Statistics and Its Interface Volume 17 (2024) 39–49

Sieve maximum likelihood estimation for generalized linear mixed models with an unknown link function

Mengdie Yuan and Guoqing Diao*

We study the generalized linear mixed models with an unknown link function for correlated outcome data. We propose sieve maximum likelihood estimation procedures by using B-splines. Specifically, we estimate the unknown link function in a sieve space spanned by the B-spline basis of the linear predictor that includes both the fixed and random terms. We establish the consistency and asymptotic normality of the proposed sieve maximum likelihood estimators. Extensive simulation studies, along with an application to an epileptic study, are provided to evaluate the finite-sample performance of the proposed method.

AMS 2000 subject classifications: Primary 62J12; secondary 62G08.
Keywords and phrases: B-splines, GLMM, Longitudinal data, Semiparametric models, Single index model.

1. INTRODUCTION

Canonical link functions are commonly used in practice for generalized linear models (GLMs) and generalized linear mixed models (GLMMs). However, there have been increasing concerns about the impact of misspecification of the link function on the inference procedure, say the consistency and robustness of the regression parameter estimators. [13] provided data analytic procedures to assess the adequacy of the assumed link function for GLMs. [1] proposed families of parametric link functions for binary outcomes by considering departures from the logistic model. [14] extended these ideas to other GLMs. However, these parametric models may still not be adequate in many applications [4]. Therefore, instead of assuming a known (or parametric) form of the link function, one can leave it completely unspecified and estimate it together with the mean model in GLMs. However, in this situation, it is challenging to estimate the unknown parameters, including an infinite-dimensional parameter in the unknown link function and the finite-dimensional regression coefficients. GLMs with unknown link functions have been well studied for cross-sectional data. Kernel-based method to estimate the unknown link function in GLMs has been studied by [18]. Other smoothing techniques without full model specification have also been considered in [8], [3] and [11], among others. GLMs have also been widely applied to longitudinal and/or correlated data.

For longitudinal data, we need to consider the variance-covariance structure to account for the within-subject correlations. Generally, there are two ways to fit the longitudinal data: marginal approach (such as GLMs) and conditional approach (such as GLMMs). GLMs construct the mean and variance-covariance structures separately. There are no subject-specific coefficients or random effects in the mean structure, and thus the inference is made on the population average. The aforementioned methodologies of dealing with GLMs with unknown link functions for cross-sectional data can also be applied to longitudinal data. [4] fitted the variance-covariance matrix as a function of the means where the function is unknown, then used the local polynomial kernel smoothing to estimate both the link function and variance-covariance function, and proposed "estimated estimating equations" (EEE) to obtain regression parameter estimators. [20] proposed profile-type estimating functions for the coefficients by applying the Kernel smoothing method to estimate the unknown link function.

An alternative method to estimate the unknown link function is to use the method of sieves, such as B-spline smoothing and truncated-power-spline smoothing, which is favorable due to its computational flexibility. [2] developed estimating equation-based procedures to estimate the unknown link function under the GLMs using the penalized truncated-power-spline smoothing. In fact, [7] compared the penalized B-spline smoothing and penalized truncated-power-spline smoothing. They found no advantage of the penalized truncated-power-spline smoothing over the penalized B-spline smoothing. On the other hand, GLMMs incorporate subject-specific random effects in the mean structure in addition to the fixed effects and allow the regression coefficients to vary across subjects through the random effects. One can also incorporate the random effects into the single index models, where we use a function of covariates to predict the outcome instead of a linear combination of fixed effects. Single index models are very useful for predictive modeling in various areas, such as econometrics and biometrics. [10], [9], and [12] have studied the single index

*Corresponding author.

models with random effects. However, for explanatory modeling, one may prefer to use regression coefficients instead of a function to interpret the relationship between the outcome and the covariates. Essentially no work on GLMM with an unknown link function is available in the literature.

In this article, we propose a sieve maximum likelihood estimation procedure to estimate the unknown parameters, including the regression coefficients, variance-covariance matrix of the random effects, and the unknown link function in GLMMs by using the B-spline smoothing. We derive the convergence rate of the sieve maximum likelihood estimators (MLEs) and establish the asymptotic normality and semiparametric efficiency of the sieve MLEs of the regression coefficients and the variance-covariance matrix of the random effects. Extensive simulation studies and an application to an epileptic study [17] are provided.

2. METHOD

Suppose there are n subjects in the study with k_i observations for the ith subject. Let Y_{ij} denote the jth outcome of the ith subject ($i = 1, \ldots, n; j = 1, \ldots, k_i$), $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp_x})^T$ and $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijp_z})^T$ denote, respectively, the corresponding p_x-dimensional and p_z-dimensional vectors of covariates. We assume that the outcome variable Y_{ij} has a probability distribution belonging to the exponential family. Specifically, the density of Y_{ij} in the general form of the exponential family is given by

$$f(y_{ij}) = \exp\left\{\frac{y_{ij}\phi_{ij} - b(\phi_{ij})}{a(\varphi)} - c(y_{ij}, \varphi)\right\}, \quad j = 1, \ldots, k_i, i = 1, \ldots, n,$$

where $a(\cdot), b(\cdot)$, and $c(\cdot)$ are known functions, and φ is a dispersion parameter, which may be either known or unknown. Here ϕ_{ij} contains covariates information and their associated parameters.

Consider the following generalized linear mixed model

$$g\left(E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)\right) = h(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i), \quad j = 1, \ldots, k_i, i = 1, \ldots, n, \tag{1}$$

where $g(\cdot)$ is a known transformation function, $h(\cdot)$ is a truncated function such that for two constants $-\infty < c_1 < c_2 < \infty$, $h(x) = h(c_1)$ if $x < c_1$ and $h(x) = h(c_2)$ if $x > c_2$, and $\boldsymbol{\beta}$ is a $p_x \times 1$ vector of unknown regression parameters Moreover, $\mathbf{b}_i$'s are p_z-dimensional subject-specific random vectors that are independently and identically distributed with a joint distribution ψ. We impose the technical condition on h because, unlike standard single index models, the linear predictor $\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i$ is not bounded. The transformation function $g(\cdot)$ can be taken as the canonical link function. For example, $g(x) = x$ for normal data, $g(x) = \log(x)$ for count data, and $g(x) = \text{logit}(x)$ for binary data. Notice that the actual link function is $h^{-1} \circ g(\cdot)$. To ensure the identifiability of the model, we do not include an intercept in the term $\mathbf{X}_{ij}^T\boldsymbol{\beta}$, and impose the constraint on $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\| = 1$ and $\beta_{p_x} > 0$, where $\|\cdot\|$ is the Euclidean norm. In this article, we assume $\mathbf{b}_i$ follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. We also assume $Y_{i1}, \ldots, Y_{ik_i}$ are mutually independent given $\mathbf{b}_i$. Let $\mu_{ij} = E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i)$. Then for the canonical link function, we have $g(\mu_{ij}) = \phi_{ij}$. We hereinafter refer to model (1) as the generalized linear mixed single index model (GLMSIM) while referring to the standard GLMM with a known link function simply as GLMM.

