some notations. Let

$\boldsymbol{\mu}_i^0 = (\mu_{i1}^0, \dots, \mu_{im_i}^0)^T$ with $\mu_{ij}^0 = \exp(\eta_{ij}^0) / (1 + \exp(\eta_{ij}^0))$, $\eta_{ij}^0 = \sum_{l=1}^p g_{0l}(x_{ijl}) + \mathbf{z}_{ij}^T \boldsymbol{\beta}_0$, where g_{0l} and $\boldsymbol{\beta}_0$ denote the true values of nonparametric function g_l and parametric vector $\boldsymbol{\beta}$. Let

$\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T)^T$, where $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_1^{0T}, \dots, \boldsymbol{\gamma}_p^{0T})^T$ and $\boldsymbol{\gamma}_l^0$ defined by (7) in Subsection 2.1. Denote $J \asymp J_l$, $l = 1, \dots, p$, where $\asymp$ represents both sides having the same order. Then, we have following theorem.

Theorem 1—Under the Conditions (C1)–(C5), there exists a local minimizer of (6) such that

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p(n^{-r/(2r+1)}).$$

As an immediate implication of Theorem 1, we have

$$\sum_{l=1}^p \|\hat{g}_l - g_{0l}\|_2^2 + \sum_{j=1}^q |\hat{\beta}_j - \beta_{0j}|^2 = O_p(n^{-2r/(2r+1)}). \quad (8)$$

Remark 1—Note that the rate of convergence for the nonparametric function is $\|\hat{g}_l - g_{0l}\|_2 = O_p(n^{-r/(2r+1)})$, which is the same as that in (33) for mean regression, and in (32) for

independent data ($m_i \equiv 1$). The nonparametric convergence rate here has been extended to the case of clustered proportion data for the PLAM-QIF approach. In practice, the main advantage of the proposed QIF approach is that it incorporates the within-cluster correlation by optimally combining estimating equations, without the need to estimate the correlation parameters.

The parametric part can be shown to be asymptotically normal under slightly stronger conditions. With some notations Ξ , $\mathcal{U}$ and $\mathcal{G}_\beta$ given in the Appendix after Lemma 6, we have the following result.

Theorem 2—Under the Conditions (C1)–(C5), we have

$$(\mathcal{G}_\beta^T \sum^{-1} \mathcal{G}_\beta)^{1/2} \sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_q),$$

$$\text{where } \sum = \text{Cov} \left(\frac{1}{\sqrt{n}} (\mathbf{I}_v \otimes \mathcal{U}^T) \Xi (\mathbf{y} - \mu^0) \right).$$

The proofs of these Theorems and the associated Lemmas are presented in the Appendix, Sections A and B, where we use the projection method to prove the asymptotic normality of the parametric part.

4. Simulation studies

In this section, we conduct some simulation studies to study the finite sample performance of our proposed method, and compare it with other available alternatives.

4.1. Simulation 1

Here, we consider the following model for the marginal expectation of our proportion response $y_{ij} \in (0, 1)$ as

$$\log \frac{\mu_{ij}}{1 - \mu_{ij}} = g_1(x_{ij1}) + g_2(x_{ij2}) + \mathbf{z}_{ij}^T \boldsymbol{\beta}, \quad i=1, \dots, n, \quad j=1, \dots, 5,$$

where $g_1(u) = 2\sin(2\pi u)$ and $g_2(u) = 8u(1-u) - 4/3$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T = (1, -1, 0.8)^T$. The covariates x_{ij1} and x_{ij2} are independently generated from uniform distribution on the interval $[0, 1]$, and $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, z_{ij3})^T$, where (z_{ij1}, z_{ij2}) has a bivariate normal distribution with mean zero, marginal variance 1 and correlation coefficient 0.5, z_{ij3} is independent of (z_{ij1}, z_{ij2}) and generated from a Bernoulli($p = 0.6$) distribution. Then, the proportion response vector $\mathbf{y}_i = (y_{i1}, \dots, y_{i5})^T$ for each subject has the above marginal mean $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{i5})^T$, with either an exchangeable (also compound symmetry, CS), or the first-order autoregressive (AR(1)) correlation structure with the correlation coefficient ρ . The sample size is set to be $n = 100, 200$ and 300 , and we consider the following two cases to obtain the proportion response y_{ij} :

Case 1. $y_{ij} = y_{ij}^* / 100$, where $y_{ij}^* | \mu_{ij} \sim \text{binomial}(100, \mu_{ij})$.

Case 2. $y_{ij} / \mu_{ij} \sim \text{beta}(\mu_{ij}\phi, (1 - \mu_{ij})\phi)$, where the dispersion parameter $\phi = 10$ in the beta distribution.

Note that for Case 1, y_{ij} can take value at 0 and 1, while it lies in the open interval (0, 1) in Case 2. We evaluate the effects of incorporating correlation structures, such as the CS, AR(1) and working independence (WI) on the estimation efficiency for each case using the integrated mean squared error (IMSE), defined as

$$\text{IMSE}(g_l) = \frac{1}{1000} \sum_{s=1}^{1000} \sqrt{\frac{1}{n_{\text{grid}}} \sum_{k=1}^{n_{\text{grid}}} \{\hat{g}_l^{(s)}(u_k) - g_l(u_k)\}^2}$$
, where $\{u_k : k = 1, \dots, n_{\text{grid}} = 101\}$ on the interval [0,1] are the grid points at which the functions $\{g(\cdot)\}$ are evaluated, and $\hat{g}_l^{(s)}(u)$ is the spline approximation to $g_l(u)$ in the s th simulation run. We finally report the

average of the integrated mean squared error as $\text{AIMSE} = \frac{1}{p} \sum_{l=1}^p \text{IMSE}(g_l)$. For the parametric part, we report the average bias, absolute bias and standard error estimate over 1000 replications for each parameter. The simulation results are summarized in Tables 1–2 and Tables C1–C4 in the Appendix.

Our numerical experiments support several conclusions. First, biases are close to zero for all settings, and the absolute bias of the parametric estimate and AIMSE for the nonparametric part becomes smaller with increasing sample size. Second, in terms of both the ABIAS and AIMSE, the gains for incorporating the working correlation are significant, compared to the WI assumption that ignore within-subject correlation. In particular, when the true correlation structure is CS, Tables 1 and 2 show that the estimators under the correct CS setting have the smallest ABIAS and AIMSE. Similar findings are also observed in the other tables (Tables C1–C4 in the Appendix). In addition, the efficiency gain become more obvious as the true correlation increases. Third, the differences between ASE and ESE are very small, which indicates that our proposed asymptotic error formula for the parametric part works well.

4.2. Simulation 2

A reviewer suggested us to illustrate the advantages of our proposed QIF approach over the GEE method of He *et al.* (27). For a fair comparison, we modify their semiparametric partially linear GEE approach to a semiparametric partially linear *additive* model for proportion responses, and apply that to the following example where the simulated proportion responses were generated from the logistic-normal distribution (5). Specifically,

$$y_{ij} = \exp(y_{ij}^*) / (1 + \exp(y_{ij}^*)), \quad y_{ij}^* = g_1(x_{ij1}) + g_2(x_{ij2}) + \mathbf{z}_{ij}^T \boldsymbol{\beta} + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, 5,$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{i5})^T$ is distributed as a zero mean multivariate normal distribution with marginal standard deviation 0.2. Following Simulation 1, we consider the two (true) correlation structures CS and AR(1) for $\boldsymbol{\varepsilon}_i$. Other settings remain as in Simulation 1. Note that the GEE approach requires assumption of the (conditional) mean and variance of the response, whereas, our quasi-likelihood based QIF approach only requires specification of the conditional mean of the response (1). Here, we assume an unstructured working

correlation matrix for the GEE approach. The simulation results are listed in Table 3–4, where RMSE denotes the root mean squared error estimate for the linear part, defined as

$RMSE = \sqrt{(\hat{\beta} - \beta)^T (\hat{\beta} - \beta) / q}$, and the AIMSE for the non-linear part is as defined in Simulation 1.

Tables 3 and 4 reveal that both the QIF (with CS and AR(1) as working correlation assumptions) outperform the WI for each scenario. Furthermore, the QIF method is better than the GEE procedure of He *et al.* (27), even under misspecification of the correlation structure in the QIF for almost all scenarios, except when the true correlation is CS, $\rho = 0.5$ and $n = 100$.

5. Application: GAAD Data

In this section, we apply our S-PLAM model to the motivating GAAD data introduced in Section 1. The GAAD study on assessing PD status of 290 study subjects was conducted at the Medical University of South Carolina (MUSC), where the primary research question was to evaluate the cross-sectional association between the proportion CAL responses, and important continuous and categorical risk factors. Note that missing teeth in our data are considered ‘missing due to past incidence of PD’, where all sites for that tooth contributed to the (completely) diseased category. Among the various subject-level covariates of interest are the continuous ‘Age’ at examination (in years, ranging from 26 to 87 years), and discrete variables such as Gender (0-male, 1-female), glycosylated hemoglobin (HbA1c) status indicator (0-controlled, < 7%; 1-uncontrolled, ≥ 7%) and Smoking status (0-nonsmoker, 1-smoker). We also considered a tooth-level variable representing each of the four tooth-types (1 = Molar, 2 = Premolar, 3 = Canine, and 4 = Incisor), with ‘Incisor’ as the baseline. The responses are continuous on [0, 1], with non-negligible proportions of 0 (9.8%) and 1 (8.1%), where 0 and 1 represent ‘completely disease-free’ and ‘highly diseased’ cases, respectively. The histogram of aggregated proportion responses and the violin plot of these responses for the four tooth type are shown in Figure 1. Note, the distribution of the responses differ significantly across the tooth-types, as revealed from the violin plots. In a regression context, PLAMs (32; 33; 34) can flexibly incorporate both discrete, and continuous covariates that may exhibit a non-linear relation with the response. For example, in the context of our motivating data example, previous research (37; 38; 39) have considered age to be linearly related to PD progression, although there is not much scientific basis to that. A recent US National Institutes of Health October 2010 Fact Sheet on PD (available here: <https://report.nih.gov/NIHfactsheets/ViewFactSheet.aspx?csid=111&key=P#P>) states that PD progression is multifactorial, and may no longer be viewed as an inevitable result of aging as the susceptibility to PD can vary with subjects. Hence, quantifying the age progression curve for our data subjects is of essence here.

