Ann Inst Stat Math (2017) 69:1099–1127
DOI 10.1007/s10463-016-0575-8

Efficient estimation of quasi-likelihood models using B-splines

Minggen Lu[1]

Received: 30 May 2014 / Revised: 7 June 2016 / Published online: 3 August 2016

Abstract We consider a simple yet flexible spline estimation method for quasi-likelihood models. We approximate the unknown function by B-splines and apply the Fisher scoring algorithm to compute the estimates. The spline estimate of the nonparametric component achieves the optimal rate of convergence under the smooth condition, and the estimate of the parametric part is shown to be asymptotically normal even if the variance function is misspecified. The semiparametric efficiency of the model can be established if the variance function is correctly specified. A direct and consistent variance estimation method based on the least-squares estimation is proposed. A simulation study is performed to evaluate the numerical performance of the spline estimate. The methodology is illustrated on a crab study.

Keywords B-spline · Least-squares estimation · Quasi-likelihood model · Semiparametric efficiency

1 Introduction

Quasi-likelihood method (Wedderburn 1974; McCullagh 1983; McCullagh and Nelder 1989) is used to model the relationship between an outcome and some covariates in cases where the exact distributional information is not available. It only requires the specification of a relationship between the mean and variance of the outcome. Quasi-likelihood methods have the similar properties to the maximal likelihood methods and thus are reasonable alternatives if the distribution of the outcome is not fully available. In parametric quasi-likelihood linear models, it is assumed that the unknown mean

✉ Minggen Lu
minggenl@unr.edu

[1] School of Community Health Sciences, University of Nevada, Reno, NV 89557, USA

function is modeled linearly via a known link function. In many practical situations, however, the underlying relationship between the response and covariates is not adequately fit by a linear function or simple parametric curves. Some components can be indeed highly nonlinear. The linear assumption may lead to substantial modeling bias and wrong conclusion. A natural extension of the quasi-likelihood linear models is to allow some covariates to be linearly associated with the response, with other covariates being modeled nonlinearly.

The statistical methodology for semiparametric quasi-likelihood estimation has been extensively discussed in the literature and can be classified into kernel smoothing (Severin and Staniswallis 1994; Härdle et al. 1998), penalized estimation (Mammen and van der Geer 1997), and local polynomial fitting (Fan et al. 1995; Fan and Chen 1999; Chen et al. 2006). Splines are well known for their numerical stability and good approximation to smooth functions, and their application to semiparametric estimation has been extensively studied, for example, Stone (1986), Kooperberg et al. (1995), Huang and Liu (2006), Lu et al. (2009), and Hua and Zhang (2012) among many others. In this manuscript, we consider the spline-based M-estimator for the unknown function ψ, which can be classified as sieve estimation. In sieve estimation, instead of maximizing a given criterion function over the whole parameter space, a sequence of increasing subspaces (sieves) that depend on the sample size n are used to approximate the large original space such that the resulting estimation problem becomes computationally less complicated. In spline quasi-likelihood estimation, the sieves are the classes of cubic B-splines and the original space is the class of bounded smooth functions.

We approximate the unknown function ψ by a cubic B-spline function:

$$\psi(z) \approx \psi_n(z) = \sum_{j=1}^{q_n} \gamma_j B_j(z).$$

The attraction of the spline estimation is that after the spline basis functions are chosen, the approximated function is totally characterized by the spline coefficients $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q_n})^\mathsf{T}$. The regression parameters $\boldsymbol{\beta}$ and the spline coefficients $\boldsymbol{\gamma}$ can then be estimated simultaneously by maximizing the spline quasi-likelihood function. Therefore, the computational burden is greatly alleviated and the standard Newton–Raphson or Fisher scoring method can be effectively adapted to the spline quasi-likelihood estimation. The proposed B-spline method is easier to implement and less computationally intensive, compared with the aforementioned existing methods. The local polynomial fitting method and the kernel estimation method are based on the profile likelihood, in which there are two (iterative) steps involved in estimating the regression parameters and the nonparametric function. The obvious extra difficulty introduced by the penalized likelihood approach is that the smoothing parameters have to be estimated by some computationally intensive approaches, such as (generalized) cross-validation. In addition to the preferable numerical properties, the proposed B-spline method also has some appealing asymptotic features. First, by allowing the number of knots to increase by the sample size at an appropriate rate, the spline estimate of ψ is uniformly consistent and can achieve the optimal rate of convergence

under the smooth condition. Second, the estimates of $\boldsymbol{\beta}$ are asymptotically normal even if the variance function is possibly misspecified, and the spline-based semiparametric model can achieve the asymptotic efficiency for $\boldsymbol{\beta}$ if the variance function is correctly specified. Finally, a consistent variance estimation method for the estimate of $\boldsymbol{\beta}$ can be derived by taking advantage of the spline approximation.

The rest of the paper is organized as follows. The spline quasi-likelihood estimator and the Fisher scoring algorithm are presented in Sect. 2 . An adaptive knots selection method is also discussed in Sect. 2. Asymptotic properties of the estimators are studied in Sect. 3. A direct and consistent variance estimation method is provided in Sect. 4. A Monte Carlo simulation study and an illustrative example are given in Sect. 5. Finally, the proofs of asymptotic results are sketched in Sect. 6.