The conditional likelihood given $\mathbf{b}_i$ for the ith subject is given by

$$\begin{aligned} &L_{ci}(\boldsymbol{\beta}, h, \varphi|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i, \mathbf{b}_i) \\ &= \prod_{j=1}^{k_i} \exp\Big[\frac{1}{a(\varphi)}\{Y_{ij}h(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i) \\ &\quad - b\left(h(\mathbf{X}_{ij}^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\mathbf{b}_i)\right)\} - c(Y_{ij}, \varphi)\Big], \end{aligned}$$

where $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T, \ldots, \mathbf{X}_{ik_i}^T)^T$, $\mathbf{Z}_i = (\mathbf{Z}_{i1}^T, \mathbf{Z}_{i2}^T, \ldots, \mathbf{Z}_{ik_i}^T)^T$, and $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik_i})^T$. It follows that the log-likelihood given the observed data $\{(\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i), i = 1, \ldots, n\}$ takes the form

$$l_n(\boldsymbol{\beta}, h, \varphi, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log \int L_{ci}(\boldsymbol{\beta}, h, \varphi|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i, \mathbf{b}_i)\psi(\mathbf{b}_i)d\mathbf{b}_i,$$

where $\psi(\cdot)$ is the multivariate normal density function with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Consider a bounded interval (c_1, c_2). We also consider K_n interior knots $v_1, v_2, \ldots, v_{K_n}$ with $K_n = O(n^\nu)$ and $\max_{0 \le j \le K_n} |v_{j+1} - v_j| = O(n^{-\nu})$, for $\nu \in (0, 0.5)$, and $c_1 = v_0 \le v_1 \le v_2 \le \cdots \le v_{K_n} \le v_{K_n+1} = c_2$. Let $\mathcal{S}_n(\mathbf{v}, K_n, M)$ denote the space of polynomial splines of order M defined in [15], where $\mathbf{v} = (v_0, v_1, \ldots, v_{K_n+1})$. Then there exists a local basis $\{B_j : 1 \le j \le K_n + M\}$ such that for any $s \in \mathcal{S}_n(\mathbf{v}, K_n, M)$, we have

$$s(t) = \sum_{j=1}^{K_n+M} \gamma_j B_{M,j}(t) = \boldsymbol{\gamma}^T\mathbf{B}_M(t),$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_{K_n+M})^T$ are the smoothing coefficients and $B_M(t) = \left(B_{M,1}(t), B_{M,2}(t), \ldots, B_{M,K_n+M}(t)\right)^T$ are the M-degree B-spline basis functions. Under some smoothness assumptions, the unknown univariate function $h(\cdot)$ can be well approximated by an M-degree B-spline in $\mathcal{S}_n(\mathbf{v}, K_n, M)$, that is,

$$h(t) \approx \begin{cases} \sum_{j=1}^{K_n+M} \gamma_j B_{M,j}(t) = \boldsymbol{\gamma}^T\mathbf{B}_M(t), & t \in [c_1, c_2]; \\ \boldsymbol{\gamma}^T\mathbf{B}_M(c_1), & t < c_1; \\ \boldsymbol{\gamma}^T\mathbf{B}_M(c_2), & t > c_2. \end{cases}$$

Replacing $h(t)$ by $\boldsymbol{\gamma}^T\widetilde{\mathbf{B}}_M(t)$, where $\widetilde{\mathbf{B}}_M(t) = \mathbf{B}_M(c_1)I(t<c_1)+\mathbf{B}_M(t)I(c_1\le t\le c_2)+\mathbf{B}_M(c_2)I(t>c_2)$, in the conditional likelihood function, we obtain

$$\begin{aligned}&L_{ci}(\boldsymbol{\beta},\boldsymbol{\gamma},\varphi|\mathbf{X}_i,\mathbf{Z}_i,\mathbf{Y}_i,\mathbf{b}_i)\\&=\prod_{j=1}^{k_i}\exp\Bigg[\frac{1}{a(\varphi)}\Big\{Y_{ij}\boldsymbol{\gamma}^T\widetilde{\mathbf{B}}_M(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b}_i)\\&\quad-b\left(\boldsymbol{\gamma}^T\widetilde{\mathbf{B}}_M(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b}_i)\right)\Big\}-c(Y_{ij},\varphi)\Bigg].\end{aligned}$$

Similarly, we can obtain the observed-data log-likelihood, denoted by $l_n(\boldsymbol{\beta},\boldsymbol{\gamma},\varphi,\boldsymbol{\Sigma})$. In order to estimate the unknown parameters, we need to maximize the observed-data log-likelihood. Notice that there is no closed form for the observed-data log-likelihood in general except for normal responses. We suggest using the Gauss-Hermite quadrature to approximate the log-likelihood function numerically. Let $\boldsymbol{\sigma} = (\sigma_1,\sigma_2,\ldots,\sigma_r)$ be a $r\times 1$ vector of parameters that contain all the unknown parameters involved in the variance-covariance matrix $\boldsymbol{\Sigma}$. To accommodate the constraint $\|\boldsymbol{\beta}\| = 1$ and $\beta_{p_x} > 0$, we use the following reparameterization

$$\beta_j = \frac{\alpha_j}{\sqrt{\sum_{l=1}^{p_x}\alpha_l^2}},\quad j=1,\ldots,p_x,$$

where $\alpha_{p_x} = 1$. Define $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\sigma})$ if φ is known and $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\sigma},\varphi)$ if φ is unknown. The score equations are then given by

$$\frac{\partial l_n(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}.$$

We use an iterative procedure to estimate the unknown parameters $\boldsymbol{\theta}$. The resulting sieve MLEs are denoted by $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\alpha}}_n,\hat{\boldsymbol{\gamma}}_n,\hat{\boldsymbol{\sigma}}_n)$ or $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\alpha}}_n,\hat{\boldsymbol{\gamma}}_n,\hat{\varphi}_n,\hat{\boldsymbol{\sigma}}_n)$ for known or unknown φ, respectively. Then $\hat{\boldsymbol{\beta}}_n$, the sieve MLE of $\boldsymbol{\beta}$, is determined by $\hat{\boldsymbol{\alpha}}_n$, and $\hat{\boldsymbol{\Sigma}}_n$, the sieve MLE of $\boldsymbol{\Sigma}$, is determined by $\hat{\boldsymbol{\sigma}}_n$. Furthermore, the sieve MLE of $h(\cdot)$ is $\hat{h}_n(t) = \hat{\boldsymbol{\gamma}}_n\widetilde{\mathbf{B}}_M(t)$. Notice that the likelihood involves the value of the basis functions at $\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i$, $j=1,\ldots,k_i, i=1,\ldots,n$. We recommend choosing the knots based on the equally spaced sample quantiles of $\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b}_k^*$, $j=1,\ldots,k_i, i=1,\ldots,n, k=1,\ldots,N_q^{p_z}$, where $N_q^{p_z}$ is the number of quadrature points and $\mathbf{b}_k^*$ are the kth abscissas of the Gauss-Hermite quadrature. As a result, the knots and basis change as $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ change. We describe the detailed algorithm in simulation studies.

To make statistical inferences of $\boldsymbol{\beta}$, φ, and $\boldsymbol{\Sigma}$, the distribution of $(\hat{\boldsymbol{\beta}}_n,\hat{\varphi}_n,\hat{\boldsymbol{\sigma}}_n)$ can be approximated by a multivariate normal distribution when the sample size is large enough. The variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_n,\hat{\varphi}_n,\hat{\boldsymbol{\sigma}}_n)$ can be estimated by the inverse of the observed information matrix under the efficient score function of $(\boldsymbol{\beta},\varphi,\boldsymbol{\sigma})$. We show that the asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_n,\hat{\varphi},\hat{\boldsymbol{\sigma}}_n)$ attains the semiparametric efficiency bound under the correct model specification. In the event of model misspecification, we may consider the sandwich estimator of the variance-covariance matrix to improve robustness. [6] suggested using the observed information matrix by taking into account the parameter $\boldsymbol{\gamma}$. In our simulation, we used the inverse matrix of the estimated variance-covariance matrix of the score functions by taking into account all the unknown parameters ($\boldsymbol{\beta}$, φ, $\boldsymbol{\sigma}$, $\boldsymbol{\gamma}$). Extensive simulations suggest that the variance-covariance estimator works well.