Initial model fitting is conducted using the QL linear model based on the QIF method, given as:

$$E y = \mu = \exp(\eta) / (1 + \exp(\eta)),$$

$$\eta = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{HbA1c} + \beta_4 \text{Smoker} + \beta_5 \text{Molar} + \beta_6 \text{Permolar} + \beta_7 \text{Canine} \quad (9)$$

Furthermore, for selecting the appropriate within-subject correlation structure, we use the value of the trace of the asymptotic covariance matrices (TACM), computed under different (correlation) conditions. A lower value of this trace reflects a better model fit. This selection criterion comes from the TGI (the trace of the Godambe Information) criterion proposed in (29). The resulting parameter estimates, their standard errors, associated p-values, and the TACM from the three correlation structures are presented in Table 5. The covariates Gender, Age, and the Tooth-type (molar, premolar, and canines) significantly explain the proportion responses. Conditional on the set of other covariates, parameter interpretation can be expressed in terms of the corresponding covariate effect directly on μ_{ij} , in particular the ratio $\frac{\mu_{ij}}{1 - \mu_{ij}}$. Here, μ_{ij} is the expected proportion of diseased sites, and $1 - \mu_{ij}$ is the complement, i.e., the expected remaining proportion to complete disease. Hence, exponentiated parameter estimates from Table 5 can be expressed as the number of times this ratio is higher/lower with every unit increase for a continuous covariate (such as Age), or a change in category, say from 0 to 1 for a discrete covariate (say, Gender). For example, under the WI assumption, this ratio for Age (often considered a strong predictor for PD) is $\exp(0.0372) = 1.037$, while, for Gender, we conclude that this ratio is 42% ($1 - \exp(-0.5514)$) lower for females as compared to males, implying females experiencing better periodontal health than males. Note that, in this clinical study, women participated at a much higher rate than men – a feature not unusual for studies on this population (40). Regarding tooth-types, with incisor as the comparison group, this ratio is 6.24 times higher for the molars. Intuitively, the rate of PD is usually higher among the posteriorly located molars, and the Gullah subjects in this study are no exception. Note that the parameter estimates and standard errors are very similar across the 3 competing choices of the correlation structure. However, the TACM marginally favors the AR(1) over the other two.

Next, to explore the nonlinear effect of the covariate ‘Age’ on the proportion response, we refit the data using a semiparametric AR(1) model, with η given as:

$$\eta = g(\text{Age}) + \beta_1 \text{Gender} + \beta_2 \text{HbA1c} + \beta_3 \text{Smoker} + \beta_4 \text{Molar} + \beta_5 \text{Permolar} + \beta_6 \text{Canine}$$

The estimated curve has been shown in Figure 2, where the number of knots is determined using the BIC knots selection approach described in Subsection 2.3. The plot clearly reveals that although Age is found to be positively related to PD incidence from our cross-sectional evaluation (revealed from Table 5), the relationship pattern is non-linear, with a distinct jump between the ages 30–40, while maintaining the increasing trend overall.

To further evaluate the gain in fitting this semiparametric model over the (parametric) model (9) linear in η , we use the following QIF-based criterion, defined as

$$Q_n(\hat{\theta}_M) + \lambda \times |M|,$$

where $|M|$ is the cardinality of the fitted model. Note that the value $\lambda = 2$ and $\log(n)$ corresponds to QAIC and QBIC, respectively. The results of this comparison is presented in Table 6. In addition, to evaluate model fit and prediction, we also reported the mean absolute fitting error (MAFE) and mean absolute prediction error (MAPE) in the table, computed using cross-validation techniques via randomly splitting the data into 200 and 90 subjects as training and testing samples, respectively, and then averaging over the random partitions. For comparison, we also report the MAFEs and MAPEs obtained from fitting the GEE method proposed in Simulation 2, and the ML-based Beta regression model with a random intercept (21), where the the Lemon-Squeezer transformation (15) is used to tackle the boundary values. From Table 6, we observe that QAIC considers the semiparametric model to be significantly better than the parametric model, whereas QBIC (which tends to penalize model complexity more heavily than QAIC) marginally favors the linear model, although the gain cannot be deemed substantial. Also, both MAFE and MAPE are uniformly lower for the semiparametric model compared to the parametric model across all models, though not substantially. Interestingly, the performances of the QIF and GEE are very similar in terms of MAFE and MAPE, and slightly better than the Beta regression model.

Finally, the parameter estimates for the parametric part of the semiparametric model is given in Table 7. Although the signs of the estimates and the list of significant covariates remained the same compared to the estimates from the AR(1) model in Table 5, the effects (as revealed by the exponentiated coefficients) are slightly different. For example, for Gender, the ratio $\frac{\mu_{ij}}{1-\mu_{ij}}$ is 36.4% (for the parametric AR(1) model) vs 38% (for the semiparametric model) lower for females, compared to males.

6. Conclusion

In this paper, we developed a QIF-based inferential framework for a semiparametric partially linear model, with responses distributed in $[0,1]$. Our QIF QL proposal models the expected (conditional) mean μ_{ij} , and can easily accommodate both 0 and 1 responses. Clinical interpretation in terms of the ratio of expected proportion of diseased sites and its complement are also available. Application of our model to the GAAD dataset demonstrated the improvement in model fit by considering the polynomial spline modeling of Age (in comparison to a linear modeling), and thus improve upon previous findings that considered linearity of Age (39). The methodology proposed is also computationally tractable, and the associated R codes are available upon request.

In our data analysis, we first used the TACM to evaluate the within-subject correlations, followed by QAIC/QBIC (computed using the penalized QIF function) to evaluate the GOF of the competing models with the selected correlation structure. Although AIC/BIC are popularly used in the literature for model selection and evaluating GOFs, they are not applicable here because the likelihood function is not fully specified. In addition, both AIC and BIC are defined under the independence model, and their adaptations to clustered data is

complicated. Additional simulations (not presented here) reveal that the performance of the QAIC is a slightly better than QBIC in terms of goodness-fit, while it is reverse for model, or variable selection.

Although other parametric alternatives exist, such as the class of zero-one-augmented beta (19) and simplex (16) models, their applications are limited. Our formulation is not restricted to these parametric assumptions. Also, this class of augmented models do not consider simultaneously modeling all the responses in $[0,1]$; instead, they mostly consider separate models for the probabilities of 0, and/or 1, and responses in $(0,1)$. That way, they model either the mean of the standard Beta regression for the responses in $(0,1)$ with the probabilities of 0 and 1 unspecified (estimated from the data), or three separate regressions that may, or may not, share random effects (for clustered data). On the contrary, our proposition using QIFs are much more appealing, and provides a direct modeling framework.

Note that our current framework only considers estimation and inference, and develops the corresponding large sample theory. To evaluate the effect of individual observations on the estimation of μ_{ij} , detailed exploration of influence diagnostics (such as, in terms of Kullback-Leibler distance, etc) are necessary. Furthermore, our nonparametric proposition based on spline approximations can be further generalized to accommodate unknown link functions for a relatively high-dimensional covariate space (such as, considering plausible interactions of model covariates with tooth-types, etc) via single-index models (41). These are important avenues for future work, and will be considered elsewhere.

Acknowledgments

We thank the anonymous Associate Editor and two reviewers, whose constructive criticism led to a significantly improved version of the manuscript. We also thank the Center for Oral Health Research at MUSC for providing the motivating data and the context of this work. This research was partially supported by National Social Science Fund of China (15BTJ027), Natural Science Fund of Nantong University (14B28), and also by grants R03DE023372 and R01DE024984 from the United States National Institutes of Health.

References

1. Song P XK, Tan M. Marginal models for longitudinal continuous proportional data. *Biometrics*. 2000; 56(2):496–502. [PubMed: 10877309]
2. Kieschnick R, McCullough BD. Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions. *Statistical Modelling*. 2003; 3(3):193–213.
3. Branscum A, Johnson W, Thurmond M. Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*. 2007; 49(3):287–301.
4. Fernandes JK, Wiegand RE, Salinas CF, Grossi SG, Sanders JJ, Lopes-Virella MF, Slate EH. Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of Periodontology*. 2009; 80(7):1062–1068. [PubMed: 19563285]
5. Lesaffre E, Rizopoulos D, Tsonaka R. The logistic transform for bounded outcomes scores. *Biostatistics*. 2007; 8(1):72– 85. [PubMed: 16597671]
6. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*. 2004; 31(7):799–815.
7. Song XK, Qiu Z, Tan M. Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*. 2004; 46(5):540–553.

8. Qiu Z, SONG PXX, Tan M. Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*. 2008; 35(4):577–596.
9. Espinheira PL, Ferrari SLP, Cribari-Neto F. Influence diagnostics in beta regression. *Computational Statistics & Data Analysis*. 2008; 52(9):4417–4431.
10. Rocha AV, Simas AB. Influence diagnostics in a general class of beta regression models. *Test*. 2011; 20(1):95–119.
11. Pereira GH, Botter DA, Sandoval MC. A regression model for special proportions. *Statistical Modelling*. 2013; 13(2):125–151.
12. Zhang P, Qiu Z. Regression analysis of proportional data using simplex distribution. *Science China Mathematics (in Chinese)*. 2014; 44(1):89–104.
13. Zhao W, Zhang R, Lv Y, Liu J. Variable selection for varying dispersion beta regression model. *Journal of Applied Statistics*. 2014; 41(1):95–108.
14. Ramalho EA, Ramalho JJS, Murteira JMR. Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*. 2011; 25(1):19–68.
15. Smithson M, Verkuilen J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*. 2006; 11(1):54. [PubMed: 16594767]
16. Bandyopadhyay D, Galvis DM, Lachos VH. Augmented mixed models for clustered proportion data. *Statistical Methods in Medical Research*. 2017; doi: 10.1177/0962280214561093
17. Cook DO, Kieschnick R, McCullough BD. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*. 2008; 15(5):860–867.
18. Ospina R, Ferrari SLP. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*. 2012; 56(6):1609–1623.
19. Galvis DM, Bandyopadhyay D, Lachos VH. Augmented mixed beta regression models for periodontal proportion data. *Statistics in Medicine*. 2014; 33(21):3759–3771. [PubMed: 24764045]
20. Hunger M, Döring A, Holle R. Longitudinal beta regression models for analyzing health-related quality of life scores over time. *BMC Medical Research Methodology*. 2011; 12(1):527–536.
21. Bonat W, Ribeiro P, Zeviani W. Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics*. 2015; 42(2):252–266.
22. Papke LE, Wooldridge JM. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*. 1996; 11(6):619–632.
23. Ramalho EA, Ramalho JJS. Moment-based estimation of nonlinear regression models with boundary outcomes and endogeneity, with applications to nonnegative and fractional responses. *Econometric Reviews*. 2014; doi: 10.1080/07474938.2014.976531
24. Murteira JMR, Ramalho JJS. Regression analysis of multivariate fractional data. *Econometric Reviews*. 2016; 35(4):515–552.
25. Zhao W, Lian H, Zhang R, Lai P. Estimation and variable selection for proportional response data with partially linear single-index models. *Computational Statistics & Data Analysis*. 2016; 96:40–56.
26. Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*. 2001; 96(455):1045–1056.
27. He X, Fung WK, Zhu Z. Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*. 2005; 100(472):1176–1184.
28. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika*. 2000; 87(4):823–836.
29. Song PXX, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine*. 2009; 28(29):3683–3696. [PubMed: 19757486]
30. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73(1):13–22.
31. De Boor, C. A Practical Guide to Splines, Applied Mathematical Sciences. 2. Vol. 27. Springer-Verlag; 2001.
32. Lian H. Semiparametric estimation of additive quantile regression models by two-fold penalty. *Journal of Business & Economic Statistics*. 2012; 30(3):337–350.