2 Spline quasi-likelihood estimation

2.1 Models

Let $\{(y_i, \mathbf{x}_i, z_i) : i = 1, \ldots, n\}$ denote the independent copies of $(y, \mathbf{x}, z)$, where y is a scalar response variable, $\mathbf{x} \in \mathbb{R}^d$, and $z \in \mathbb{R}$. Denote $\mathbf{w} = (\mathbf{x}^\mathsf{T}, z)^\mathsf{T}$ and $\mathbf{v} = (\mathbf{w}^\mathsf{T}, y)^\mathsf{T}$. Assume

$$\mu(\mathbf{w}; \boldsymbol{\beta}, \psi) = E(y|\mathbf{w}) = F(\mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \psi(z)), \tag{1}$$

where F is a known monotone function, $\boldsymbol{\beta}$ is an unknown $d \times 1$ parameter vector, and ψ is an unknown function. Assume further that the conditional variance of y only depends on $\sigma^2 V(\mu)$, where σ^2 is an unknown parameter and $V(\mu)$ is a known positive function. Because in practice the information about the variance function is not always available, we relax the assumption that $V(\mu)$ is correctly specified. Thus, our results may be used in case of model misspecification. Denote $\boldsymbol{\tau} = (\boldsymbol{\beta}^\mathsf{T}, \psi)^\mathsf{T}$. The quasi-likelihood function is defined as follows:

$$Q(\boldsymbol{\tau}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - s}{\sigma^2 V(s)} \mathrm{d}s, \tag{2}$$

where $\mu_i \equiv \mu(\mathbf{w}_i; \boldsymbol{\tau}) = F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \psi(z_i))$.

We assume that z takes values in $[a, b]$, where a and b are finite numbers. Let $\mathcal{T}_n = \{t_i\}_1^{m_n+2l}$, with

$$a = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{m_n+l} < t_{m_n+l+1} = \cdots = t_{m_n+2l} = b,$$

be a sequence of knots that subdivide the interval $[a, b]$ into $m_n + 1$ subintervals $I_i = [t_{l+i}, t_{l+1+i}]$, $i = 0, \ldots, m_n$. The spline of order $l \geq 1$ with the knot sequence $\mathcal{T}_n$ is a polynomial of degree $l-1$ within any subinterval $[t_{l+i}, t_{l+i+1}]$. A spline of $l = 4$ is a piecewise cubic polynomial with continuous second order derivative. Let $\mathcal{S}_n(\mathcal{T}_n, l)$ be the class of splines of order $l \geq 1$ with knots $\mathcal{T}_n$. According to Corollary 4.10 of Schumaker (1981), $\mathcal{S}_n(\mathcal{T}_n, l)$ can be linearly spanned by spline basis functions, that is,

for any $s \in \mathcal{S}_n(\mathcal{T}_n, l)$, there exists a set of B-spline basis functions $\{B_j : 1 \le j \le q_n\}$ such that $s = \sum_{j=1}^{q_n} \gamma_j B_j$, where $q_n = m_n + l$ is the number of basis functions.

If $\psi(z)$ is smooth enough, we can approximate $\psi(z)$ by a B-spline function $\psi_n(z) \in \mathcal{S}_n$:

$$\psi(z) \approx \psi_n(z) = \sum_{j=1}^{q_n} \gamma_j B_j(z).$$

Replacing $\psi(z)$ by $\psi_n(z)$ in (2), we obtain the spline quasi-likelihood function for $\boldsymbol{\vartheta} = (\boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\gamma}^\mathsf{T})^\mathsf{T}$, namely,

$$Q(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \int_{y_i}^{\bar{\mu}_i} \frac{y_i - s}{\sigma^2 V(s)} \mathrm{d}s, \tag{3}$$

where $\bar{\mu}_i \equiv \mu(\mathbf{w}_i; \boldsymbol{\vartheta}) = F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \mathbf{b}_i^\mathsf{T}\boldsymbol{\gamma})$ and $\mathbf{b}_i = (B_1(z_i), \ldots, B_{q_n}(z_i))^\mathsf{T}$. The advantage of this reparametrization is that we can estimate the regression parameters $\boldsymbol{\beta}$ and the spline coefficients $\boldsymbol{\gamma}$ simultaneously, and hence release the computational burden. Let $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\beta}}^\mathsf{T}, \hat{\boldsymbol{\gamma}}^\mathsf{T})^\mathsf{T}$ be the values that maximize the spline quasi-likelihood function (3). The spline estimator of $\psi(z)$ is defined as $\sum_{j=1}^{q_n} \hat{\gamma}_j B_j(z)$.

2.2 Computation of the estimates

Let $\xi(\mathbf{w}; \boldsymbol{\vartheta}) = f(\mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \mathbf{b}^\mathsf{T}\boldsymbol{\gamma})$, where f is the first derivative of F. Denote $\xi_i = \xi(\mathbf{w}_i; \boldsymbol{\vartheta})$ and $V_i = V(\bar{\mu}_i)$, $i = 1, \ldots, n$. Let $\mathbf{X}^\mathsf{T} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{B}^\mathsf{T} = (\mathbf{b_1}, \ldots, \mathbf{b_n})$. Denote $\mathbf{V} = \mathrm{diag}\{V_1, \ldots, V_n\}$ and $\boldsymbol{\xi} = \mathrm{diag}\{\xi_1, \ldots, \xi_n\}$. Let $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$, $\bar{\boldsymbol{\mu}} = (\bar{\mu}_1, \ldots, \bar{\mu}_n)^\mathsf{T}$, and $\mathbf{D} = (\mathbf{X}, \mathbf{B})$. Some calculation yields the score vector

$$\nabla Q(\boldsymbol{\vartheta}) = \sigma^{-2}\mathbf{D}^\mathsf{T}\boldsymbol{\xi}\mathbf{V}^{-1}(\mathbf{y} - \bar{\boldsymbol{\mu}})$$

and the expected information matrix

$$\mathbf{E}(\boldsymbol{\vartheta}) = \sigma^{-2}\mathbf{D}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{D}.$$

We now give some expressions that will be used in the variance estimation. Let

$$\mathbf{D}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{D} = \begin{pmatrix} \mathbf{X}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{B} \\ \mathbf{B}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{X} & \mathbf{B}^\mathsf{T}\boldsymbol{\xi}^2\mathbf{V}^{-1}\mathbf{B} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}.$$

It follows from the formula of block matrix inverse that

$$\begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{E}^{11} & \mathbf{E}^{12} \\ \mathbf{E}^{21} & \mathbf{E}^{22} \end{pmatrix} = \begin{pmatrix} \mathbf{E}_{11.2}^{-1} & -\mathbf{E}_{11.2}^{-1}\mathbf{E}_{12}\mathbf{E}_{22}^{-1} \\ -\mathbf{E}_{22.1}^{-1}\mathbf{E}_{21}\mathbf{E}_{11}^{-1} & \mathbf{E}_{22.1}^{-1} \end{pmatrix},$$

where $\mathbf{E}_{11.2} = \mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{E}_{22}^{-1}\mathbf{E}_{21}$ and $\mathbf{E}_{22.1} = \mathbf{E}_{22} - \mathbf{E}_{21}\mathbf{E}_{11}^{-1}\mathbf{E}_{12}$.