3. ASYMPTOTIC PROPERTIES

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0,\varphi_0,\boldsymbol{\Sigma}_0,h_0)$ denote the true parameters. Before establishing the asymptotic properties of the sieve MLEs, we impose the following regularity conditions.

C1. The true values $(\boldsymbol{\beta}_0,\varphi_0,\boldsymbol{\Sigma}_0)$ for parameters $(\boldsymbol{\beta},\varphi,\boldsymbol{\Sigma})$ belong to the interior of a compact set

$$\begin{aligned}\mathcal{B}_0 = \{&(\boldsymbol{\beta},\varphi,\boldsymbol{\Sigma}):\boldsymbol{\beta}\in\mathcal{R}^p,\varphi\in\mathcal{R}^+,\beta_{p_x}>0,\|\boldsymbol{\beta}\|=1,\\&\text{and }\boldsymbol{\Sigma}\text{ is positive definite and its eigenvalues}\\&\text{are bounded away from 0 and }\infty\}.\end{aligned}$$

C2. The covariate vectors $\mathbf{X}$ and $\mathbf{Z}$ are bounded almost surely, and both $E(\mathbf{X}\mathbf{X}^T)$ and $E(\mathbf{Z}\mathbf{Z}^T)$ are positive definite.

C3. The number of observations k for each subject is random. Moreover, there exists a positive integer k_0 such that $1\le k\le k_0$ and $Pr(k\ge 2)>0$.

C4. The true function $h_0(\cdot)$ for h belongs to $\mathcal{H}^q$, where the functional space $\mathcal{H}^q$ is the collection of all bounded functions h on a bounded interval $[c_1,c_2]$ with bounded jth derivative $h^{(j)}$, $j=1,\ldots,k$, such that $h^{(k)}$ satisfies the Lipschitz continuity condition with exponent m $(0<m\le 1)$:

$$|h^{(k)}(s)-h^{(k)}(t)|\le L|s-t|^m,\ \text{for } s,t\in[c_1,c_2],$$

where $L<\infty$ is a positive constant, and $q=k+m\ge 3$.

C5. For some $\eta\in(0,1)$, $u^TVar(\mathbf{X}|\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})u \ge \eta u^TE(\mathbf{X}\mathbf{X}^T|\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})u$ and $u^TVar(\mathbf{Z}|\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})u\ge\eta u^TE(\mathbf{Z}\mathbf{Z}^T|\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})u$ almost surly for all $u\in\mathcal{R}^p$.

C6. $E[\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})\}^2\mathbf{X}\mathbf{X}^T]$ and $E[\{h_0'(\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})\}^2\mathbf{Z}\mathbf{Z}^T]$ are nonsingular.

Remark 3.1. Condition C1 is a common regularity assumption in the literature. Conditions C1 and C2 ensure the identifiability of the model. The restriction $q\ge 3$ in Condition C4 is needed to provide desirable control for the B-spline approximation error rates of h_0 as well as the first and second derivatives of h_0 for the proof of the normality. Conditions C5 and C6 are technical assumptions and can be justified in many applications. Condition C3 implies that the number of observations for each subject is bounded and some of them have at least two observations. Note that

although the domain of $\mathbf{X}^T\boldsymbol{\beta}+\mathbf{Z}^T\mathbf{b}$ is $(-\infty,\infty)$, we focus on the estimation and inference of h on a bounded interval $[c_1,c_2]$. In practice, we may choose c_1 and c_2 such that $P(\mathbf{X}_{ij}^T\boldsymbol{\beta}+\mathbf{Z}_{ij}^T\mathbf{b}_i\in(c_1,c_2))$ is greater than a pre-specified threshold, e.g., 99%. In addition, Condition C6 ensures that the semiparametric efficiency information matrix is invertible and is also necessary to ensure model identifiability.

Let $\mathcal{H}_n^q=\mathcal{H}^q\cap\mathcal{S}_n$. Clearly, we have $\mathcal{H}_n^q\subseteq\mathcal{H}_{n+1}^q\subseteq\cdots\subseteq\mathcal{H}^q$ for all $n\geq 1$. Denote $\Theta_n^q=\mathcal{B}_0\times\mathcal{H}_n^q$ as the sieve space. The sieve MLEs are the maximizer of the log-likelihood function over the sieve space Θ_n^q. Define the norm $\|\cdot\|_2$ over the space $\mathcal{H}^q$ as

$$\|h\|_2=\left[E\{h^2(\mathbf{X}^T\boldsymbol{\beta}_0+\mathbf{Z}^T\mathbf{b})\}\right]^{\frac{1}{2}}.$$

Let $\|\cdot\|_\infty$ denote the supremum norm. We define the distance over the space $\Theta^q=\mathcal{B}_0\times\mathcal{H}^p$ as follows

$$\begin{aligned}d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=&(\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|^2+|\varphi_1-\varphi_2|^2\\&+\|\boldsymbol{\sigma}_1-\boldsymbol{\sigma}_2\|^2+\|h_1-h_2\|_2^2)^{\frac{1}{2}},\end{aligned}$$

for $\boldsymbol{\theta}_1=(\boldsymbol{\beta}_1,\varphi_1,\boldsymbol{\sigma}_1,h_1),\boldsymbol{\theta}_2=(\boldsymbol{\beta}_2,\varphi_2,\boldsymbol{\sigma}_2,h_2)\in\Theta^q$.

Recall that $K_n=O(n^\nu)$. We restrict ν to be in $(\frac{1}{2(1+q)},\frac{1}{2q})$, where q is the smoothness parameter in C4. The following theorem establishes the convergence rate of the sieve MLEs $\hat{\boldsymbol{\theta}}_n=(\hat{\boldsymbol{\beta}}_n,\hat{\varphi}_n,\hat{\boldsymbol{\sigma}}_n,\hat{h}_n)$ to the true parameter, with a slight abuse of notation, still denoted by $\boldsymbol{\theta}_0$.

Theorem 3.1. Under the Conditions C1–C6,

$$d(\hat{\boldsymbol{\theta}}_n,\boldsymbol{\theta}_0)=O_p(n^{-\min(q\nu,\frac{1-\nu}{2})}).$$

The outlines of the proofs of Theorem 3.1 and the two subsequent theorems concerning the large sample distributions of the sieve MLEs of the finite-dimensional parameters are deferred to the Appendix.

Before describing the next two theorems, we introduce some notation in the context of empirical processes. Let $\mathbb{P}_n$ and $\mathbb{P}$ be the empirical measure and the population distribution of n i.i.d. observations $\mathbf{O}_1,\mathbf{O}_2,\ldots,\mathbf{O}_n$ for $\mathbf{O}=(\mathbf{Y},\mathbf{X},\mathbf{Z})$, $i=1,\ldots,n$. Let $\mathbb{G}_n=\sqrt{n}(\mathbb{P}_n-\mathbb{P})$ denote the empirical process. Then for any measurable function $h(\mathbf{O})$,

$$\mathbb{P}_nh(\mathbf{O})=\frac{1}{n}\sum_{i=1}^n h(\mathbf{O}_i),\mathbb{P}h(\mathbf{O})=E_{\mathbb{P}}[h(\mathbf{O})],$$

and

$$\mathbb{G}_nh(\mathbf{O})=\sqrt{n}(\mathbb{P}_nh(\mathbf{O})-\mathbb{P}h(\mathbf{O})).$$

The following theorem establishes the asymptotic normality of the sieve MLEs of the finite-dimensional parameters, denoted by $\boldsymbol{\zeta}\equiv(\boldsymbol{\beta},\varphi,\boldsymbol{\sigma})$. Their corresponding true values and sieve MLEs are denoted by $\boldsymbol{\zeta}_0$ and $\hat{\boldsymbol{\zeta}}_n$, respectively.