33. Lian H, Liang H, Wang L. Generalized additive partial linear models for clustered data with diverging number of covariates using GEE. *Statistica Sinica*. 2014; 24(1):173–196.
34. Wang L, Xue L, Qu A, Liang H, et al. Estimation and model selection in generalized additive partial linear models for correlated data with diverging number of covariates. *The Annals of Statistics*. 2014; 42(2):592–624.
35. Xue L, Qu A, Zhou J. Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*. 2010; 105(492):1518–1530.
36. Wang HJ, Zhu Z, Zhou J. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*. 2009; 37(6):3841–3866.
37. El-Qaderi S, Quteish Ta'ani D. Assessment of periodontal knowledge and periodontal status of an adult population in Jordan. *International Journal of Dental Hygiene*. 2004; 2(3):132–136. [PubMed: 16451477]
38. Chung SY, Song KB, Lee SG, Choi YH. The strength of age effect on tooth loss and periodontal condition in Korean elderly. *Archives of Gerontology and Geriatrics*. 2011; 53(2):243–248.
39. Busby M, Martin J, Matthews R, Burke F, Chapple I. The relationship between oral health risk and disease status and age, and the significance for general dental practice funding by capitation. *British Dental Journal*. 2014; 217(10):1–5. [PubMed: 25012305]
40. Johnson-Spruill I, Hammond P, Davis B, McGee Z, Loudon D. Health of Gullah families in South Carolina with Type 2 diabetes diabetes self-management analysis from Project Sugar. *The Diabetes Educator*. 2009; 35(1):117–123. [PubMed: 19244567]
41. Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association*. 1997; 92(438):477–489.
42. He X, Shi P. Bivariate tensor-product B-splines in a partly linear model. *Journal of Multivariate Analysis*. 1996; 58(2):162–181.

APPENDIX A: Associated Lemmas

We start with some notations. Denote $\mathbf{S}_k(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik}(\boldsymbol{\theta})$, where $\mathbf{S}_{ik}(\boldsymbol{\theta}) = \mathbf{U}_i^T \mathbf{M}_{ki}(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}))$, decomposed as

$$\mathbf{S}_{ik}(\boldsymbol{\theta}) = \mathbf{S}_{ik1} + \mathbf{S}_{ik2} + \mathbf{S}_{ik3}(\boldsymbol{\theta}), \quad k=1, \dots, v,$$

where $\mathbf{S}_{ik1} = \mathbf{U}_i^T \mathbf{M}_{ki}(\mathbf{y}_i - \boldsymbol{\mu}_i^0)$, $\mathbf{S}_{ik2} = \mathbf{U}_i^T \mathbf{M}_{ki}(\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0))$ and $\mathbf{S}_{ik3} = \mathbf{U}_i^T \mathbf{M}_{ki}(\boldsymbol{\mu}_i(\boldsymbol{\theta}_0) - \boldsymbol{\mu}_i(\boldsymbol{\theta}))$. Next, we denote

$$\mathbf{S}_{k1} = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik1}, \quad \mathbf{S}_{k2} = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik2} \quad \text{and} \quad \mathbf{S}_{k3}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik3}(\boldsymbol{\theta}), \quad (\text{A-1})$$

and define $\mathbf{C}_{kk'}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik}(\boldsymbol{\theta}) \mathbf{S}_{ik'}^T(\boldsymbol{\theta})$ and $\tilde{\mathbf{C}}_{kk'} = \frac{1}{n} \sum_{i=1}^n \mathbf{S}_{ik1} \mathbf{S}_{ik'1}^T$ as the (k, k') -block of $\mathbf{C}_n(\boldsymbol{\theta}) = \{\mathbf{C}_{kk'}(\boldsymbol{\theta})\}_{k,k'=1}^v$ and $\tilde{\mathbf{C}}_n = \{\tilde{\mathbf{C}}_{kk'}\}_{k,k'=1}^v$, respectively. Denote

$$\tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) = n \mathbf{S}^T(\boldsymbol{\theta}) \tilde{\mathbf{C}}_n^{-1} \mathbf{S}(\boldsymbol{\theta}), \quad (\text{A-2})$$

as an approximation of $\mathbf{Q}_n(\boldsymbol{\theta})$ defined in (6). In the following, we use c and C to denote generic positive constants. The following lemmas are necessary to prove the theorems.

Lemma 1

Assume the conditions (C1)–(C4) hold. There exist two constants $0 < c_1 < c_2$, such that, excluding an event whose probability tends to zero as $n \rightarrow \infty$,

$$c_1 \|\boldsymbol{\gamma}\|_2^2 \leq \|\mathbf{B}^T \boldsymbol{\gamma}\|_2^2 \leq c_2 \|\boldsymbol{\gamma}\|_2^2$$

for any vector $\boldsymbol{\gamma}$ with length $\sum_{l=1}^p J_l$. Furthermore, there also exists two constants $0 < c'_1 \leq c'_2$, such that

$$c'_1 \|\boldsymbol{\gamma}\|_2^2 \leq \|\boldsymbol{\Pi}^T \boldsymbol{\gamma}\|_n^2 \leq c'_2 \|\boldsymbol{\gamma}\|_2^2,$$

excluding an event whose probability tends to 0 as $n \rightarrow \infty$.

Proof

The results can be proved along the line of Lemmas 3 and 6 in (35).

Next, define the set $\Theta_n(C) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq C \sqrt{J/n}\}$, for a sufficiently large constant C .

Lemma 2

Assume the Conditions (C1)–(C5) hold. Then, we have

$$\|\mathbf{S}_{k1}\|_2 = O_p(1/\sqrt{n}), \quad \|\mathbf{S}_{k2}\|_2 = O_p(J^{-r}).$$

Furthermore, there exist two constants $0 < c_1 < c_2$ such that for any $\boldsymbol{\theta} \in \Theta_n(C)$,

$$c_1 C \sqrt{J/n} \leq \|\mathbf{S}_{k3}(\boldsymbol{\theta})\|_2 \leq c_2 C \sqrt{J/n}$$

excluding an event whose probability goes to 0 as $n \rightarrow \infty$, where $\mathbf{S}_{k1}$, $\mathbf{S}_{k2}$ and $\mathbf{S}_{k3}$ are defined in (A-1).

Proof

By the definition of $\mathbf{S}_{k1}$, for any $\boldsymbol{\nu} \in \mathbb{R}^{\sum_{l=1}^p J_l + q}$ with $\boldsymbol{\nu}^T \boldsymbol{\nu} = 1$, we have

$$E(\boldsymbol{\nu}^T \mathbf{S}_{k1})=0 \text{ and } \text{Var}(\boldsymbol{\nu}^T \mathbf{S}_{k1})=\frac{1}{n}E(\boldsymbol{\nu}^T \mathbf{S}_{ik1} \mathbf{S}_{ik1}^T \boldsymbol{\nu}).$$

Note that

$$\frac{1}{n}E(\boldsymbol{\nu}^T \mathbf{S}_{ik1} \mathbf{S}_{ik1}^T \boldsymbol{\nu})=\frac{1}{n}E(\boldsymbol{\nu}^T \mathbf{U}_i^T \mathbf{M}_{ki} \boldsymbol{\varepsilon}_i^0 \boldsymbol{\varepsilon}_i^{0T} \mathbf{M}_{ki}^T \mathbf{U}_i \boldsymbol{\nu}),$$

where $\boldsymbol{\varepsilon}_i^0 = \mathbf{y}_i - \boldsymbol{\mu}_i^0$. By the definition of $\boldsymbol{\varepsilon}_i^0$ and Conditions (C3), (C4) and (C5) together with Lemma 1, we have

$$\frac{1}{n}E(\boldsymbol{\nu}^T \mathbf{S}_{ik1} \mathbf{S}_{ik1}^T \boldsymbol{\nu})=O(1/n).$$

Therefore, we obtain $\|\mathbf{S}_{k1}\|_2 = \sup_{\boldsymbol{\nu}^T \boldsymbol{\nu}=1} \boldsymbol{\nu}^T \mathbf{S}_{k1} = O_p(1/\sqrt{n})$.

We next consider $\mathbf{S}_{k2}$. Denote $h(u) = \exp(u)/(1 + \exp(u))$. By the Taylor's expansion, we have

$$h(\eta_{ij}^0) = h(\mathbf{U}_{ij}^T \boldsymbol{\theta}_0) + h'(\zeta)(\eta_{ij}^0 - \mathbf{U}_{ij}^T \boldsymbol{\theta}_0),$$

where ζ lies between η_{ij}^0 and $\mathbf{U}_{ij}^T \boldsymbol{\theta}_0$. Since $h'(u) = h(u)(1 - h(u))$, therefore $\max |h'(\zeta)| \leq 1$, and together with $|\eta_{ij}^0 - \mathbf{U}_{ij}^T \boldsymbol{\theta}_0|_\infty = O_p(J^{-r})$ by (7), it follows that $|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0)|_\infty = O_p(J^{-r})$. Under Conditions (C4) and (C5), and applying the Cauchy-Schwartz inequality, there exist a constant c such that

$$\begin{aligned} |\boldsymbol{\nu}^T \mathbf{S}_{k2}| &= \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{\nu}^T \mathbf{U}_i^T \mathbf{M}_{ki} (\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0)) \right| \\ &\leq \left[\frac{1}{n} \sum_i \boldsymbol{\nu}^T \mathbf{U}_i^T \mathbf{M}_{ki} \mathbf{M}_{ki}^T \mathbf{U}_i \boldsymbol{\nu} \right]^{1/2} \left[\frac{1}{n} \sum_i (\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0))^T (\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\boldsymbol{\theta}_0)) \right]^{1/2} \\ &\leq c (\boldsymbol{\nu}^T \boldsymbol{\nu})^{1/2} \times O_p(J^{-r}) \\ &= O_p(J^{-r}). \end{aligned}$$

Thus, we have $\|\mathbf{S}_{k2}\|_2 = O_p(J^{-r})$. Similarly, by the Cauchy-Schwartz inequality, on the set $\Theta_n(C)$, we can prove that there exist a constant c_2 such that $\|\mathbf{S}_{k3}(\boldsymbol{\theta})\|_2 \leq c_2 C \sqrt{J/n}$.

On the other hand, take $\boldsymbol{\nu} = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2$, then Condition (C5) entail that there exist a positive constant c such that

$$\begin{aligned}
|\boldsymbol{\nu}^T \mathbf{S}_{k3}(\boldsymbol{\theta})| &= \frac{1}{n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2} \sum_{i=1}^n (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \mathbf{U}_i^T \mathbf{M}_{ki} \mathbf{U}_i (\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\
&\geq \frac{c}{n\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2} \sum_{i=1}^n (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \mathbf{U}_i^T \mathbf{U}_i (\boldsymbol{\theta}-\boldsymbol{\theta}_0) \\
&= \frac{c}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2} \|\mathbf{U}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_n^2,
\end{aligned}$$

According to the Condition (C4) and Lemma 1, there exist a positive constant c such that $\|\mathbf{U}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\|_n^2 \geq c_1 \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2^2$. Therefore, we have $|\boldsymbol{\nu}^T \mathbf{S}_{k3}(\boldsymbol{\theta})| \geq c_1 \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_2$. Consequently, we obtain $\|\mathbf{S}_{k3}(\boldsymbol{\theta})\|_2 \geq c_1 C \sqrt{J/n}$. This completes the proof.