Remark 1 For $F(s) = s$ and $V(\mu) = 1$, the quasi-likelihood function reduces to the log-likelihood function for normally distributed data. We can explicitly estimate $\boldsymbol{\vartheta}$ by

$$\hat{\boldsymbol{\vartheta}} = (\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\mathbf{y}. \tag{4}$$

Remark 2 Let $F(s) = \exp(s)$ and $V(\mu) = \sigma^2\mu$, $0 < \mu < \infty$. For $\sigma^2 = 1$, the quasi-likelihood function becomes the Poisson log-likelihood function. Otherwise, the resulting likelihood function is the log-likelihood function with data according to a Poisson distribution with over-dispersion (under-dispersion) parameter σ^2.

Because σ^2 is a nuisance parameter, and the estimation of $\boldsymbol{\vartheta}$ is independent of the estimator of σ^2 and depends only on the first two moments of the outcome, we propose to estimate σ^2 by the moment estimation method, namely,

$$\hat{\sigma}^2 = \frac{1}{n - d - q_n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{5}$$

where $\hat{\mu}_i = F(\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}} + \sum_{j=1}^{q_n}\hat{\gamma}_j B_j(z_i))$. The correction term $1/(n-d-q_n)$ is important for bias adjustment when $d + q_n$ is relatively large compared with n.

We apply a modified Fisher scoring method to simultaneously calculate the spline estimate $\hat{\boldsymbol{\vartheta}}$. The iterative optimization procedure is described as follows:
Step 1 Use (4) to obtain the initial value $\boldsymbol{\vartheta}^{(0)}$.
Step 2 Apply the Fisher scoring method to update $\boldsymbol{\vartheta}^{(k)}$ in the kth iteration

$$\boldsymbol{\vartheta}^{(k)} = \boldsymbol{\vartheta}^{(k-1)} + [E(\boldsymbol{\vartheta}^{(k-1)})]^{-1}\nabla Q(\boldsymbol{\vartheta}^{(k-1)}).$$

Repeat the iteration until the convergence criterion

$$\|\boldsymbol{\vartheta}^{(k)} - \boldsymbol{\vartheta}^{(k-1)}\| < \varepsilon = 10^{-6}$$

is met.

Remark 3 Step 2 is equivalent to performing iteratively weighted linear regression of $\mathbf{D}\boldsymbol{\vartheta} + \boldsymbol{\xi}^{-1}(\mathbf{y} - \bar{\boldsymbol{\mu}})$ on $\mathbf{D}$ with weight $\boldsymbol{\xi}^2\mathbf{V}^{-1}$.

Remark 4 In our simulation study and real data analysis, because the outcome is the count, we use $\hat{\boldsymbol{\vartheta}} = (\mathbf{D}^{\mathsf{T}}\mathbf{D})^{-1}\mathbf{D}^{\mathsf{T}}\log(1 + \mathbf{y})$ as the initial value. This approach leads to good performance for most cases in our numerical study. In some cases where the algorithm does not converge due to the singularity of the information matrix, the Levenberg–Marquardt method is applied to remedy the algorithm.

2.3 Knot selection

It is well known that the performance of spline estimation depends on the selection of knots sequence $\mathcal{T}_n$. The methodology for selecting the knots of splines has been extensively discussed in literature, for example, Wahba (1990), Stone et al. (1997), and Ruppert (2002). We adopt the quantile method (Rosenbeg 1995) and some model selection criteria, such as Akaike information criteria (AIC), to adaptively select knots. Assume the true function ψ_0 has rth continuous derivative, $r \geq 1$. According to Theorem 2 in Sect. 3, the number of knots is chosen to be of order $n^{1/(1+2r)}$ to achieve the optimal rate of convergence of $\hat{\psi}$. Therefore, we choose the number of inner knots from a neighborhood of $n^{1/(1+2r)}$, such as $\left[0.5N_r, \min(4N_r, n^{1/2})\right]$, where $N_r = \text{ceiling}(n^{1/(1+2r)})$. In our simulation study and real data analysis, r is chosen to be 1. The optimal number of interior knots, m_n^*, is chosen to minimize the AIC value

$$\text{AIC}(m_n) = -2Q(\hat{\boldsymbol{\vartheta}}; m_n) + 2(m_n + l + d),$$

where l is the order of spline and $m_n + l$ is the number of B-spline basis functions. For a given number of interior knots m_n, the interior knots $t_{l+k}, k = 1, \ldots, m_n$, correspond to the $k/(m_n + 1)$ quantile of z. The similar method was used in Xue and Liang (2010) and Lu and Loomis (2013).