Theorem 3.2. Under the Conditions C1–C6,

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\zeta}}_n-\boldsymbol{\zeta}_0)\to N(\mathbf{0},\mathbf{I}^{-1}(\boldsymbol{\zeta}_0))$$

in distribution, where $\mathbf{I}(\boldsymbol{\zeta}_0)=\mathbb{P}\{\tilde{l}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_0;\mathbf{O})^{\otimes 2}\}$, and $\tilde{l}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}_0;\mathbf{O})$ is the efficient score function of $\boldsymbol{\zeta}_0$ based on observations $\mathbf{O}$ from a generic subject.

Finally, a consistent estimator for the asymptotic variance-covariance matrix is provided in the following theorem.

Theorem 3.3. Under the Conditions C1–C6,

$$\mathbb{P}_n\left\{\tilde{l}_{\boldsymbol{\zeta}}(\hat{\boldsymbol{\theta}}_n;\mathbf{O})^{\otimes 2}\right\}\to\mathbb{P}\left\{\tilde{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0;\mathbf{O})^{\otimes 2}\right\}$$

in probability, where $\tilde{l}_{\boldsymbol{\zeta}}(\hat{\boldsymbol{\theta}}_n;\mathbf{O})$ is the plug-in estimator of $\tilde{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0;\mathbf{O})$.

4. SIMULATION STUDIES

We conduct simulation studies to examine the finite-sample performance of the proposed method. In the study, we consider the longitudinal count data. We include three covariates ($p_x=3$) X_1, X_2 and X_3 in the model. Specifically, X_1 is generated from a normal distribution with mean 1 and variance 0.5^2, X_2 is generated from the uniform distribution $U(-1,2.25)$, and X_3 is generated from a Bernoulli distribution with success probability 0.5. We set the true values $\boldsymbol{\beta}_0=\frac{1}{\sqrt{3}}(1,1,1)$ and $g(t)=\log(t)$. We include a random intercept in the model that follows $N(0,0.6^2)$. The outcome Y is generated from a Poisson distribution with mean $\exp\{h(\mathbf{X}^T\boldsymbol{\beta}_0+b)\}$. In the simulations, we consider two different scenarios for $h(\cdot)$: (i) $h(t)=t$; (ii) $h(t)=\cos(t)$. For scenario (i), we divide X_1, X_2 and X_3 by 2. The log-likelihood is approximated by the Gauss-Hermite quadrature with $N_q=10$ quadrature points. Suppose $a_k,k=1,\ldots,N_q$ are the abscissas. Let σ denote the standard deviation of the random intercept b. Then we adopt three data-adaptive interior knots that were placed at the equally spaced quantiles of $\mathbf{X}_{ij}^T\boldsymbol{\beta}+\sqrt{2}\sigma a_k$, $j=1,\ldots,k_i,i=1,\ldots,n,k=1,\ldots,N_q$ given $\boldsymbol{\beta}$ and σ, and employ the cubic ($M=4$) B-spline to approximate the unknown function. We obtain the sieve MLEs iteratively by using the following iterative procedures.

(1) Choose the initial values $\boldsymbol{\beta}$ and σ.
(2) Given $\boldsymbol{\beta}$, σ and the abscissas $a_k,k=1,\ldots,N_q$ of the Gauss-Hermite quadrature, calculate the knots and basis functions.
(3) Given the knots and basis functions, obtain $\hat{\boldsymbol{\gamma}}$ by solving $\partial l_n(\boldsymbol{\beta},\sigma,\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}=\mathbf{0}$.
(4) Given the knots, basis functions and $\hat{\boldsymbol{\gamma}}$, obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ by solving $\partial l_n(\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{\gamma})/\partial\boldsymbol{\beta}=\mathbf{0}$ and $\partial l_n(\boldsymbol{\beta},\sigma,\boldsymbol{\gamma})/\partial\sigma=0$.
(5) Set the initial values for next iteration as $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$, $\sigma=\hat{\sigma}$.

Table 1. Summary Statistics of the proposed sieve MLEs with $h(t) = t$ for Poisson data

n	Parameter	True	Bias	SE	SEE	MSE	CP
120	β_1	0.577	−0.0161	0.146	0.158	0.0217	0.938
	β_2	0.577	−0.0091	0.110	0.119	0.0121	0.953
	β_3	0.577	−0.0226	0.145	0.158	0.0215	0.940
	σ	0.600	−0.0201	0.139	0.140	0.0197	0.932
200	β_1	0.577	−0.0042	0.113	0.116	0.0128	0.938
	β_2	0.577	−0.0137	0.089	0.089	0.0080	0.930
	β_3	0.577	−0.0113	0.113	0.117	0.0129	0.941
	σ	0.600	−0.0223	0.096	0.102	0.0097	0.920

True is the true value for the parameter; Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; SEE is the average of the standard error estimates; MSE is the mean squared error; and CP is the coverage probability of the 95% confidence interval estimates.

(6) Repeat steps (2)–(5) until the maximum of the component-wise absolute value difference of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ between two consecutive iterations is less than 10^{-3}.

The maximization in steps (3) and (4) is carried out by the quasi-Newton algorithm. In our experience, the above algorithm converges quickly and is reasonably robust to the choices of the initial estimates.

We estimate the variance-covariance matrix for the regression parameter $\boldsymbol{\beta}$ using the inverse of the consistent variance-covariance matrix estimator for the score functions. We compare the proposed sieve MLEs with the sieve MLEs based on model (1) but without random effects, referred to as the generalized linear single index model (GL-SIM). We also consider the standard Poisson GLMM with a random intercept and a canonical log link function. For fair comparisons, we standardize the estimates of the regression parameters (excluding the fixed intercept) of the GLMM such that the Euclidean norm is 1. For each case, we considered the sample sizes $n = 120$, 200 with 5 observations for each subject. The results are based on 1000 replicates.

Tables 1 and 3 present the summary statistics of the sieve MLEs of the unknown regression parameters $\boldsymbol{\beta}$ and σ using our method with $h(t) = t$ and $h(t) = \cos(t)$, respectively. In these tables, Bias is the sampling average of the biases of the estimates; SE denotes the empirical standard deviations of the parameter estimates; SEE denotes the average of the standard error estimates; CP is the coverage probability of 95% confidence interval estimates; and MSE is the mean squared error of the parameter estimates. We observe that the biases of the proposed sieve MLEs are small under all simulation settings. The standard error estimates exhibit the true variation well. The coverage probabilities of the 95% confidence interval estimates are close to the nominal level.

Tables 2 and 4 present the results from the GL-SIM and the Poisson GLMM with a random intercept and a log link function. Notice that the results for σ under the GL-SIM are not applicable. The column RE is the MSE relative efficiency, defined as the ratio of the MSE of the competing estimators and the MSE of the proposed sieve MLEs. In Table 2, the relative efficiencies for all parameters compared to the GL-SIM are greater than 1. These results are expected because the GL-SIM ignores the within-subject correlations. As expected, the proposed sieve MLEs are less efficient than the MLEs from the GLMM with the correctly specified link function in most cases, with the efficiency loss of less than 10%. The loss of efficiency is caused by estimating the unknown link function in our method. We also notice that when the sample size increases, the relative efficiency compared to the GLMM improves. The sieve MLEs appear to be more efficient with $h(t) = \cos(t)$ in Table 4 compared to both the estimators based on GL-SIM and the standard GLMM. Additionally, the estimates under the GLMM in Table 4 have larger biases compared to those of the proposed estimators in Table 3.

5. EXAMPLE

In this section, we apply the proposed method to an epileptic data study [17]. The objective of this study is to explore the treatment/drug effects on patients' seizures. Fifty-nine epileptics were enrolled in this study. The numbers of seizures for each patient suffering from epileptic episodes were recorded at the baseline, then followed every two weeks for an eight-week period. Patients in this study were randomized to the test group to receive the drug Progabide (Trt = 1) or the control group to receive a placebo (Trt = 0). Besides the treatment, two additional covariates are also of interest, including Age in years and Base in the logarithm of the baseline counts divided by 4. Variable Age was log-transformed.