Lemma 3

Assume that the Conditions (C1)–(C5) hold, the eigenvalues of $\tilde{\mathbf{C}}_n$ are bounded from 0 and infinite when n is large enough. And, on the set $\Theta_n(C)$, we have $\|\mathbf{S}_n(\boldsymbol{\theta})\|_2 = O_p(\sqrt{J/n} + J^{-r})$ and $\tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) = O_p(n(\sqrt{J/n} + J^{-r})^2)$.

Proof

By Conditions (C4) and (C5) and the definition of $\tilde{\mathbf{C}}_n$, we can easily conclude that $\tilde{\mathbf{C}}_n$ has eigenvalues bounded away from 0 and infinity. In addition, $\tilde{\mathbf{S}} = O_p(\sqrt{J/n} + J^{-r})$ follows directly from Lemma 2, and $\tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) = O_p(n(\sqrt{J/n} + J^{-r})^2)$ follows from the above two results.

Lemma 4

Assume that the Conditions (C1)–(C5) hold, on the set $\Theta_n(C)$ for some C sufficiently large, if $\hat{\mathcal{P}}/n \rightarrow 0$ as $n \rightarrow \infty$, we have

$$\sup_{\boldsymbol{\theta} \in \Theta_n(C)} |\mathbf{Q}_n(\boldsymbol{\theta}) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta})| = o_p(n(\sqrt{J/n} + J^{-r})^2). \quad (\text{A-3})$$

Proof

Notice that

$$\begin{aligned}
\|\mathbf{C}_n^{-1}(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_n^{-1}\| &= \|\mathbf{C}_n^{-1}(\boldsymbol{\theta})[\mathbf{C}_n(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_n]\tilde{\mathbf{C}}_n^{-1}\| \\
&\leq \|\mathbf{C}_n^{-1}(\boldsymbol{\theta})\| \|\mathbf{C}_n(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_n\| \|\tilde{\mathbf{C}}_n^{-1}\|.
\end{aligned} \quad (\text{A-4})$$

Consider the difference of (k, k') -block of $\mathbf{C}_n(\boldsymbol{\theta})$ and $\tilde{\mathbf{C}}_n$

$$\mathbf{C}_{kk'}(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_{kk'} = \frac{1}{n} \sum_{i=1}^n \mathbf{U}_i^T \mathbf{M}_{ki} [\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T - \boldsymbol{\varepsilon}_i^0 \boldsymbol{\varepsilon}_i^{0T}] \mathbf{M}_{ki} \mathbf{U}_i,$$

where $\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) = \mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})$. By Lemma 6.2 of (42), we have

$$\sup_{1 \leq i \leq n, \boldsymbol{\nu}^T \mathbf{U}_i = 1} \|\boldsymbol{\nu}^T \mathbf{U}_i\| = O_p(\sqrt{J}). \quad (\text{A-5})$$

Therefore, by Condition (C5) and (A-5), we have

$$\|\mathbf{C}_{kk'}(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_{kk'}\| = O_p(J) \times O_p\left(\left\|\frac{1}{n} \sum_{i=1}^n [\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_i^T(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}_i^0 \boldsymbol{\varepsilon}_i^{0T}]\right\|\right).$$

Since we can write

$$\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) \boldsymbol{\varepsilon}_i^T(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}_i^0 \boldsymbol{\varepsilon}_i^{0T} = \boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) (\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}_i^0)^T + (\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}_i^0) \boldsymbol{\varepsilon}_i^{0T},$$

from definitions of $\boldsymbol{\varepsilon}_i(\boldsymbol{\theta})$ and $\boldsymbol{\varepsilon}_i^0$, we have

$$\max \|\boldsymbol{\varepsilon}_i(\boldsymbol{\theta})\| \leq 1 \quad \text{and} \quad \max \|\boldsymbol{\varepsilon}_i^0\| \leq 1,$$

and by Taylor's expansion, we can prove

$$\sup_{\boldsymbol{\theta} \in \Theta_n(C)} \|\boldsymbol{\varepsilon}_i(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}_i^0\| = O_p(\sqrt{J/n} + J^{-r}),$$

which leads to

$$\|\mathbf{C}_{kk'}(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_{kk'}\| = O_p(J) \times O_p(\sqrt{J/n} + J^{-r}) = O_p(J^{3/2} / \sqrt{n} + J^{-(r-1)}).$$

Consequently, we obtain

$$\|\mathbf{C}_n(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_n\| = O_p(J^{3/2} / \sqrt{n} + J^{-(r-1)}).$$

Since the eigenvalues of $\tilde{\mathbf{C}}_n$ are bounded from 0 and infinity from Lemma 3, thus the eigenvalues of $\mathbf{C}_n(\boldsymbol{\theta})$ are also bounded from 0 and infinity, leading to

$$\|\mathbf{C}_n^{-1}(\boldsymbol{\theta})\| = O_p(1) \text{ and } \|\tilde{\mathbf{C}}_n^{-1}\| = O_p(1).$$

Hence, by (A-4), we get

$$\|\mathbf{C}_n^{-1}(\boldsymbol{\theta}) - \tilde{\mathbf{C}}_n^{-1}\| = O_p(J^{3/2} / \sqrt{n} + J^{-(r-1)}). \quad (\text{A-6})$$

Finally, from the definitions of $\mathbf{Q}_n(\boldsymbol{\theta})$ and $\tilde{\mathbf{Q}}_n(\boldsymbol{\theta})$, and $\|\mathbf{S}_n(\boldsymbol{\theta})\|_2 = O_p(\sqrt{J/n} + J^{-r})$ by Lemma 3, (A-3) holds from (A-6) and the condition $\tilde{\mathcal{P}}/n = o(1)$. This completes the proof.

Lemma 5

Under the Conditions (C1)–(C5), for any small positive ε , there exists a constant ϱ and $C = C(\varepsilon)$ such that

$$P \left\{ \inf_{\boldsymbol{\theta} \in \Theta_n(c)} \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) > \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}_0) + \varrho J \right\} > 1 - \varepsilon$$

for n large enough.

Proof

Let $\boldsymbol{\theta} = \mathbf{S}(\boldsymbol{\theta}) - \mathbf{S}(\boldsymbol{\theta}_0)$. Then, we can write

$$\frac{1}{n} \left\{ \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}_0) \right\} = \boldsymbol{\Delta}^T(\boldsymbol{\theta}) \tilde{\mathbf{C}}_n^{-1} \boldsymbol{\Delta}(\boldsymbol{\theta}) + 2 \boldsymbol{\Delta}^T(\boldsymbol{\theta}) \tilde{\mathbf{C}}_n^{-1} \mathbf{S}(\boldsymbol{\theta}_0).$$

According to the definitions of $\mathbf{S}(\boldsymbol{\theta})$ and $\mathbf{S}(\boldsymbol{\theta}_0)$, we have $\boldsymbol{\Delta}(\boldsymbol{\theta}) = (\mathbf{S}_{13}^T(\boldsymbol{\theta}), \dots, \mathbf{S}_{v3}^T(\boldsymbol{\theta}))^T$. By Lemma 3, the eigenvalues of $\tilde{\mathbf{C}}_n^{-1}$ are bounded; hence, there exist a constant $c > 0$, such that on the set $\Theta_n(C)$

$$\begin{aligned} |\boldsymbol{\Delta}^T(\boldsymbol{\theta}) \tilde{\mathbf{C}}_n^{-1} \mathbf{S}(\boldsymbol{\theta}_0)| &\leq c \sum_{k=1}^v \left| (\mathbf{S}_{k3}(\boldsymbol{\theta}))^T (\mathbf{S}_{k1} + \mathbf{S}_{k2}) \right| \\ &= O_p \left\{ \sum_{k=1}^v \|\mathbf{S}_{k3}(\boldsymbol{\theta})\|_2 (\|\mathbf{S}_{k1}\|_2 + \|\mathbf{S}_{k2}\|_2) \right\}. \end{aligned}$$

Thus, by Lemma 2 we have

$$|\Delta^T(\boldsymbol{\theta})\tilde{\mathbf{C}}_n^{-1}\mathbf{S}(\boldsymbol{\theta}_0)|=O_p(J/n)$$

when n is large enough. On the other hand, there exist a constant $c' > 0$ such that

$$\Delta^T(\boldsymbol{\theta})\tilde{\mathbf{C}}_n^{-1}\Delta(\boldsymbol{\theta}) \geq c' \sum_{k=1}^v \|\mathbf{S}_{k3}(\boldsymbol{\theta})\|_2^2.$$

Hence, we have

$$\Delta^T(\boldsymbol{\theta})\tilde{\mathbf{C}}_n^{-1}\Delta(\boldsymbol{\theta})=O_p(J/n),$$

again by Lemma 2. Choosing C sufficiently large, we can obtain the desired result. This completes the proof.

Lemma 6

Under the conditions in Lemma 4, for any $\varepsilon > 0$, when n is large enough, there exists $C = C(\varepsilon)$ such that

$$P \left\{ \inf_{\boldsymbol{\theta} \in \Theta_n(C)} \mathbf{Q}_n(\boldsymbol{\theta}) > \mathbf{Q}_n(\boldsymbol{\theta}_0) \right\} > 1 - \varepsilon.$$

Proof

Note that we can write

$$\mathbf{Q}_n(\boldsymbol{\theta}) - \mathbf{Q}_n(\boldsymbol{\theta}_0) = \mathbf{Q}_n(\boldsymbol{\theta}) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) + \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}_0) + \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}_0) - \mathbf{Q}_n(\boldsymbol{\theta}_0).$$

By previous lemmas, we have both

$$|\mathbf{Q}_n(\boldsymbol{\theta}) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta})| = o_p(J) \quad \text{and} \quad |\mathbf{Q}_n(\boldsymbol{\theta}_0) - \tilde{\mathbf{Q}}_n(\boldsymbol{\theta}_0)| = o_p(J),$$

and together with Lemma 5, the result is obvious. This completes the proof.