3 Asymptotic properties of the estimators

3.1 Assumptions

Denote by $\boldsymbol{\tau}_0 = (\boldsymbol{\beta}_0^\intercal, \psi_0)^\intercal$ the true value of $\boldsymbol{\tau} = (\boldsymbol{\beta}^\intercal, \psi)^\intercal$. Let the regression parameter space Θ be the interior of some compact set in $\mathbb{R}^d$, and let

$$\Psi = \{\psi : \text{the } r\text{th derivative of } \psi \text{ is Lipschitz on a compact subset } \Im,\ r \geq 3\}$$

be the nonparametric space. Let $\|\cdot\|$ and $\|\cdot\|_2$ be the Euclidean norm of $\mathbb{R}^d$ and L_2-norm, respectively. Also let $\|\cdot\|_\infty$ denote the supremum norm. Define L_2-norm $\|\cdot\|_2$ on $\Theta \times \Psi$ as follows:

$$\|\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1\|_2^2 = \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2 + \|\psi_2 - \psi_1\|_2^2.$$

The following regularity conditions are assumed to derive the asymptotic properties of the spline estimators:

- C1. The maximum spacing of the knots is assumed to be $O(n^{-\nu})$, $0 < \nu < 1/2$. Moreover, the ratio of maximum and minimum spacing of knots is uniformly bounded.
- C2. The true parameter $\boldsymbol{\tau}_0$ is in the interior of $\Theta \times \Psi$.
- C3. The support of z is an interval within $\Im$ and the second moment of z is finite.
- C4. The vector $\mathbf{x}$ takes values in a convex set $\mathcal{X} \subset \mathbb{R}^d$ and the fourth moment of $\mathbf{x}$ is finite.

C5. Let $\mathcal{M}$ be a compact set of $\mathbb{R}$ such that $\mathbf{x}^\intercal\boldsymbol{\beta} + \psi(z) \in \mathcal{M}$, for all $z \in \Im$, $\psi \in \Psi$, $\mathbf{x} \in \mathcal{X}$, and $\boldsymbol{\beta} \in \Theta$, and let $\mathcal{F} = F(\mathcal{M})$. The variance function $V(\mu)$ is bounded away from 0 and ∞ on $\mathcal{F}$.
C6. For $k = 1, 2$, the derivatives $\partial^k V(\mu)/\partial\mu^k$ exist and are bounded on $\mathcal{F}$. For $l = 1, 2, 3$, the derivatives $\partial^l F(m)/\partial m^l$ exist and are bounded on $\mathcal{M}$.
C7. Write $\varepsilon = y - E(y|\mathbf{w})$. Given $\mathbf{w}$, ε is sub-Gaussian, that is, for some constants $0 < M_1, M_2 < \infty$, $E(\exp(|\varepsilon|/M_1)|\mathbf{w}) < M_2$, almost surely.
C8. For any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $\Pr(\mathbf{x}^\intercal\boldsymbol{\beta} \neq \mathbf{x}^\intercal\boldsymbol{\beta}_0) > 0$.

Remark 5 Condition C1 is a mild assumption on knots and is required to derive the asymptotic consistency and the rate of convergence of $\hat{\boldsymbol{\tau}}$. Condition C2 is the standard assumption in semiparametric estimation. Conditions C3–C6 are required for entropy calculation in the proofs of Theorems 2–4. Here, condition C4 relaxes the restrictive boundedness assumption of $\mathbf{x}$ in the literature. In addition, condition C6 is needed to guarantee the smoothness of the least favorable direction $\boldsymbol{\phi}^*$ defined in (6). Condition C7 is essential to calculate the bracketing integral with respect to the Bernstein norm (van der Vaart and Wellner 1996). Finally, condition C8 is used to establish the identifiability of the model.

3.2 Semiparametric efficient score and information bound

The efficient score and information bound serve as the benchmark for asymptotic behavior of $\hat{\boldsymbol{\beta}}$. The efficient score and information bound for quasi-likelihood function presented in the following theorem are derived without specifying the distribution of outcome variable. Let $\|f\|_{L_2} = \left[\int_{\Im} f^2(z)dz\right]^{1/2}$ denote the L_2-norm of square integrable function $f(z)$ on $\Im$. Denote $f \in L_2(\Im)$ if $\|f\|_{L_2} < \infty$.

Theorem 1 *Under model (1), the efficient score for $\boldsymbol{\beta}$ at $\boldsymbol{\tau}_0$ is given by*

$$\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v}) = (\mathbf{x} - \boldsymbol{\phi}^*)\Delta_0\Sigma_0^{-1}(y - \mu_0),$$

where $\Delta_0 = f(\mathbf{x}^\intercal\boldsymbol{\beta}_0 + \psi_0(z))$, $\mu_0 = F(\mathbf{x}^\intercal\boldsymbol{\beta}_0 + \psi_0(z))$, and $\Sigma_0 = Var(y|\mathbf{w})$. The least favorable direction $\boldsymbol{\phi}^*$ satisfies

$$E[(\mathbf{x} - \boldsymbol{\phi}^*)\Delta_0^2\Sigma_0^{-1}h] = 0,$$

for any $h \in L_2(\Im)$ and has a closed form

$$\boldsymbol{\phi}^*(z) = \frac{E_{\mathbf{x}|z}[\mathbf{x}\Delta_0^2\Sigma_0^{-1}|z]}{E_{\mathbf{x}|z}[\Delta_0^2\Sigma_0^{-1}|z]}. \tag{6}$$

The semiparametric information bound for $\boldsymbol{\beta}$ at $\boldsymbol{\tau}_0$ is given by

$$\mathbf{I}(\boldsymbol{\beta}_0) = E[\ell^*_{\boldsymbol{\beta}}(\mathbf{v}; \boldsymbol{\tau}_0)]^{\otimes 2} = E[(\mathbf{x} - \boldsymbol{\phi}^*)^{\otimes 2}\Delta_0^2\Sigma_0^{-1}],$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\intercal$, $\mathbf{a} \in \mathbb{R}^d$.

Remark 6 Because $f(z)$ is assumed to be smooth on $\mathcal{M}$, the least favorable direction $\boldsymbol{\phi}^*(z)$ is smooth on $\Im$ if the true variance function Σ_0 is assumed to be smooth and positive on $\mathcal{F}$. The smoothness of $\boldsymbol{\phi}^*(z)$ is required to derive the asymptotic results of $\hat{\boldsymbol{\tau}}$.