In our analysis, we include Age, Trt, Base and the interaction Trt*Base as covariates, and evaluate the covariate

Table 2. Comparison of the proposed sieve MLEs with the sieve MLEs assuming independence (GL-SIM) and the MLEs from the GLMM with $h(t) = t$ for Poisson data

		GL-SIM				GLMM			
n	Parameter	Bias	SE	MSE	RE	Bias	SE	MSE	RE
120	β_1	−0.0449	0.194	0.0396	1.83	−0.0164	0.142	0.0203	0.94
	β_2	−0.0080	0.146	0.0214	1.77	−0.0063	0.108	0.0117	0.97
	β_3	−0.0317	0.189	0.0367	1.71	−0.0222	0.139	0.0198	0.92
	σ					−0.0050	0.125	0.0157	0.80
200	β_1	−0.0225	0.139	0.0198	1.54	−0.0046	0.111	0.0123	0.96
	β_2	−0.0070	0.112	0.0127	1.58	−0.0098	0.090	0.0082	1.01
	β_3	−0.0163	0.142	0.0205	1.59	−0.0141	0.111	0.0125	0.97
	σ					−0.0022	0.100	0.0099	1.03

Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the indicating model.

Table 3. Summary statistics of the proposed sieve MLEs with $h(t) = \cos(t)$ for Poisson data

n	Parameter	True	Bias	SE	SEE	MSE	CP
120	β_1	0.577	−0.0128	0.091	0.093	0.0084	0.952
	β_2	0.577	0.0024	0.067	0.068	0.0045	0.948
	β_3	0.577	−0.0073	0.086	0.090	0.0075	0.946
	σ	0.600	−0.0095	0.116	0.106	0.0135	0.942
200	β_1	0.577	−0.0074	0.068	0.071	0.0046	0.961
	β_2	0.577	0.0031	0.050	0.052	0.0025	0.953
	β_3	0.577	−0.0056	0.066	0.069	0.0043	0.958
	σ	0.600	−0.0026	0.081	0.079	0.0065	0.942

True is the true value for the parameter; Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; SEE is the average of the standard error estimates; MSE is the mean squared error; and CP is the coverage probability of the 95% confidence interval estimates.

Table 4. Comparison of the proposed sieve MLEs with the sieve MLEs assuming independence (GL-SIM) and the MLEs from the GLMM with $h(t) = \cos(t)$ for Poisson data

		GL-SIM				GLMM			
n	Parameter	Bias	SE	MSE	RE	Bias	SE	MSE	RE
120	β_1	−0.0129	0.104	0.0110	1.31	−0.0433	0.106	0.0131	1.56
	β_2	0.0001	0.077	0.0060	1.33	0.0282	0.079	0.0070	1.55
	β_3	−0.0110	0.102	0.0106	1.40	−0.0116	0.103	0.0116	1.43
	σ					−0.0125	0.123	0.0154	1.14
200	β_1	−0.0073	0.079	0.0062	1.34	−0.0436	0.085	0.0092	1.97
	β_2	−0.0003	0.058	0.0033	1.31	0.0316	0.064	0.0051	2.02
	β_3	−0.0055	0.075	0.0056	1.29	−0.0063	0.082	0.0068	1.58
	σ					−0.0260	0.097	0.0100	1.54

Bias is the bias of the parameter estimate; SE is the empirical standard deviation of the parameter estimates; MSE is the mean squared error; and RE is the relative efficiency of the proposed sieve MLEs with the estimators under the indicating model.

Table 5. Results for the analysis of the epileptic data

	Est	SE	P-value	Est	SE	P-value
	New			New_0		
Age	0.478	0.291	1.00E-01	0.622	0.032	<1.00E-06**
Trt	−0.297	0.168	7.70E-02	−0.245	0.034	<1.00E-06**
Base	0.766	0.184	3.05E-05**	0.688	0.039	<1.00E-06**
Trt*Base	0.310	0.253	2.20E-01	0.282	0.045	<1.00E-06**
σ	0.448	0.108	1.67E-05**			
	GLMM			GLM		
Intercept	1.795	0.104	<1.00E-06**	1.860	0.041	<1.00E-06**
Age	0.481	0.346	1.64E-01	0.888	0.117	<1.00E-06**
Trt	−0.334	0.147	2.33E-02**	−0.346	0.061	<1.00E-06**
Base	0.883	0.131	<1.00E-06**	0.949	0.044	<1.00E-06**
Trt*Base	0.339	0.202	9.40E-02	0.562	0.064	<1.00E-06**
σ	0.501	0.058	<1.00E-06**			

** indicates significant effect at the 5% significance level.

effects using the proposed method with a random intercept. For comparison, we also consider three other methods: a) the GLM based on the Poisson distribution with an unknown link function (GL-SIM); b) the standard GLM based on the Poisson distribution with a log link function; and c) the standard GLMM based on the Poisson distribution with a random intercept and a log link function. The parameter estimates are summarized in Table 5 for the four different approaches. The results under the GLM-SIM and the GL-SIM do not include the intercept because of the constraint on the regression parameters for the model identifiability. The standard GLM and GLMM were fitted by routines glm() and glmer() in R, respectively.

The estimate of the standard error σ of the random intercept and the standard error estimate for the estimator are 0.448 and 0.108 under the GLM-SIM, and are 0.501 and 0.058 under the GLMM. Both models detected significant within-subject correlations. By comparing the Wald test statistic for testing $H_0 : \sigma = 0$ against a half-half mixture of a point mass at 0 and a χ^2_1 distribution, we obtain the p-values of 1.67E−05 and $<$ 1.0E−6 for the GLM-SIM and the GLMM, respectively. As expected, the tests based on the standard GLM and the GL-SIM tend to be liberal in the presence of within-subject correlations. At the significance level of 0.05, the proposed method detected a significant Base effect whereas the GLMM detected both a significant Base effect and a significant treatment effect. The directions of the estimates are the same from these two models with a random intercept. Notice that the interpretation of the parameter estimates that are obtained by our method is intrinsically tied to the link function.

Figure 1 presents the estimated curve $\hat{h}_n(t)$ in the region of $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n + \hat{b}_{in}$, $i = 1, \ldots, n, j = 1, \ldots, k_i$, where $\hat{b}_{in} = E(b_i|\mathbf{Y}_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_n)$ is the predictor of the random intercept for the ith subject given the observed data and the sieve MLEs $\hat{\boldsymbol{\theta}}_n$. It appears that $\hat{h}_n(t)$ is close to a straight line within the interval $[-2, 2]$ suggesting a link function close to the canonical log link function in the Poisson GLMM. To check the fit of the proposed method to the epileptic data, in Figure 2, we also plot the model-fitted mean curve $\exp\{\hat{h}_n(t)\}$ and compare it with the observed data as well as the empirically estimated mean curve. The model-fitted mean curve and the empirically estimated mean curve agree very well, indicating a good fit of the GLM-SIM.

6. DISCUSSION

We have proposed a sieve maximum likelihood estimation approach for the GLMMs with an unknown link function using B-spline smoothing and have established the asymptotic properties of the proposed sieve MLEs. The sieve MLEs of the regression coefficients and the variance-covariance matrix of the random effects achieve the semiparametric efficiency bound. The simulation studies demonstrate that the GLM-SIM outperforms the standard GLMMs when the link function is misspecified and there is little loss of efficiency with a correctly specified link function. We have applied our method to the epileptic data. A model-checking procedure of comparing the model-fitted mean curve with the empirically estimated mean curve suggests that the proposed method can handle the longitudinal data well and yield satisfying results. The proposed methods are implemented in a C program that is available upon request.