We next establish the asymptotic normality for parametric component. Denote

$$\mathbf{G}_{k\gamma} = \frac{1}{n} \sum_i \mathbf{U}_i^T \mathbf{M}_{ki} \boldsymbol{\Gamma}_i \boldsymbol{\pi}_i \quad \text{and} \quad \mathbf{G}_{k\beta} = \frac{1}{n} \sum_i \mathbf{U}_i^T \mathbf{M}_{ki} \boldsymbol{\Gamma}_i \mathbf{z}_i \quad \text{where} \quad \boldsymbol{\Gamma}_i = \text{diag}(\boldsymbol{\mu}_i^0(1 - \boldsymbol{\mu}_i^0)).$$

Let $\mathbf{G}_k = (\mathbf{G}_{k\gamma}, \mathbf{G}_{k\beta})$, and

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_1 \\ \vdots \\ \mathbf{G}_v \end{pmatrix}, \quad \mathbf{G}_\gamma = \begin{pmatrix} \mathbf{G}_{1\gamma} \\ \vdots \\ \mathbf{G}_{v\gamma} \end{pmatrix}, \quad \mathbf{G}_\beta = \begin{pmatrix} \mathbf{G}_{1\beta} \\ \vdots \\ \mathbf{G}_{v\beta} \end{pmatrix}$$

In matrix notations, let $\mathbf{U} = (\mathbf{U}_1^T, \dots, \mathbf{U}_n^T)^T$, $\mathbf{\Gamma} = \text{diag}(\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_n)$ and $\mathbf{\Xi}_k = \text{diag}(\mathbf{M}_{k1}, \dots, \mathbf{M}_{kn})$, $k = 1, \dots, v$. Then, we can write

$$\mathbf{G}_k = \frac{1}{n} \mathbf{U}^T \mathbf{\Xi}_k \mathbf{\Gamma} \mathbf{U}, \quad \mathbf{G}_{k\gamma} = \frac{1}{n} \mathbf{U}^T \mathbf{\Xi}_k \mathbf{\Gamma} \mathbf{\Pi}, \quad \mathbf{G}_{k\beta} = \frac{1}{n} \mathbf{U}^T \mathbf{\Xi}_k \mathbf{\Gamma} \mathbf{Z},$$

and

$$\mathbf{G} = \frac{1}{n} (\mathbf{I}_v \otimes \mathbf{U}^T) \mathbf{\Xi} \mathbf{\Gamma} \mathbf{U}, \quad \mathbf{G}_\gamma = \frac{1}{n} (\mathbf{I}_v \otimes \mathbf{U}^T) \mathbf{\Xi} \mathbf{\Gamma} \mathbf{\Pi}, \quad \mathbf{G}_\beta = \frac{1}{n} (\mathbf{I}_v \otimes \mathbf{U}^T) \mathbf{\Xi} \mathbf{\Gamma} \mathbf{Z},$$

where $\mathbf{\Xi} = (\mathbf{\Xi}_1^T, \dots, \mathbf{\Xi}_v^T)^T$ and $\otimes$ denotes the Kronecker product of matrix. Let the

transformation matrix $\mathbf{P} = (\mathbf{G}_\gamma^T \tilde{\mathbf{C}}_n^{-1} \mathbf{G}_\gamma)^{-1} \mathbf{G}_\gamma^T \tilde{\mathbf{C}}_n^{-1} \mathbf{G}_\beta$, where $\tilde{\mathbf{C}}_n$ is defined by (A-2) in Appendix. Note that the transformation matrix $\mathbf{P}$ can be seen as the projection of the design matrix $\mathbf{Z}$ in linear part into the spline basis space $\mathbf{\Pi}$, such that

$$\mathbf{P} = \arg\min_{\Omega} (\mathbf{Z} - \mathbf{\Pi} \mathbf{\Omega})^T (\mathbf{\Gamma} \mathbf{\Xi}^T (\mathbf{I}_v \otimes \mathbf{U}) \tilde{\mathbf{C}}_n^{-1} (\mathbf{I}_v \otimes \mathbf{U}^T) \mathbf{\Xi} \mathbf{\Gamma}) (\mathbf{Z} - \mathbf{\Pi} \mathbf{\Omega}).$$

Define matrix $\mathbf{R} = \begin{pmatrix} \mathbf{I}_{\sum_{l=1}^p K_l} & -\mathbf{P} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}$ and denote $\mathcal{U}_i = \mathbf{U}_i \mathbf{R} = (\boldsymbol{\pi}_i, \mathbf{z}_i^*)$, $\mathbf{z}_i^* = \mathbf{z}_i - \boldsymbol{\pi}_i \mathbf{P}$ and $\mathcal{G}_\beta = \frac{1}{n} (\mathbf{I}_v \otimes \mathcal{U}^T) \mathbf{\Xi} \mathbf{\Gamma} \mathbf{Z}^*$, where $\mathbf{Z}^* = (\mathbf{z}_1^{*T}, \dots, \mathbf{z}_n^{*T})^T$ and $\mathcal{U} = (\mathcal{U}_1^T, \dots, \mathcal{U}_n^T)^T$. Denote

$$\begin{aligned} \mathcal{S}_{ik}(\boldsymbol{\theta}) &= \mathbf{R} \mathbf{S}_{ik}(\boldsymbol{\theta}), \quad k=1, \dots, v, i=1, \dots, n, \\ \mathcal{S}_i(\boldsymbol{\theta}) &= (\mathcal{S}_{i1}^T(\boldsymbol{\theta}), \dots, \mathcal{S}_{iv}^T(\boldsymbol{\theta}))^T, \\ \mathcal{Q}_n(\boldsymbol{\theta}) &= n \mathcal{S}^T(\boldsymbol{\theta}) \{\mathcal{E}_n(\boldsymbol{\theta})\}^{-1} \mathcal{S}(\boldsymbol{\theta}), \end{aligned}$$

where

$$\mathcal{S}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathcal{S}_i(\boldsymbol{\theta}), \quad \mathcal{E}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathcal{S}_i(\boldsymbol{\theta}) \mathcal{S}_i^T(\boldsymbol{\theta}).$$

From the above notations, we can write

$$\mathcal{S}(\boldsymbol{\theta}) = (\mathbf{I}_v \otimes \mathbf{R})\mathbf{S}(\boldsymbol{\theta}), \quad \mathcal{C}_n(\boldsymbol{\theta}) = (\mathbf{I}_v \otimes \mathbf{R})\mathbf{C}_n(\boldsymbol{\theta})(\mathbf{I}_v \otimes \mathbf{R})^T.$$

Hence, we obtain

$$\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathcal{Q}_n(\boldsymbol{\theta}) = n\mathbf{S}^T(\boldsymbol{\theta})(\mathbf{I}_v \otimes \mathbf{R}^T) \{(\mathbf{I}_v \otimes \mathbf{R})\mathbf{C}_n(\boldsymbol{\theta})(\mathbf{I}_v \otimes \mathbf{R}^T)\}^{-1}(\mathbf{I}_v \otimes \mathbf{R})\mathbf{S}(\boldsymbol{\theta}).$$

From the definition of the invertible matrix $\mathbf{R}$, and by checking the matrix multiplication for $\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})$, we can conclude

$$\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathbf{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) \quad (\text{A-7})$$

for any $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. On the other hand, note that we can write

$$\mathbf{U}_i \boldsymbol{\theta} = \boldsymbol{\pi}_i(\boldsymbol{\gamma} + \mathbf{P}\boldsymbol{\beta}) + (\mathbf{z}_i - \boldsymbol{\pi}_i \mathbf{P})\boldsymbol{\beta} \triangleq \boldsymbol{\mathcal{U}}_i \boldsymbol{\theta}^*,$$

where $\boldsymbol{\mathcal{U}}_i = \mathbf{U}_i \mathbf{R} = (\boldsymbol{\pi}_i, \mathbf{z}_i^*)$, $\mathbf{z}_i^* = \mathbf{z}_i - \boldsymbol{\pi}_i \mathbf{P}$, $\boldsymbol{\theta}^* = (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)^T$ and

$$\begin{cases} \boldsymbol{\gamma}^* = \boldsymbol{\gamma} + \mathbf{P}\boldsymbol{\beta}, \\ \boldsymbol{\beta}^* = \boldsymbol{\beta}, \end{cases}$$

which is a one to one parameter transformation $(\boldsymbol{\gamma}, \boldsymbol{\beta}) \rightarrow (\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$. Therefore, we can also write $\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})$ as $\mathcal{Q}_n(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$, since

$$\frac{1}{n} \sum_{i=1}^n \boldsymbol{\mathcal{U}}_i^T \boldsymbol{\mathcal{U}}_i = \mathbf{R}^T \left(\frac{1}{n} \sum_{i=1}^n \mathbf{U}_i^T \mathbf{U}_i \right) \mathbf{R}. \quad (\text{A-8})$$

From Condition (C4), we can conclude that the smallest and largest eigenvalues of (A-7) are also bounded away from zero and infinity. Denote $\hat{\boldsymbol{\theta}}^* = \operatorname{argmin}_{\boldsymbol{\theta}^*} \mathcal{Q}_n(\boldsymbol{\theta}^*)$. Following Theorem 1, we can prove that

$$\hat{\boldsymbol{\theta}}^* = \boldsymbol{\theta}_0^* + O_p(n^{-r/(2r+1)}),$$

where $\boldsymbol{\theta}_0^*$ is the true value of $\boldsymbol{\theta}^*$. As a result, we have

$$\hat{\gamma} = \gamma_0 + O_p(n^{-r/(2r+1)}) \quad \text{and} \quad \hat{\beta} = \beta_0 + O_p(n^{-r/(2r+1)}).$$

Let $\tilde{\mathcal{S}}(\gamma, \beta) = \mathcal{S}(\gamma, \beta_0) - \mathcal{S}(\beta - \beta_0)$, $\tilde{\mathcal{Q}}(\gamma, \beta) = n\tilde{\mathcal{S}}^T(\gamma, \beta)\tilde{\mathcal{C}}_n^{-1}\tilde{\mathcal{S}}(\gamma, \beta)$ and $\tilde{\mathcal{C}}_n = (\mathbf{I}_v \otimes \mathbf{R})\tilde{\mathbf{C}}_n(\mathbf{I}_v \otimes \mathbf{R})^T$. We first prove some lemmas.