For a specific case, assume that the log-likelihood function of y given covariate $\mathbf{x}$ takes the form

$$\ell(\boldsymbol{\tau}; \mathbf{v}) = \sigma^{-2}[y\theta(\mathbf{w}) - b(\theta(\mathbf{w}))] + c(y, \sigma^2), \tag{7}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions and σ^2 is possibly unknown. Note that $E(y|\mathbf{w}) = b'(\theta)$ and $Var(y|\mathbf{w}) = \sigma^2 b''(\theta)$. Under the canonical link $\theta = \mathbf{x}^\intercal\boldsymbol{\beta} + \psi(z)$ and model (1), $\mu = b'(\theta) = F(\mathbf{x}^\intercal\boldsymbol{\beta} + \psi(z))$. The score function $\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\tau}; \mathbf{v})$ is the partial derivative of $\ell(\boldsymbol{\tau}; \mathbf{v})$ with respect to $\boldsymbol{\beta}$, namely,

$$\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\tau}; \mathbf{v}) = \frac{\partial \ell(\boldsymbol{\tau}; \mathbf{v})}{\partial \boldsymbol{\beta}} = \sigma^{-2}[y - b'(\theta)]\mathbf{x}.$$

Consider the parametric smooth submodel $(\boldsymbol{\beta}, \psi_t)$, where $\psi_t|_{t=0} = \psi$ and $\partial\psi_t/\partial t|_{t=0} = h$. Let $\mathcal{H} \subset L_2(\Im)$ be the class of such h on $\Im$. The score operator for ψ is defined as

$$\dot{\ell}_{\psi}(\boldsymbol{\tau}; \mathbf{v})[h] = \left.\frac{\partial \ell(\boldsymbol{\beta}, \psi_t; \mathbf{v})}{\partial t}\right|_{t=0} = \sigma^{-2}[y - b'(\theta)]h.$$

The efficient score for $\boldsymbol{\beta}$ at $\boldsymbol{\tau}_0$ is given by

$$\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v}) = \dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v}) - \dot{\ell}_{\psi}(\boldsymbol{\tau}_0; \mathbf{v})[\boldsymbol{\psi}^*],$$

where $\boldsymbol{\psi}^* \in \mathcal{H}^d$ minimizes $\rho(\mathbf{h}) \equiv \|\dot{\ell}_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v}) - \dot{\ell}_{\psi}(\boldsymbol{\tau}_0; \mathbf{v})[\mathbf{h}]\|_2^2$ over $\mathcal{H}^d$. It follows that

$$\boldsymbol{\psi}^*(z) = \frac{E_{\mathbf{x}|z}[\mathbf{x}b''(\theta_0)|z]}{E_{\mathbf{x}|z}[b''(\theta_0)|z]},$$

where $\theta_0 = \mathbf{x}^\intercal\boldsymbol{\beta}_0 + \psi_0(z)$. Under model (7), $\Delta_0 = b''(\theta_0)$ and $\Sigma_0 = \sigma^2 b''(\theta_0)$, and hence $\boldsymbol{\phi}^*$ reduces to $\boldsymbol{\psi}^*$. Therefore, the efficient score function for $\boldsymbol{\beta}$ at $\boldsymbol{\tau}_0$ is given by

$$\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v}) = \frac{y - b'(\theta_0)}{\sigma^2}\left(\mathbf{x} - \boldsymbol{\psi}^*\right)$$

and the efficient information takes the form of

$$\mathbf{I}(\boldsymbol{\beta}_0) = E[\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0; \mathbf{v})]^{\otimes 2} = \sigma^{-2}E\left[b''(\theta_0)\left(\mathbf{x} - \boldsymbol{\psi}^*\right)^{\otimes 2}\right].$$

3.3 Asymptotic results

Theorem 2 (Uniform consistency and rate of convergence) *Let* $q_n = O(n^{\nu})$, *for* $1/(2r+2) < \nu < 1/(2r)$. *Suppose conditions C1–C8 hold. Then,*

$$\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2 = O_p\left(n^{-\min(r\nu,(1-\nu)/2)}\right).$$

Consequently, by Lemma 7 of Stone (1986), $\|\hat{\psi} - \psi_0\|_\infty = o_p(1)$. *Furthermore, if* $\nu = 1/(1+2r)$, $O_p(n^{-\min(r\nu,(1-\nu)/2)}) = O_p(n^{-r/(1+2r)})$, *which is the optimal rate of convergence in semiparametric regression.*

For a single observation $\mathbf{v}$, its log density for $\boldsymbol{\tau}$ is given by

$$Q(\boldsymbol{\tau}; \mathbf{v}) = \int_y^{\mu} \frac{y-s}{\sigma^2 V(s)} \mathrm{d}s.$$

In the following we use the similar notations as those in Huang (1996) and Wellner and Zhang (2007) with the objective function $Q(\boldsymbol{\tau}; \mathbf{v})$. The score function for $\boldsymbol{\beta}$ is the vector of partial derivative of $Q(\boldsymbol{\tau}; \mathbf{v})$ with respect to $\boldsymbol{\beta}$, namely,

$$m_1(\boldsymbol{\tau}; \mathbf{v}) \equiv \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\tau}; \mathbf{v}) = \left(\frac{\partial Q(\boldsymbol{\tau}; \mathbf{v})}{\partial \beta_1}, \ldots, \frac{\partial Q(\boldsymbol{\tau}; \mathbf{v})}{\partial \beta_d}\right)^{\mathsf{T}}.$$

Consider the parametric smooth submodel $(\boldsymbol{\beta}, \psi_t)$, where $\psi_t|_{t=0} = \psi$ and $\partial\psi_t/\partial t|_{t=0} = h$. The score operator for ψ is defined as

$$m_2(\boldsymbol{\tau}; \mathbf{v})[h] = \left.\frac{\partial Q(\boldsymbol{\beta}, \psi_t; \mathbf{v})}{\partial t}\right|_{t=0}.$$