The GLMMs are useful in exploratory modeling. It is often of interest to interpret the model by increasing one unit in one covariate. Usually the interpretation can be straightforward with the canonical link. Discussion on the interpretation of a nonparametric link function can be found in [5] and [4]. One of the potential issues with using the splines method to approximate the unknown function is the well-known undersmoothing problem. A penalty term can be introduced into the likelihood to control for the potential undersmoothing problem, which is known as penalized splines (P-splines) smoothing. As future work, it would be interesting to develop an estimation procedure based on the

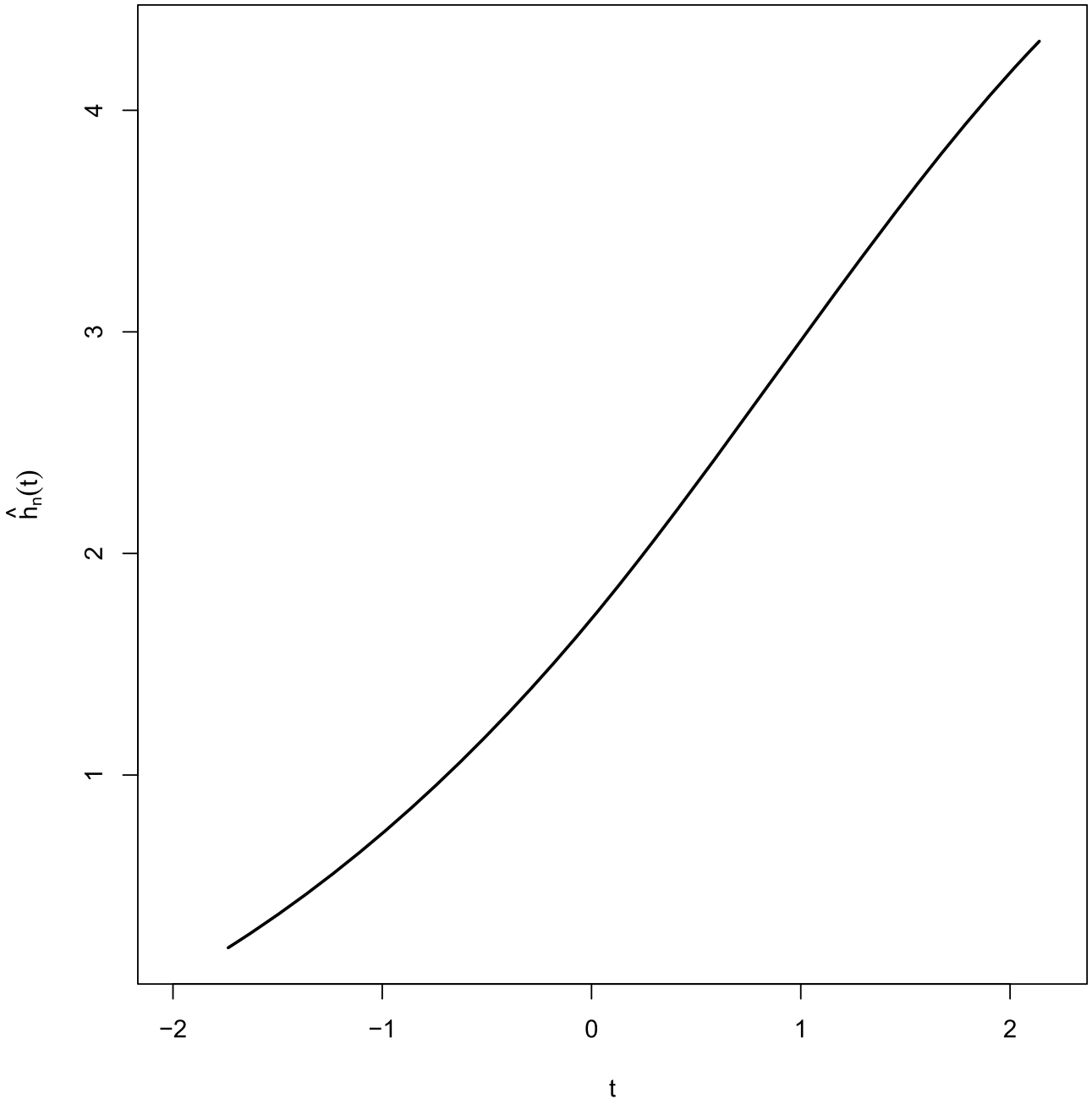

Figure 1. Estimated curve $\hat{h}_n(t)$ in the region of $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n + \hat{b}_{in}$, $i = 1, \ldots, n, j = 1, \ldots, k_i$, where $\hat{b}_{in} = E(b_i|\mathbf{Y}_i, \mathbf{X}_i, \hat{\boldsymbol{\theta}}_n)$ is the predictor of the random intercept for the ith subject given the observed data and the sieve MLEs $\hat{\boldsymbol{\theta}}_n$.

penalized likelihood and compare the results. Note that our methods can not only provide the estimation for the coefficients, but also obtain an estimation for the link function. Therefore, it is feasible to further develop a formal procedure to test $H_0 : h'(t) = c, \forall t \in [c_1, c_2]$ for some constant c to detect if the canonical link function is deviated from the true link function. Future research is warranted toward this direction.

We impose some conditions (C4 in Section 3) on the true link function h. Under these conditions, we establish the consistency and convergence rate of the estimator of h. Similar conditions have been imposed in the literature to use B-splines to approximate an unknown function. However, the estimator of h may be biased if these conditions are violated. It warrants future research to examine the performance of the proposed methods when the conditions on h are not satisfied.