Lemma 7

Assume the Conditions (C1)–(C5) hold, for any constant C , we have

$$\sup_{\|\gamma - \gamma_0\| \leq C\sqrt{J/n}, \|\beta - \beta_0\| \leq C/\sqrt{n}} \|\mathcal{S}(\gamma, \beta) - \tilde{\mathcal{S}}(\gamma, \beta)\|_2 = O_p(1/\sqrt{n}).$$

Proof

By the definition of $\mathcal{S}(\gamma, \beta)$ and $\tilde{\mathcal{S}}(\gamma, \beta)$, we write

$$\mathcal{S}(\gamma, \beta) - \tilde{\mathcal{S}}(\gamma, \beta) = \frac{1}{n} \sum_{i=1}^n (\mathbf{I}_v \otimes \mathcal{W}_i^T) \Xi_i [\mu_i(\gamma, \beta_0) - \mu_i(\gamma, \beta)] + \frac{1}{n} \sum_{i=1}^n (\mathbf{I}_v \otimes \mathcal{W}_i^T) \Xi_i \Gamma_i \mathbf{z}_i (\beta - \beta_0),$$

where $\mu_i(\gamma, \beta) = \exp(\Pi_i \gamma + \mathbf{z}_i \beta) / (1 + \exp(\Pi_i \gamma + \mathbf{z}_i \beta))$. It follows that

$$\|\mu_i(\gamma, \beta_0) - \mu_i(\gamma, \beta)\| = O_p(\|\beta - \beta_0\|).$$

By the similar proof of $\mathbf{S}_{k3}$ in Lemma 2, it is easy to obtain

$$\sup_{\|\gamma - \gamma_0\| \leq C\sqrt{J/n}, \|\beta - \beta_0\| \leq C/\sqrt{n}} \|\mathcal{S}(\gamma, \beta) - \tilde{\mathcal{S}}(\gamma, \beta)\|_2 = O_p(1/\sqrt{n}).$$

Lemma 8

Assume the Conditions (C1)–(C5) hold, we have

$$\|\mathcal{S}(\gamma, \beta)\|_2 = O_p(\sqrt{J/n}) \quad \text{and} \quad \|\tilde{\mathcal{C}}_n^{-1} - \mathcal{C}_n^{-1}(\gamma, \beta)\| = O_p(J^{3/2}/\sqrt{n} + J^{-(r-1)}).$$

In addition, the eigenvalues of $\tilde{\mathcal{C}}_n$ are also bounded from 0 and infinity, when n is large enough.

Proof

From the previous definitions, $\mathcal{S}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{I}_v \otimes \mathbf{R})\mathbf{S}(\boldsymbol{\theta})$, $\tilde{\mathcal{C}}_n = (\mathbf{I}_v \otimes \mathbf{R})\tilde{\mathbf{C}}_n(\mathbf{I}_v \otimes \mathbf{R})^T$ and $\tilde{\mathcal{C}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = (\mathbf{I}_v \otimes \mathbf{R})\mathbf{C}_n(\boldsymbol{\theta})(\mathbf{I}_v \otimes \mathbf{R})^T$. With $\mathbf{R}$ an invertible matrix, and using results in the proof of Lemma 3 and Lemma 4, it is obvious that the desired results hold.

Lemma 9

Assume the Conditions (C1)–(C5) hold, we have

$$\sup_{\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leq C\sqrt{J/n}, \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \leq C/\sqrt{n}} |\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) - \tilde{\mathcal{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})| = o_p(1).$$

Proof

This result can be directly obtained from Lemmas 7 and 8.

Denote $\tilde{\boldsymbol{\beta}}$ as the minimizer of $\tilde{\mathcal{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for given $\boldsymbol{\gamma}$ ($\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\| \leq C\sqrt{J/n}$), we next prove that $\tilde{\boldsymbol{\beta}}$ is asymptotically normal.

Lemma 10

Assume the Conditions (C1)–(C5) hold, then we have

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0}, (\mathcal{G}_{\boldsymbol{\beta}}^T \sum_n^{-1} \mathcal{G}_{\boldsymbol{\beta}})^{-1}\right).$$

Proof

Recall that

$$\tilde{\mathcal{J}}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathcal{J}(\boldsymbol{\gamma}, \boldsymbol{\beta}_0) - \mathcal{G}_{\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \quad \text{and} \quad \tilde{\mathcal{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = n\tilde{\mathcal{J}}^T(\boldsymbol{\gamma}, \boldsymbol{\beta})\tilde{\mathcal{C}}_n^{-1}\tilde{\mathcal{J}}(\boldsymbol{\gamma}, \boldsymbol{\beta}).$$

Notice that the first derivative of $\tilde{\mathcal{Q}}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is given as

$$\frac{\partial \tilde{\mathcal{Q}}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2n\{\mathcal{G}_{\boldsymbol{\beta}}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{G}_{\boldsymbol{\beta}}\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - 2n\mathcal{G}_{\boldsymbol{\beta}}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{J}(\boldsymbol{\gamma}, \boldsymbol{\beta}_0),$$

Thus, we have

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \{\mathcal{G}_{\boldsymbol{\beta}}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{G}_{\boldsymbol{\beta}}\}^{-1} \mathcal{G}_{\boldsymbol{\beta}}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{J}(\boldsymbol{\gamma}, \boldsymbol{\beta}_0),$$

since

$$\mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{J}(\gamma, \beta_0) = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} (\mathbf{I}_v \otimes \mathcal{U}^T) \Xi(\mathbf{y} - \boldsymbol{\mu}(\gamma, \beta_0)) \triangleq \mathbf{W}_1 + \mathbf{W}_2,$$

where

$$\mathbf{W}_1 = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} \sum_{i=1}^n (\mathbf{I}_v \otimes \mathcal{U}_i^T) \Xi_i(\mathbf{y}_i - \boldsymbol{\mu}_i^0),$$

and

$$\mathbf{W}_2 = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} \sum_{i=1}^n (\mathbf{I}_v \otimes \mathcal{U}_i^T) \Xi_i[\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i(\gamma, \beta_0)],$$

We next need to prove $\sqrt{n}\mathbf{W}_1$ is asymptotically normal, and $\mathbf{W}_2 = o_p(n^{-1/2})$. For $\mathbf{W}_2$, by the Taylor's expansion for $h(\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0)$, we have

$$h(\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0) = h(\eta_{ij}^0) + h'(\eta_{ij}^0)(\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0 - \eta_{ij}^0) + h''(\varpi)(\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0 - \eta_{ij}^0)^2,$$

where ϖ lies between η_{ij}^0 and $\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0$. We can write

$$\boldsymbol{\pi}_{ij}^T \gamma + \mathbf{z}_{ij}^T \beta_0 - \eta_{ij}^0 = \boldsymbol{\pi}_{ij}^T (\gamma - \gamma_0) + \left[\boldsymbol{\pi}_{ij}^T \gamma_0 - \sum_{l=1}^p g_l(x_{ijl}) \right],$$

and from Lemma 1,

$$\|\boldsymbol{\Pi}(\gamma_0 - \gamma)\|_n^2 = O(\|\gamma_0 - \gamma\|_n^2) = O(J/n).$$

Hence, we can write

$$\mathbf{W}_2 = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} (\mathbf{I}_v \otimes \mathcal{U}^T) [\Xi \Gamma \boldsymbol{\Pi}(\gamma_0 - \gamma) + O_p(J/n)] = \mathbf{W}_{21} + \mathbf{W}_{22},$$

where $\mathbf{W}_{21} = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} \mathcal{G}_{\gamma}(\gamma_0 - \gamma)$, $\mathbf{W}_{22} = \frac{1}{n} \mathcal{G}_{\beta}^T \tilde{\mathcal{C}}_n^{-1} (\mathbf{I}_v \otimes \mathcal{U}^T) \times \sqrt{J/n}$.

By the definition of $\mathbf{P}$, we have $\mathcal{G}_\beta^T(\tilde{\mathcal{C}}_n)^{-1}\mathcal{G}_\gamma=0$, which leads to $\mathbf{W}_{21}=0$. For $\mathbf{W}_{22}$, the eigenvalues of $\tilde{\mathcal{C}}_n$ are bounded by checking the regularity conditions. Combining (7), we can easily obtain $\mathbf{W}_2=O_p(\sqrt{J}/n)=o_p(n^{-1/2})$.

For $\mathbf{W}_1$, it is obvious that $E(\mathbf{W}_1)=0$. From the definition of $\sum=\text{Cov}(\frac{1}{\sqrt{n}}(\mathbf{I}_v \otimes \mathcal{U}^T)\Xi\epsilon^0)$, and the law of large numbers $\tilde{\mathcal{C}}_n \rightarrow \Sigma$ in probability, it follows that

$$\text{Cov}(\sqrt{n}\mathbf{W}_1)=\mathcal{G}_\beta^T\tilde{\mathcal{C}}_n^{-1}\mathcal{G}_\beta=\mathcal{G}_\beta^T\sum^{-1}\mathcal{G}_\beta+o_p(1).$$

We next show that $\forall \mathbf{v} \in \mathbb{R}^{\sum_{i=1}^p J_i+q}$, $\mathbf{v}^T\mathbf{v}=1$,

$$\mathbf{v}^T\tilde{\mathcal{C}}_n^{-1/2}\frac{1}{n}\sum_{i=1}^n(\mathbf{I}_v \otimes \mathcal{U}_i^T)\Xi_i(\mathbf{y}_i-\boldsymbol{\mu}_i^0) \xrightarrow{d} N(0,1).$$

Denote $H_{ni}=\frac{1}{\sqrt{n}}\mathbf{v}^T\tilde{\mathcal{C}}_n^{-1/2}(\mathbf{I}_v \otimes \mathcal{U}_i^T)\Xi_i\epsilon_i^0$ with $\epsilon_i^0=\mathbf{y}_i-\boldsymbol{\mu}_i^0$. Using Cauchy-Schwarz inequality and Condition (C5), we have

$$H_{ni}^2 \leq C\omega_{ni}\|\epsilon_i^0\|_2^2,$$

where $\omega_{ni}=n^{-1}(\tilde{\mathcal{C}}_n)^{-1/2}(\mathbf{I}_v \otimes \mathcal{U}_i^T)(\mathbf{I}_v \otimes \mathcal{U}_i)(\tilde{\mathcal{C}}_n)^{-1/2}$. By (A-5) and the previous lemmas, we can conclude

$$\sup_i \omega_{ni}=n^{-1} \times O_p(J)=o_p(1). \quad (\text{A-9})$$

To finish the proof of asymptotic normality, we check the Lyapunov conditions for $\sum_{i=1}^n H_{ni}$, i.e., for arbitrary $\delta > 0$, we want to show $\sum_{i=1}^n E(H_{ni}^{2+\delta}) \rightarrow 0$, as $n \rightarrow \infty$.

Due to the boundedness of $\|\epsilon_i^0\|$, we have

$$\begin{aligned} \sum_{i=1}^n E(H_{ni}^{2+\delta}) &\leq \sum_{i=1}^n E(C^{1+\delta/2}\omega_{ni}^{1+\delta/2}\|\epsilon_i^0\|_2^{2+\delta}), \\ &\leq C\sup_i \omega_{ni}^{\delta/2} \sum_{i=1}^n \omega_{ni}. \end{aligned}$$

Finally, according to (A-9) and (A-8), it follows that $\sum_{i=1}^n E(H_{ni}^{2+\delta})=o_p(1)$. This completes the proof.

APPENDIX B: Proof of Theorems

Proof of Theorem 1

Proof

Lemma 6 implies that $\mathbf{Q}_n(\boldsymbol{\theta})$ has a local minimum in the ball $\{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq C\sqrt{J/n}\}$. Hence, we conclude

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_p(\sqrt{J/n}).$$

The desired results can be obtained by applying the Lemma 1. This completes the proof.