Define

$$m_{11}(\boldsymbol{\tau}; \mathbf{v}) = \nabla^2_{\boldsymbol{\beta}} Q(\boldsymbol{\tau}; \mathbf{v}), \quad m_{12}(\boldsymbol{\tau}; \mathbf{v})[h] = \left.\frac{\partial m_1(\boldsymbol{\beta}, \psi_t; \mathbf{v})}{\partial t}\right|_{t=0},$$
$$m_{21}(\boldsymbol{\tau}; \mathbf{v})[h] = \nabla_{\boldsymbol{\beta}} m_2(\boldsymbol{\tau}; \mathbf{v})[h], \quad m_{22}(\boldsymbol{\tau}; \mathbf{v})[h_1, h_2] = \left.\frac{\partial m_2(\boldsymbol{\beta}, \psi_{t_2}; \mathbf{v})[h_1]}{\partial t_2}\right|_{t_2=0}.$$

Moreover, for $\mathbf{h} = (h_1, \ldots, h_d)^{\mathsf{T}} \in \mathcal{H}^d$, denote

$$\begin{aligned}
m_2(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}] &= (m_2(\boldsymbol{\tau}; \mathbf{v})[h_1], \ldots, m_2(\boldsymbol{\tau}; \mathbf{v})[h_d])^{\mathsf{T}},\\
m_{12}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}] &= (m_{12}(\boldsymbol{\tau}; \mathbf{v})[h_1], \ldots, m_{12}(\boldsymbol{\tau}; \mathbf{v})[h_d])^{\mathsf{T}},\\
m_{21}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}] &= (m_{21}(\boldsymbol{\tau}; \mathbf{v})[h_1], \ldots, m_{21}(\boldsymbol{\tau}; \mathbf{v})[h_d])^{\mathsf{T}},\\
m_{22}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}, h] &= (m_{22}(\boldsymbol{\tau}; \mathbf{v})[h_1, h], \ldots, m_{22}(\boldsymbol{\tau}; \mathbf{v})[h_d, h])^{\mathsf{T}}.
\end{aligned}$$

Let

$$\mathbf{h}^* = \arg\min_{\mathbf{h}\in\mathcal{H}^d} E\|m_1(\boldsymbol{\tau}_0;\mathbf{v}) - m_2(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}]\|_2^2.$$

Observe that $E\{[m_1(\boldsymbol{\tau}_0;\mathbf{v}) - m_2(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}^*]]^\mathsf{T} m_2(\boldsymbol{\tau}_0,\mathbf{v})[\mathbf{h}]\} = 0$, for any $\mathbf{h}\in\mathcal{H}^d$. It is readily to show that

$$\mathbf{h}^* = \frac{E_{\mathbf{x}|z}[\mathbf{x}\Delta_0 V^{-1}(\mu_0)|z]}{E_{\mathbf{x}|z}[\Delta_0 V^{-1}(\mu_0)|z]}.$$

Denote $m^*(\boldsymbol{\tau}_0;\mathbf{v}) = m_1(\boldsymbol{\tau}_0;\mathbf{v}) - m_2(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}^*]$. When the variance function is correctly specified, that is, $\Sigma_0 = \sigma^2 V(\mu_0)$, $\mathbf{h}^*$ reduces to $\boldsymbol{\phi}^*$ and $m^*(\boldsymbol{\tau}_0;\mathbf{v})$ reduces to the efficient score function $\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0;\mathbf{v})$ accordingly. Define $\mathbf{A}_0 = -P[m_{11}(\boldsymbol{\tau}_0;\mathbf{v}) - m_{12}(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}^*]]$ and $\mathbf{B}_0 = E[m^*(\boldsymbol{\tau}_0;\mathbf{v})]^{\otimes 2}$. For a measurable function f, define Pf and $\mathbb{P}_n f$ as the expectation of f under the measure P and the empirical measure $\mathbb{P}_n$, respectively. The empirical process evaluated at f is defined as $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)f$.

Theorem 3 (Asymptotic normality) *Suppose conditions C1–C8 hold and* $\mathbf{A}_0$ *is nonsingular. Then*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{A}_0^{-1}\mathbb{G}_n[m^*(\boldsymbol{\tau}_0;\mathbf{v})] + o_p(1) \to N(0, \mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}),$$

in distribution, as $n\to\infty$. *If the variance function is correctly specified, then under model* (1), $\mathbf{A}_0 = \mathbf{B}_0$. *The theorem reduces to*

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{B}_0^{-1}\mathbb{G}_n[\ell^*_{\boldsymbol{\beta}}(\boldsymbol{\tau}_0;\mathbf{v})] + o_p(1) \to N(0, \mathbf{B}_0^{-1}),$$

in distribution, as $n\to\infty$.

Remark 7 If $V(\mu)$ is correctly specified, then $\mathbf{B}_0 = E[(y-\mu_0)^2\Delta_0\Sigma_0{}^{-2}(\mathbf{x}-\boldsymbol{\psi}^*)^{\otimes 2}]$ reduces to $\mathbf{I}(\boldsymbol{\beta}_0)$. Therefore, the spline estimator $\hat{\boldsymbol{\beta}}$ achieves the semiparametric information bound and hence is efficient when the variance function is correctly specified.