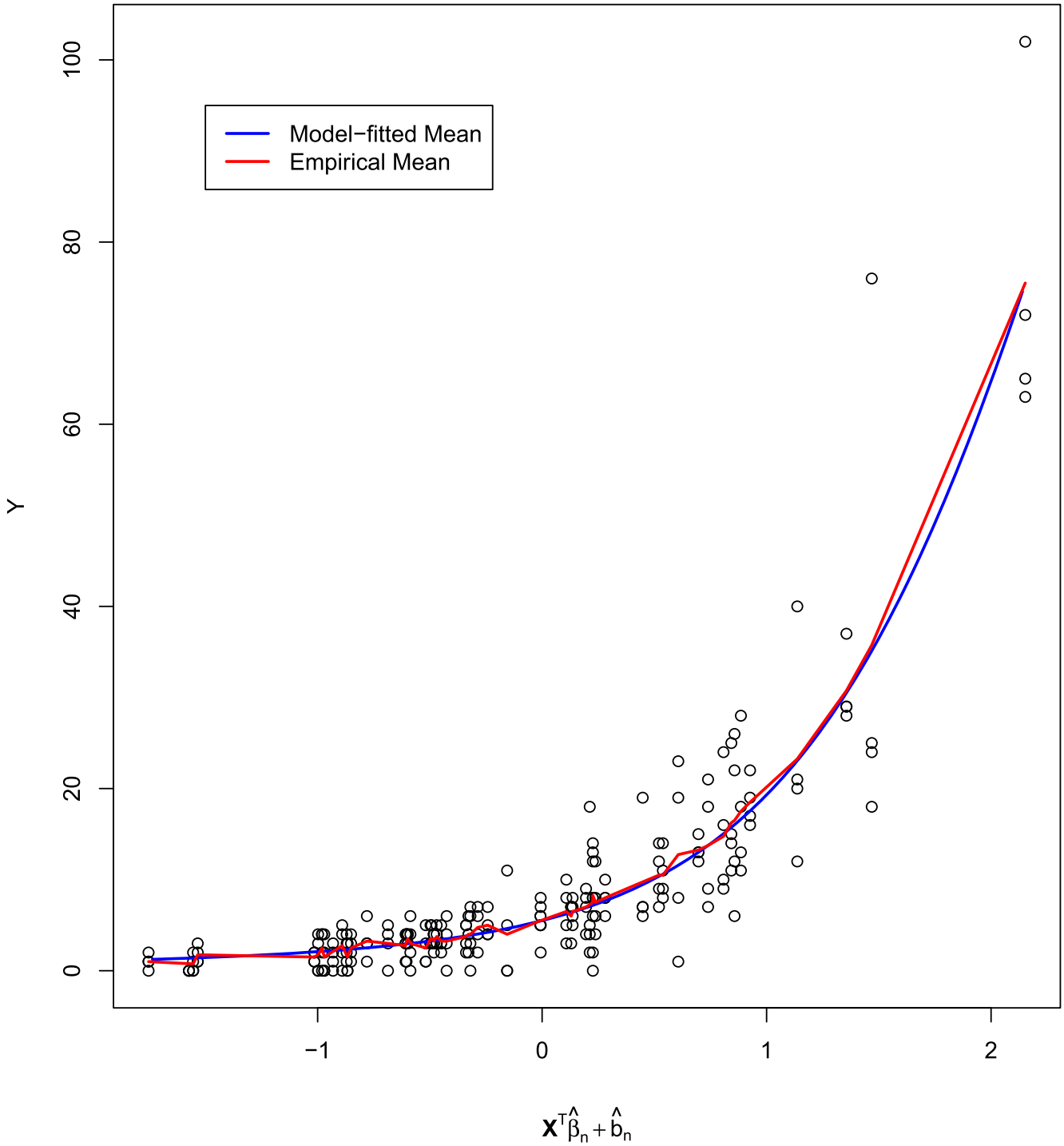

Figure 2. Model fitted mean curve $\exp\{\hat{h}_n(t)\}$ *(blue solid curve) and the empirical estimated mean curve (red solid curve). The circles correspond to the observed data* Y_{ij} *against* $\mathbf{X}_{ij}^T\hat{\boldsymbol{\beta}}_n + \hat{b}_{in}$, $i = 1, \ldots, n, j = 1, \ldots, k_i$.

APPENDIX

A.1 Proof outline for Theorem 3.1

The convergence rate of $\hat{\boldsymbol{\theta}}_n$ is obtained by verifying the assumptions in Theorem 1 in [16]. According to [6], the sieve space $\mathcal{H}_n^q$ does not have to be restricted to $\mathcal{S}_n$ as long as the estimator $\hat{\boldsymbol{\theta}}_n$ satisfies the following assumptions corresponding to those in Theorem 1 in [16].

1) $\inf_{\{d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\geq\varepsilon,\boldsymbol{\theta}\in\Theta_n^p\}} \mathbb{P}(l(\boldsymbol{\theta}_0;\mathbf{O}) - l(\boldsymbol{\theta};\mathbf{O})) \gtrsim \varepsilon$.

2) $\mathbb{P}\{l(\boldsymbol{\zeta}, h; \mathbf{O}) - l(\boldsymbol{\zeta}_0, h_0; \mathbf{O})\}^2 \le C_2(\|\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\|^2 + \|h - h_0\|_2^2)$.
3) Let $\boldsymbol{\theta}_{0,n} = (\boldsymbol{\zeta}_0, h_{0,n})$, and $\mathcal{F}_n = \{l(\boldsymbol{\theta}; \mathbf{O}) - l(\boldsymbol{\theta}_{0,n}; \mathbf{O}) : \boldsymbol{\theta} \in \Theta_n^p\}$, where

$$\|h_{0,n} - h_0\|_\infty = O(q_n^{-q}) = O(n^{-q\nu}).$$

Then the L_∞-metric entropy of the space $\mathcal{F}_n$ satisfies

$$\begin{aligned} H(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) &= \log N_{[]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \\ &\le c_7(cq_n + p)\log(1/\varepsilon) \le c_8 n^\nu \log(1/\varepsilon). \end{aligned}$$

With these three conditions verified, it then follows from Theorem 1 in [16] that

$$\begin{aligned} d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) &= O_p\left(\max(n^{-\frac{1-\nu}{2}}, n^{-q\nu}, n^{-q\nu})\right) \\ &= O_p(n^{-\min(q\nu, \frac{1-\nu}{2})}). \end{aligned}$$

We first consider model (1) without random effects, i.e., the GL-SIM,

$$g(\mu_i) = h(\mathbf{X}_i^T \boldsymbol{\beta}), \quad i = 1, \ldots, n. \tag{2}$$

In other words, we take $\boldsymbol{\Sigma} = \mathbf{0}$ in model (1), in which case the integral with respect to $\mathbf{b}$ in the likelihood is reduced to a fixed point at $\mathbf{0}$. It is easy to see the three assumptions are satisfied for model (2).

We then generalize the results to model (1) where the individual likelihood is integrated over the distribution of the random effects. In fact, the integral of a bounded function against a probability measure is still bounded. In the methodology, $h(\cdot)$ is defined on a bounded interval. To make the integral meaningful, we extended $h(\cdot)$ to the whole real line with $h(x) = 0$ if x is outside the bounded interval. Then under Conditions C1–C4, the likelihood and all of its derivatives above are continuous and bounded. Due to this fact, the properties of the log-likelihood and its derivatives remain the same. In addition, Condition C1 indicates that the eigenvalues of $\boldsymbol{\Sigma}_0$ are also bounded. Therefore, the conditions in [16] can be verified.

A.2 Proof outline for Theorem 3.2

To prove Theorem 3.2, we use Theorem 6.1 in [19]. Specifically, we need to verify the consistency and rate of convergence, positive information, stochastic equicontinuity of the estimators and the smoothness of the model. The key idea is that we need to construct ε-baskets in which the target functions can be bounded by ε times a positive number. Following the proof of Theorem 6.1 in [19], we only need to verify the following assumptions.

(i) $d(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta}_0) = O_p(n^{-\delta})$ for some $\delta > 0$.
(ii) $\mathbb{P}\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O}) = 0$ and $\mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h] = 0$ for all $h \in \mathbf{H}$.
(iii) There exists an $\mathbf{h}^* = (h_1^*, h_2^*, \ldots, h_{(p_x + p_z^2)^*})^T$, where $h_j^* \in \mathbf{H}$ for $j = 1, \ldots, p_x + p_z^2$, such that $\mathbb{P}\ddot{l}_{\boldsymbol{\zeta}_h}(\boldsymbol{\theta}_0; \mathbf{O})[h] - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*, h] = \mathbf{0}$, for all $h \in \mathbf{H}$. Furthermore, the matrix $\mathbb{P}\{\ddot{l}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O}) - \ddot{l}_{h\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\}$ is nonsingular.
(iv) $\mathbb{P}_n\dot{l}_{\boldsymbol{\zeta}}(\hat{\boldsymbol{\theta}}_n, \mathbf{O}) = o_p(n^{-\frac{1}{2}})$ and $\mathbb{P}_n\dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[\mathbf{h}^*] = o_p(n^{-\frac{1}{2}})$.
(v) For any $c > 0$,

$$\begin{aligned} &\sup_{\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le cn^{-\delta}, \boldsymbol{\theta} \in \Theta_n^p\}} \left|\mathbb{G}_n\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}; \mathbf{O}) - \mathbb{G}_n\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O})\right| \\ &= o_p(1), \end{aligned}$$

and

$$\begin{aligned} &\sup_{\{d(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \le cn^{-\delta}, \boldsymbol{\theta} \in \Theta_n^p\}} \left|\mathbb{G}_n\dot{l}_g(\boldsymbol{\theta}; \mathbf{O})[\mathbf{h}^*] - \mathbb{G}_n\dot{l}_g(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\right| \\ &= o_p(1). \end{aligned}$$