Proof of Theorem 2

Proof

Note that $(\hat{\boldsymbol{\gamma}}^T, \hat{\boldsymbol{\beta}}^T)^T$ is the minimizer of $\mathbf{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})$. Since $\mathcal{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathbf{Q}_n(\boldsymbol{\gamma}, \boldsymbol{\beta})$ by (A-7), $\hat{\boldsymbol{\beta}}$ is also the minimizer of $\mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta})$. According to the Lemma 10, we only need to prove that $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\| = o_p(1/\sqrt{n})$. To this end, it is sufficient to prove for any $\varepsilon > 0$,

$$\inf_{\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2 \geq \varepsilon/\sqrt{n}} \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) > 0 \quad (\text{B-1})$$

with probability tending to 1 as $n \rightarrow \infty$.

Note that we can decompose the term $\mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})$ into the following three parts

$$\mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) = \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) + \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) + \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) - \mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}).$$

(B-2)

By the quadratic form of $\tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta})$ in terms of $\boldsymbol{\beta}$, we obtain

$$|\tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})| \geq Cn\|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\|_2^2. \quad (\text{B-3})$$

On the other hand, by Lemma 9, we have both

$$|\mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta}) - \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \boldsymbol{\beta})| = o_p(1) \quad \text{and} \quad |\mathcal{Q}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}}) - \tilde{\mathcal{Q}}_n(\hat{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}})| = o_p(1).$$

From (B-2) and (B-3), we can conclude $\|\hat{\beta} - \tilde{\beta}\| = o_p(1/\sqrt{n})$. This completes the proof.

APPENDIX C: Tables

Table C1

Results from Simulation 1: Reported values are AIMSE (average integrated mean squared error) for Cases 1 & 2, when the true correlation structure is AR(1)

ρ	n	Case 1			Case 2		
		WI	CS	AR(1)	WI	CS	AR(1)
0.5	100	0.0437	0.0436	0.0416	0.0573	0.0571	0.0544
	200	0.0343	0.0332	0.0319	0.0427	0.0412	0.0392
	300	0.0303	0.0295	0.0285	0.0370	0.0355	0.0340
0.8	100	0.0436	0.0355	0.0341	0.0573	0.0463	0.0436
	200	0.0339	0.0288	0.0277	0.0423	0.0345	0.0330
	300	0.0305	0.0265	0.0257	0.0372	0.0307	0.0295

Table C2

Results from Simulation 1: Here, BIAS = average bias, ABIAS = absolute bias, ASE = asymptotic standard error estimate (derived from the asymptotic normality results), and ESE = the empirical standard error estimates, calculated over 1000 replications, when the true correlation structure is AR(1) for Case 1.

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.5	100	BIAS	-0.0005	-0.0004	0.0007	0.0012	-0.0018	0.0015	0.0010	-0.0013	0.0012
		ABIAS	0.0120	0.0118	0.0195	0.0121	0.0119	0.0189	0.0114	0.0112	0.0177
		ASE	0.0149	0.0149	0.0232	0.0127	0.0127	0.0197	0.0118	0.0118	0.0181
		ESE	0.0152	0.0149	0.0244	0.0151	0.0150	0.0239	0.0143	0.0140	0.0224
	200	BIAS	0.0006	-0.0009	0.0012	0.0012	-0.0012	0.0016	0.0009	-0.0010	0.0014
		ABIAS	0.0088	0.0090	0.0133	0.0085	0.0085	0.0128	0.0079	0.0081	0.0117
		ASE	0.0106	0.0106	0.0166	0.0094	0.0093	0.0146	0.0087	0.0087	0.0135
		ESE	0.0110	0.0110	0.0167	0.0105	0.0106	0.0160	0.0098	0.0101	0.0147
	300	BIAS	0.0004	-0.0001	0.0006	0.0005	-0.0004	0.0006	0.0004	-0.0003	0.0002
		ABIAS	0.0069	0.0070	0.0111	0.0065	0.0065	0.0103	0.0061	0.0060	0.0095
		ASE	0.0087	0.0087	0.0135	0.0078	0.0078	0.0121	0.0073	0.0073	0.0112
		ESE	0.0087	0.0088	0.0140	0.0081	0.0082	0.0131	0.0077	0.0076	0.0119
0.8	100	BIAS	-0.0005	0.0000	0.0008	0.0011	-0.0014	0.0021	0.0011	-0.0015	0.0020
		ABIAS	0.0119	0.0118	0.0188	0.0091	0.0096	0.0142	0.0087	0.0090	0.0133
		ASE	0.0149	0.0149	0.0232	0.0099	0.0099	0.0152	0.0091	0.0091	0.0138
		ESE	0.0150	0.0149	0.0237	0.0115	0.0119	0.0178	0.0108	0.0111	0.0166

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
	200	BIAS	0.0000	-0.0001	0.0000	0.0009	-0.0012	0.0007	0.0007	-0.0008	0.0007
		ABIAS	0.0084	0.0088	0.0133	0.0063	0.0066	0.0094	0.0058	0.0060	0.0088
		ASE	0.0106	0.0106	0.0164	0.0074	0.0074	0.0112	0.0068	0.0068	0.0102
		ESE	0.0108	0.0109	0.0165	0.0078	0.0083	0.0118	0.0073	0.0076	0.0110
	300	BIAS	0.0001	-0.0006	0.0001	0.0005	-0.0006	0.0001	0.0006	-0.0008	0.0000
		ABIAS	0.0070	0.0067	0.0107	0.0053	0.0052	0.0080	0.0048	0.0047	0.0072
		ASE	0.0087	0.0087	0.0135	0.0061	0.0061	0.0093	0.0056	0.0056	0.0085
		ESE	0.0087	0.0083	0.0135	0.0066	0.0065	0.0099	0.0060	0.0059	0.0090

Table C3

Results from Simulation 1: Here, BIAS = average bias, ABIAS = absolute bias, ASE = asymptotic standard error estimate (derived from the asymptotic normality results), and ESE = the empirical standard error estimates, calculated over 1000 replications, when the true correlation structure is compound symmetry (CS) for Case 2

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.5	100	BIAS	0.0014	-0.0006	0.0007	0.0056	-0.0048	0.0048	0.0054	-0.0048	0.0045
		ABIAS	0.0167	0.0166	0.0269	0.0159	0.0155	0.0238	0.0165	0.0165	0.0257
		ASE	0.0210	0.0209	0.0326	0.0159	0.0159	0.0245	0.0168	0.0168	0.0259
		ESE	0.0208	0.0212	0.0336	0.0189	0.0189	0.0294	0.0200	0.0204	0.0318
	200	BIAS	0.0002	-0.0001	0.0000	0.0024	-0.0024	0.0021	0.0022	-0.0021	0.0018
		ABIAS	0.0119	0.0123	0.0189	0.0107	0.0107	0.0156	0.0111	0.0115	0.0171
		ASE	0.0149	0.0149	0.0233	0.0118	0.0118	0.0182	0.0125	0.0125	0.0194
		ESE	0.0149	0.0155	0.0236	0.0132	0.0133	0.0196	0.0139	0.0143	0.0213
	300	BIAS	0.0004	-0.0002	-0.0008	0.0018	-0.0017	0.0007	0.0016	-0.0016	0.0005
		ABIAS	0.0098	0.0100	0.0156	0.0083	0.0085	0.0134	0.0091	0.0091	0.0143
		ASE	0.0122	0.0122	0.0190	0.0098	0.0098	0.0151	0.0104	0.0104	0.0161
		ESE	0.0120	0.0124	0.0193	0.0103	0.0105	0.0166	0.0111	0.0113	0.0178
0.8	100	BIAS	0.0014	-0.0011	-0.0010	0.0060	-0.0062	0.0031	0.0058	-0.0058	0.0032
		ABIAS	0.0171	0.0172	0.0278	0.0123	0.0128	0.0174	0.0132	0.0139	0.0207
		ASE	0.0210	0.0210	0.0325	0.0122	0.0121	0.0178	0.0135	0.0134	0.0201
		ESE	0.0217	0.0215	0.0346	0.0142	0.0148	0.0215	0.0157	0.0164	0.0256
	200	BIAS	0.0005	-0.0008	0.0013	0.0032	-0.0036	0.0034	0.0033	-0.0036	0.0037
		ABIAS	0.0124	0.0120	0.0187	0.0085	0.0086	0.0121	0.0094	0.0094	0.0139
		ASE	0.0150	0.0150	0.0232	0.0090	0.0091	0.0132	0.0101	0.0101	0.0151
		ESE	0.0156	0.0150	0.0237	0.0102	0.0101	0.0148	0.0115	0.0112	0.0170

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
	300	BIAS	0.0005	-0.0007	0.0007	0.0024	-0.0025	0.0016	0.0024	-0.0026	0.0017
		ABIAS	0.0100	0.0094	0.0150	0.0068	0.0069	0.0092	0.0075	0.0076	0.0106
		ASE	0.0123	0.0122	0.0190	0.0075	0.0075	0.0110	0.0084	0.0084	0.0126
		ESE	0.0123	0.0120	0.0188	0.0081	0.0082	0.0115	0.0091	0.0092	0.0132

Table C4

Results from Simulation 1: Here, BIAS = average bias, ABIAS = absolute bias, ASE = asymptotic standard error estimate (derived from the asymptotic normality results), and ESE = the empirical standard error estimates, calculated over 1000 replications, when the true correlation structure is AR(1) for Case 2

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.5	100	BIAS	0.0003	-0.0003	0.0001	0.0041	-0.0044	0.0032	0.0037	-0.0038	0.0027
		ABIAS	0.0178	0.0174	0.0255	0.0180	0.0172	0.0255	0.0174	0.0169	0.0250
		ASE	0.0209	0.0209	0.0327	0.0178	0.0178	0.0276	0.0166	0.0166	0.0255
		ESE	0.0221	0.0217	0.0323	0.0222	0.0218	0.0327	0.0212	0.0208	0.0311
	200	BIAS	0.0014	-0.0017	0.0007	0.0033	-0.0033	0.0020	0.0030	-0.0030	0.0019
		ABIAS	0.0117	0.0119	0.0187	0.0115	0.0113	0.0181	0.0106	0.0107	0.0170
		ASE	0.0149	0.0149	0.0232	0.0132	0.0132	0.0204	0.0123	0.0123	0.0189
		ESE	0.0148	0.0147	0.0236	0.0139	0.0138	0.0226	0.0129	0.0131	0.0213
	300	BIAS	0.0003	-0.0006	0.0005	0.0019	-0.0022	0.0020	0.0016	-0.0020	0.0015
		ABIAS	0.0101	0.0099	0.0152	0.0096	0.0095	0.0145	0.0091	0.0089	0.0135
		ASE	0.0122	0.0122	0.0189	0.0109	0.0109	0.0169	0.0102	0.0102	0.0157
		ESE	0.0127	0.0124	0.0191	0.0118	0.0118	0.0180	0.0112	0.0110	0.0168
0.8	100	BIAS	0.0010	-0.0012	0.0011	0.0060	-0.0060	0.0052	0.0054	-0.0054	0.0039
		ABIAS	0.0178	0.0181	0.0266	0.0152	0.0150	0.0202	0.0140	0.0143	0.0192
		ASE	0.0209	0.0210	0.0326	0.0142	0.0142	0.0215	0.0130	0.0131	0.0196
		ESE	0.0223	0.0227	0.0335	0.0180	0.0178	0.0254	0.0167	0.0170	0.0240
	200	BIAS	0.0010	-0.0006	0.0017	0.0035	-0.0033	0.0031	0.0032	-0.0030	0.0027
		ABIAS	0.0121	0.0120	0.0192	0.0096	0.0092	0.0136	0.0091	0.0086	0.0125
		ASE	0.0150	0.0149	0.0232	0.0105	0.0105	0.0159	0.0098	0.0097	0.0146
		ESE	0.0153	0.0151	0.0238	0.0116	0.0111	0.0168	0.0109	0.0105	0.0153
	300	BIAS	0.0004	-0.0003	0.0006	0.0019	-0.0017	0.0020	0.0017	-0.0017	0.0023
		ABIAS	0.0098	0.0099	0.0152	0.0077	0.0074	0.0111	0.0071	0.0071	0.0104
		ASE	0.0122	0.0122	0.0189	0.0087	0.0087	0.0131	0.0081	0.0081	0.0121
		ESE	0.0123	0.0124	0.0192	0.0095	0.0092	0.0139	0.0087	0.0087	0.0132

Figure 1.