4 Consistent estimation of standard error

To consistently estimate the standard error of $\hat{\boldsymbol{\beta}}$, we need to estimate $\mathbf{h}^* = (h_1^*,\ldots,h_d^*)^\mathsf{T}$. We again apply the spline-based sieve method to approximate $\mathbf{h}^*$, that is, $h_s^* \approx h_{n,s}^* = \sum_{j=1}^{q_n}\gamma_{j,s}B_j$, where q_n may depend on s, $s = 1,\ldots,d$. The coefficients $\boldsymbol{\gamma}_s = (\gamma_{1,s},\ldots,\gamma_{q_n,s})^\mathsf{T}$ can be estimated by minimizing

$$\mathbb{P}_n[m_{1,s}(\hat{\boldsymbol{\tau}};\mathbf{v}) - m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[h_{n,s}^*]]^2, \tag{8}$$

where $m_{1,s}(\hat{\boldsymbol{\tau}};\mathbf{v})$ is the sth element of $m_1(\hat{\boldsymbol{\tau}};\mathbf{v})$. Observe that $m_2(\boldsymbol{\tau};\mathbf{v})[h]$ is a linear operator for h. Therefore, the optimization problem in (8) is equivalent to a least-squares problem by solving $\hat{\boldsymbol{\gamma}}_s = (\hat{\gamma}_{1,s},\ldots,\hat{\gamma}_{q_n,s})^\mathsf{T}$ that minimizes

$$\mathbb{P}_n\left[m_{1,s}(\hat{\boldsymbol{\tau}};\mathbf{v})-\sum_{j=1}^{q_n}\gamma_{j,s}m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[B_j]\right]^2.$$

The estimate of h_s^* is defined as $\hat{h}_s^*=\sum_{j=1}^{q_n}\hat{\gamma}_{j,s}B_j$. Denote $\hat{\mathbf{h}}^*=(\hat{h}_1^*,\ldots,\hat{h}_d^*)^\mathsf{T}$. Standard least-squares calculation leads to

$$\hat{\boldsymbol{\gamma}}_s=[m_2^\mathsf{T}(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]]^{-1}m_2^\mathsf{T}(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]m_1^{(s)}(\hat{\boldsymbol{\tau}};\mathbf{v}),$$

where $m_1^{(s)}(\hat{\boldsymbol{\tau}};\mathbf{v})$ is a vector with ith element, $m_{1,s}(\hat{\boldsymbol{\tau}};\mathbf{v}_i)$, $i=1,\ldots,n$ and $m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]$ is an $n\times q_n$ matrix with (j,k)th entry, $m_2(\hat{\boldsymbol{\tau}};\mathbf{v}_j)[B_k]$, $j=1,\ldots,n$ and $k=1,\ldots,q_n$. Denote $\hat{\mathbf{A}}=-\mathbb{P}_n[m_{11}(\hat{\boldsymbol{\tau}};\mathbf{v})-m_{12}(\hat{\boldsymbol{\tau}};\mathbf{v})[\hat{\mathbf{h}}^*]]$ and $\hat{\mathbf{B}}=\mathbb{P}_n[m_1(\hat{\boldsymbol{\tau}};\mathbf{v})-m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[\hat{\mathbf{h}}^*]]^{\otimes 2}$.

Theorem 4 (Variance estimation) *Under the same conditions assumed in Theorem 3, $\hat{\mathbf{h}}^*$ is a consistent estimate of $\mathbf{h}^*$. Consequently, $\hat{\mathbf{A}}\to\mathbf{A}_0$ and $\hat{\mathbf{B}}\to\mathbf{B}_0$, in probability, as $n\to\infty$.*

Some straightforward calculation leads to

$$\hat{\mathbf{B}}=\mathbb{P}_n\left[m_1(\hat{\boldsymbol{\tau}};\mathbf{v})-m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[\hat{\boldsymbol{\phi}}^*]\right]^{\otimes 2}=\hat{\mathcal{O}}_{11}-\hat{\mathcal{O}}_{12}\hat{\mathcal{O}}_{22}^{-1}\hat{\mathcal{O}}_{21},$$

where

$$\hat{\mathcal{O}}_{11}=\mathbb{P}_n[m_1(\hat{\boldsymbol{\tau}};\mathbf{v})^{\otimes 2}],\quad \hat{\mathcal{O}}_{12}=\mathbb{P}_n[m_1(\hat{\boldsymbol{\tau}};\mathbf{v})m_2^\mathsf{T}(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]],$$
$$\hat{\mathcal{O}}_{21}=\hat{\mathcal{O}}_{12}^\mathsf{T},\quad \hat{\mathcal{O}}_{22}=\mathbb{P}_n[m_2(\hat{\boldsymbol{\tau}};\mathbf{v})[\mathbf{B}]^{\otimes 2}].$$

Denote $\hat{\mathcal{E}}_{11}=E_{y|\mathbf{w}}(\hat{\mathcal{O}}_{11}|\mathbf{w})$, $\hat{\mathcal{E}}_{12}=E_{y|\mathbf{w}}(\hat{\mathcal{O}}_{12}|\mathbf{w})$, $\hat{\mathcal{E}}_{21}=\hat{\mathcal{E}}_{12}^\mathsf{T}$, and $\hat{\mathcal{E}}_{22}=E_{y|\mathbf{w}}(\hat{\mathcal{O}}_{22}|\mathbf{w})$. Corollary 1 shows that the conditional expected information

$$\hat{\mathcal{E}}_n=\hat{\mathcal{E}}_{11}-\hat{\mathcal{E}}_{12}\hat{\mathcal{E}}_{22}^{-1}\hat{\mathcal{E}}_{21}$$

is a consistent estimator of $\mathbf{I}(\boldsymbol{\beta}_0)$. Specifically, with the notations defined in Sect. 2, $\mathbf{I}(\boldsymbol{\beta}_0)$ can be consistently estimated by

$$\begin{aligned}\hat{\mathcal{E}}_n&=n^{-1}\hat{\sigma}^2\hat{\mathbf{E}}^{11}\\&=n^{-1}\hat{\sigma}^2\left[\mathbf{X}^\mathsf{T}\hat{\boldsymbol{\xi}}^2\hat{\mathbf{V}}^{-1}\mathbf{X}-\mathbf{X}^\mathsf{T}\hat{\boldsymbol{\xi}}^2\hat{\mathbf{V}}^{-1}\mathbf{B}\left(\mathbf{B}^\mathsf{T}\hat{\boldsymbol{\xi}}^2\hat{\mathbf{V}}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^\mathsf{T}\hat{\boldsymbol{\xi}}^2\hat{\mathbf{V}}^{-1}\mathbf{X}\right].\end{aligned}\tag{9}$$

Here $\hat{\mathbf{E}}^{11}$, $\hat{\boldsymbol{\xi}}$, and $\hat{\mathbf{V}}$ represent $\mathbf{E}^{11}$, $\boldsymbol{\xi}$, and $\mathbf{V}$ evaluated at $\hat{\boldsymbol{\vartheta}}$, respectively.