(vi) For some $\zeta > 1$ and $\delta\zeta > 1/2$, consider a neighborhood of $\boldsymbol{\theta}_0$:

$$\left\{\boldsymbol{\theta}_c : |\boldsymbol{\zeta}_c - \boldsymbol{\zeta}_0| + \|h_c - h_0\|_2 \le cn^{-\delta}\right\}.$$

Then

$$\begin{aligned} &\left|\mathbb{P}\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}; \mathbf{O}) - \mathbb{P}\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O}) - \mathbb{P}\ddot{l}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O})(\boldsymbol{\zeta}_c - \boldsymbol{\zeta}_0)\right. \\ &\left. - \mathbb{P}\ddot{l}_{\boldsymbol{\zeta}h}(\boldsymbol{\theta}_0; \mathbf{O})[h_c - h_0]\right| \\ &= O\left((|\boldsymbol{\zeta}_c - \boldsymbol{\zeta}_0| + \|h_c - h_0\|_2)^\zeta\right), \end{aligned}$$

and

$$\begin{aligned} &\left|\mathbb{P}\dot{l}_h(\boldsymbol{\theta}; \mathbf{O})[\mathbf{h}^*] - \mathbb{P}\dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]\right. \\ &- \mathbb{P}\ddot{l}_{h\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O})(\boldsymbol{\zeta} - \boldsymbol{\zeta}_0)[\mathbf{h}^*](\boldsymbol{\zeta} - \boldsymbol{\zeta}_0) \\ &\left. - \mathbb{P}\ddot{l}_{hh}(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*, h - h_0]\right| \\ &= O\left((|\boldsymbol{\zeta}_c - \boldsymbol{\zeta}_0| + \|h_c - h_0\|_2)^\zeta\right). \end{aligned}$$

Notice that the term $\tilde{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O})$ in Theorem 3.2 is given by $\dot{l}_{\boldsymbol{\zeta}}(\boldsymbol{\theta}_0; \mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[\mathbf{h}^*]$.

Again it is easier to first show the asymptotic normality of the sieve MLEs for Model (2) by verifying the above conditions. To extend the result to model (1), again we use the fact that the integral of a bounded function over a probability space is still bounded, therefore will not change the fact that the above conditions are still satisfied, which provides the proof of Theorem 3.2.

A.3 Proof outline for Theorem 3.3

We first can prove the theorem for model (2), in which case $\boldsymbol{\zeta}$ is reduced to $\boldsymbol{\xi} \equiv (\boldsymbol{\beta}, \varphi)$. Define

$$\begin{aligned} I_{jk}(\boldsymbol{\theta}_0) &= \mathbb{P}\Big[\Big\{\dot{l}_{\xi_j}(\boldsymbol{\theta}_0; \mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h_j^*]\Big\} \\ &\quad \times \Big\{\dot{l}_{\xi_k}(\boldsymbol{\theta}_0; \mathbf{O}) - \dot{l}_h(\boldsymbol{\theta}_0; \mathbf{O})[h_k^*]\Big\}\Big] \\ &= \mathbb{P}A_{jk}(\boldsymbol{\theta}_0; \mathbf{O}) \end{aligned}$$

and

$$\begin{aligned} I_{jkn}(\hat{\boldsymbol{\theta}}_n) &= \mathbb{P}_n\Big[\Big\{\dot{l}_{\xi_j}(\hat{\boldsymbol{\theta}}_n; \mathbf{O}) - \dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_j^*]\Big\} \\ &\quad \times \Big\{\dot{l}_{\xi_k}(\hat{\boldsymbol{\theta}}_n; \mathbf{O}) - \dot{l}_h(\hat{\boldsymbol{\theta}}_n; \mathbf{O})[h_k^*]\Big\}\Big] \\ &= \mathbb{P}_n A_{jkn}(\hat{\boldsymbol{\theta}}_n; \mathbf{O}). \end{aligned}$$

Then

$$\begin{aligned} I_{jkn}(\hat{\boldsymbol{\theta}}_n) - I_{jk}(\boldsymbol{\theta}_0) &= \mathbb{G}_n A_{jkn}(\hat{\boldsymbol{\theta}}_n; \mathbf{O}) \\ &\quad + \mathbb{P}\Big\{A_{jkn}(\hat{\boldsymbol{\theta}}_n; \mathbf{O}) - A_{jk}(\boldsymbol{\theta}_0; \mathbf{O})\Big\} \\ &= I_{1n} + I_{2n}. \end{aligned}$$

Using the similar argument in the verification of the assumption (iv) in the proof of the asymptotic normality, we can show $I_{1n} = o_p(1)$ and $I_{2n} = o_p(1)$. This is for model (2). Using the similar boundedness arguments, we obtain the consistency of the variance-covariance matrix estimator for model (1).

ACKNOWLEDGMENT

The authors thank Dr. Wanli Qiao for reading the manuscript and providing valuable comments and suggestions. The authors also thank the Associate Editor and two anonymous referees for their valuable comments that have improved the presentation of the paper.

Received 15 December 2022

REFERENCES

[1] Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* **68** 357–363. MR0626394

[2] Bai, Y., Fung, W. K., and Zhu, Z. Y. (2009). Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis* **100**, 152–161. MR2460484

[3] Chiou, J. M. and Müller, H. G. (1998). Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association* **93**, 1376–1387. MR1666634

[4] Chiou, J. M. and Müller, H. G. (2005). Estimated estimating equations: semiparametric inference for clustered and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 531–553. MR2168203

[5] Climov, D., Hart, J., and Simar, L. (2002). Automatic smoothing and estimation in single index poisson regression. *Journal of Nonparametric Statistics* **14**, 307–323. MR1905754

[6] Ding, Y. and Nan, B. (2011). A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of statistics* **39**, 2795. MR3012400

[7] Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 637–653.

[8] Fan, J. Q., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141–150. MR1325121

[9] Gu, C. and Ma, P. (2005). Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. *Journal of Computational and Graphical Statistics* **14**, 485–504. MR2161625

[10] Karcher, P. and Wang, Y. (2001). Generalized nonparametric mixed effects models. *Journal of Computational and Graphical Statistics* **10**, 641–655. MR1938972

[11] Muggeo, V. M. R. and Ferrara, G. (2007). Fitting generalized linear models with unspecified link function: A p-spline approach. *Computational Statistics and Data Analysis* **52**, 2529–2537. MR2411956

[12] Pang, Z. and Xue, L. (2012). Estimation for the single-index models with random effects. *Computational Statistics & Data Analysis* **56**, 1837–1853. MR2892381

[13] Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* **9**, 705–724. MR0619277

[14] Scallan, A., Gilchrist, R., and Green, M. (1984). Fitting parametric link functions in generalised linear models. *Computational Statistics and Data Analysis* **2**, 37–49.

[15] Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. Wiley New York. MR0606200

[16] Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* **22**, 580–615. MR1292531

[17] Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671. MR1085814

[18] Weisberg, S. and Welsh, A. H. (1994). Adapting for the missing link. *The Annals of Statistics* **22**, 1674–1700. MR1329165

[19] Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics* **35**, 2106–2142. MR2363965

[20] Xu, P. and Zhu, L. (2012). Estimation for a marginal generalized single-index longitudinal model. *Journal of Multivariate Analysis* **105**, 285–299. MR2877518

Mengdie Yuan
Department of Statistics
George Mason University
Fairfax, Virginia
U.S.A.
E-mail address: myuan2@gmu.edu

Guoqing Diao
Department of Biostatistics and Bioinformatics
George Washington University
Washington, District of Columbia
U.S.A.
E-mail address: gdiao@gwu.edu