GAAD data: Left panel plots the proportion CAL responses, with the ‘pins’ at the extremes representing proportions of zeros (9.8%) and ones (8.1%). Right panel presents violin plots of the proportion responses across the four tooth-types, where the ‘white’ dots represent the respective medians, with the interquartile range denoted by the surrounding ‘black’ boxes.

Figure 2. Plot of estimated $g(\text{Age})$ (red line), along with the 95% pointwise confidence intervals (blue dotted line), obtained after fitting the semi-parametric model with the AR(1) correlation structure to the GAAD data.

Results from Simulation 1: Reported values are AIMSE (average integrated mean squared error) for Cases 1 & 2, when the true correlation structure is CS.

Table 1

		Case 1			Case 2		
ρ	n	WI	CS	AR(1)	WI	CS	AR(1)
0.5	100	0.0434	0.0394	0.0417	0.0576	0.0516	0.0556
	200	0.0341	0.0310	0.0322	0.0430	0.0379	0.0398
	300	0.0305	0.0280	0.0289	0.0370	0.0329	0.0343
0.8	100	0.0435	0.0318	0.0347	0.0567	0.0408	0.0448
	200	0.0344	0.0268	0.0284	0.0425	0.0313	0.0338
	300	0.0304	0.0250	0.0261	0.0372	0.0282	0.0301

Table 2

Results from Simulation 1: Here, BIAS = average bias, ABIAS = absolute bias, ASE = asymptotic standard error estimate (derived from the asymptotic normality results), and ESE = the empirical standard error estimates, calculated over 1000 replications, when the true correlation structure is compound symmetry (CS) for Case 1.

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.5	100	BIAS	0.0007	-0.0004	0.0013	0.0024	-0.0016	0.0020	0.0024	-0.0017	0.0020
		ABIAS	0.0121	0.0125	0.0192	0.0109	0.0113	0.0170	0.0117	0.0121	0.0183
		ASE	0.0149	0.0149	0.0233	0.0113	0.0113	0.0175	0.0119	0.0119	0.0185
		ESE	0.0151	0.0155	0.0237	0.0136	0.0140	0.0212	0.0145	0.0149	0.0228
	200	BIAS	0.0005	-0.0004	0.0003	0.0007	-0.0008	0.0010	0.0007	-0.0008	0.0005
		ABIAS	0.0092	0.0089	0.0134	0.0074	0.0076	0.0112	0.0082	0.0082	0.0117
		ASE	0.0106	0.0106	0.0166	0.0084	0.0084	0.0129	0.0089	0.0089	0.0138
		ESE	0.0115	0.0112	0.0168	0.0093	0.0095	0.0141	0.0102	0.0103	0.0148
	300	BIAS	-0.0001	-0.0002	0.0004	0.0005	-0.0005	0.0006	0.0006	-0.0006	0.0008
		ABIAS	0.0072	0.0074	0.0110	0.0060	0.0060	0.0089	0.0065	0.0065	0.0096
		ASE	0.0087	0.0087	0.0135	0.0070	0.0070	0.0107	0.0074	0.0074	0.0114
		ESE	0.0090	0.0091	0.0138	0.0074	0.0074	0.0113	0.0080	0.0080	0.0122
0.8	100	BIAS	0.0009	-0.0011	0.0005	0.0021	-0.0021	0.0018	0.0022	-0.0021	0.0017
		ABIAS	0.0124	0.0123	0.0194	0.0082	0.0083	0.0120	0.0095	0.0094	0.0139
		ASE	0.0149	0.0149	0.0233	0.0084	0.0084	0.0126	0.0094	0.0094	0.0143
		ESE	0.0156	0.0154	0.0244	0.0101	0.0102	0.0148	0.0118	0.0116	0.0173
	200	BIAS	0.0006	-0.0002	0.0000	0.0010	-0.0008	0.0007	0.0010	-0.0008	0.0003
		ABIAS	0.0091	0.0090	0.0134	0.0057	0.0054	0.0079	0.0065	0.0062	0.0095
		ASE	0.0106	0.0106	0.0166	0.0063	0.0063	0.0093	0.0071	0.0070	0.0107
		ESE	0.0113	0.0111	0.0171	0.0072	0.0068	0.0100	0.0081	0.0077	0.0120
300	BIAS	-0.0001	0.0000	0.0000	0.0000	0.0006	-0.0006	0.0005	0.0005	-0.0007	0.0004
		ABIAS	0.0071	0.0071	0.0111	0.0044	0.0044	0.0066	0.0051	0.0050	0.0077

ρ	n	Measure	WI			CS			AR(1)		
			β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
		ASE	0.0086	0.0086	0.0135	0.0052	0.0052	0.0077	0.0058	0.0058	0.0089
		ESE	0.0088	0.0089	0.0139	0.0055	0.0055	0.0082	0.0063	0.0062	0.0096

Results from Simulation 2: Reported values are AIMSE (average integrated mean squared error), derived from the fits of our QIF approach (under various correlation assumptions), and the GEE approach of He *et al.* (2005, *JASA*)

Table 3

ρ	n	CS					AR(1)				
		QIF-WI	QIF-CS	QIF-AR(1)	GEE		QIF-WI	QIF-CS	QIF-AR(1)	GEE	
0.5	100	0.0898	0.0799	0.0863	0.0843	0.0908	0.0866	0.0858	0.0881		
	200	0.0777	0.0706	0.0714	0.0740	0.0760	0.0726	0.0710	0.0739		
	300	0.0697	0.0648	0.0649	0.0672	0.0711	0.0684	0.0668	0.0696		
0.8	100	0.0901	0.0699	0.0720	0.0772	0.0918	0.0764	0.0747	0.0795		
	200	0.0752	0.0630	0.0632	0.0684	0.0769	0.0670	0.0634	0.0682		
	300	0.0726	0.0611	0.0616	0.0653	0.0703	0.0634	0.0611	0.0640		

Table 4

Results from Simulation 2: Reported values are RMSE (root mean squared error), derived from the fits of our QIF approach (under various correlation assumptions), and the GEE approach of He *et al.* (2005, *JASA*)

ρ	n	CS					AR(1)				
		QIF-WI	QIF-CS	QIF-AR(1)	GEE	GEE	QIF-WI	QIF-CS	QIF-AR(1)	GEE	GEE
0.5	100	0.0539	0.0414	0.0444	0.0438	0.0438	0.0552	0.0463	0.0455	0.0470	0.0470
	200	0.0424	0.0343	0.0350	0.0365	0.0365	0.0444	0.0384	0.0371	0.0396	0.0396
	300	0.0415	0.0339	0.0343	0.0354	0.0354	0.0403	0.0363	0.0354	0.0373	0.0373
0.8	100	0.0587	0.0354	0.0375	0.0385	0.0385	0.0555	0.0376	0.0365	0.0403	0.0403
	200	0.0434	0.0307	0.0314	0.0362	0.0362	0.0461	0.0335	0.0322	0.0347	0.0347
	300	0.0424	0.0301	0.0305	0.0348	0.0348	0.0426	0.0339	0.0321	0.0340	0.0340

Table 5

Parameter estimates, their standard errors (denoted by Std. Er), associated p-values, and TACM obtained after fitting our model with working independence (WI), compound symmetry (CS) and autoregressive with order 1 (AR(1)) correlation assumptions to the GAAD data.

	WI			CS			AR(1)		
	Estimate	Std. Er	p-value	Estimate	Std. Er	p-value	Estimate	Std. Er	p-value
Intercept	-2.4138	0.3886	<0.0001	-2.4138	0.3728	<0.0001	-2.1373	0.3703	<0.0001
Age	0.0372	0.0062	<0.0001	0.0372	0.0061	<0.0001	0.0325	0.0057	<0.0001
Gender	-0.5514	0.1680	0.0010	-0.5514	0.1680	0.0010	-0.4532	0.1562	0.0037
HbA1c	0.1444	0.1382	0.2961	0.1444	0.1382	0.2959	0.1160	0.1359	0.3935
Smoker	0.0346	0.1578	0.8264	0.0346	0.1577	0.8263	0.0152	0.1559	0.9222
Molar	1.8313	0.0773	<0.0001	1.8313	0.1001	<0.0001	1.7873	0.0753	<0.0001
Premolar	0.5886	0.0657	<0.0001	0.5886	0.0641	<0.0001	0.5492	0.0691	<0.0001
Canine	-0.2609	0.0461	<0.0001	-0.2609	0.0498	<0.0001	-0.2838	0.0445	<0.0001
TACM		0.2357			0.2278			0.2168	

Table 6

Reported QAIC, QBIC, MAFE and MAPE values for model without $g(\text{Age})$ (9), and model with $g(\text{Age})$. Both MAFE and MAPE are calculated, using the AR(1) working correlation matrix for our QIF method and unstructured working correlation matrix for the GEE. BETA denotes the random intercept Beta regression model.

	QIF				GEE				BETA			
	QAIC	QBIC	MAFE	MAPE	MAFE	MAPE	MAFE	MAPE	MAFE	MAPE	MAFE	MAPE
Model without $g(\text{Age})$	52.9611	78.65031	0.2453	0.2486	0.2458	0.2484	0.2462	0.2487	0.2462	0.2487	0.2462	0.2487
Model with $g(\text{Age})$	35.0369	79.07545	0.2310	0.2374	0.2312	0.2371	0.2313	0.2386	0.2313	0.2386	0.2313	0.2386

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 7
Regression coefficient estimates from fitting the semiparametric model to the GAAD data with AR(1) working

	Gender	HbA1c	Smoker	Molar	Premolar	Canine
Estimate	−0.4784	0.0836	0.0316	1.7998	0.5542	−0.2826
Std.	0.1523	0.1365	0.1592	0.0752	0.0712	0.0450
<i>p</i> -value	0.0017	0.5403	0.8428	<0.0001	<0.0001	<0.0001