Corollary 1 *Under the same conditions assumed in Theorem 3, if the variance function is correctly specified, then $\hat{\mathcal{E}}_n$ is asymptotically consistent to $\mathbf{I}(\boldsymbol{\beta}_0)$.*

The finite sample performance of the variance estimation method is evaluated in simulation study and the approach is applied in real application.

5 Simulation study and data analysis

5.1 Simulation

An extensive simulation study is carried out to evaluate the finite sample performance of the method in this section. In each simulation, we generate n i.i.d. observations $\{(y_i, \mathbf{x}_i, z_i) : i = 1, \ldots, n\}$ with $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^\mathsf{T}$. For each subject i, the data are generated by the following scheme: covariates $x_{i1}, x_{i2}, x_{i3} \sim N(0, 1)$, $x_{i4} \sim B(0.5)$, and $z_i \sim U[0, 1]$. The outcome y_i is generated from a Poisson distribution with conditional mean $\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_0 + \psi_0(z_i))$. For all of generated data, $\sigma^2 = 1$, $\boldsymbol{\beta}_0 = (0.5, 0.5, 0.5, -1)^\mathsf{T}$, and $\psi_0(z) = 2 \times 10^5 z^{11}(1-z)^5 + 10^4 z^2(1-z)^{10}$. The efficient information $\mathbf{I}(\boldsymbol{\beta}_0)$ given in Sect. 3 for this simulation setting reduces to

$$E\left[\exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0 + \psi_0(z))\left\{\mathbf{x} - \frac{E[\mathbf{x}\exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0)]}{E[\exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0)]}\right\}^{\otimes 2}\right].$$

The inverse of asymptotic variance $\mathbf{I}^{-1}(\boldsymbol{\beta}_0)$ can be approximated by

$$\begin{pmatrix} 0.0743 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0743 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0743 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.3780 \end{pmatrix}.$$

Thus, the standard error of $\hat{\boldsymbol{\beta}}$ can be approximated by $n^{-1/2}(0.2725, 0.2725, 0.2725, 0.6148)^\mathsf{T}$, which is served as the benchmark for variance study.

Cubic B-splines are used to approximate ψ_0 in the simulation. The number and the location of knots $\mathcal{T}_n$ are determined according to the AIC criteria and the quantile method, respectively, as discussed in Sect. 2.3. The proposed B-spline quasi-likelihood method is compared with the local polynomial quasi-likelihood method and the penalized quasi-likelihood method. The Monte Carlo sample means, biases, standard deviations, and mean squared errors of $\hat{\boldsymbol{\beta}}$ over 1000 replications for $n = 200$ or 400 are summarized in Table 1. The standard deviations and the mean squared errors of $\hat{\boldsymbol{\beta}}$ for the B-spline fit are almost identical to those for the penalized fit, and are smaller than those based on the local polynomial method. The results also indicate that the standard deviations of the estimates for B-spline fit decrease at a rate of $n^{-1/2}$. In addition, the B-spline method for estimation of σ^2 works reasonably well. To evaluate the accuracy of the spline estimator of $\psi_0(z)$, we compute the estimates of $\psi_0(z)$ at points $z = 0.05, 0.15 \ldots, 0.95$. The pointwise biases, standard deviations, and mean squared errors of $\hat{\psi}(z)$ are given in Table 2. It shows that the B-spline method and the penalized method yield similar results, while the local polynomial method demonstrates the larger biases, standard deviations, and mean squared errors, compared to its alternatives. The biases, standard deviations, and mean squared errors decrease as n increases, indicating the consistency of the B-spline estimate. Figure 1 displays the curve estimates of ψ_0 and the corresponding confidence bands over 1000 Monte Carlo samples for $n = 200$ or 400. The fitted curves are reasonably close to the true curve,

Table 1 Summary of parameter estimation for simulation study

Parameters	Model	$n = 200$				$n = 400$			
		Mean	Bias	sd	mse	Mean	Bias	sd	mse
β_1	BQ	0.5003	0.0003	0.0238	0.0005	0.5001	0.0001	0.0152	0.0002
	PQ	0.5002	0.0000	0.0239	0.0005	0.5000	0.0000	0.0153	0.0002
	LQ	0.5395	0.0395	0.0448	0.0035	0.5443	0.0443	0.0288	0.0027
β_2	BQ	0.5000	−0.0000	0.0240	0.0005	0.5000	0.0000	0.0153	0.0002
	PQ	0.5000	−0.0000	0.0241	0.0005	0.5000	−0.0000	0.0153	0.0002
	LQ	0.5415	0.0415	0.0436	0.0036	0.5411	0.0411	0.0287	0.0025
β_3	BQ	0.5003	0.0003	0.0237	0.0005	0.5000	−0.0000	0.0155	0.0002
	PQ	0.5000	−0.0000	0.0237	0.0005	0.4999	−0.0001	0.0156	0.0002
	LQ	0.5401	0.0401	0.0465	0.0038	0.5432	0.0432	0.0273	0.0026
β_4	BQ	−1.0004	−0.0004	0.0500	0.0025	−1.0002	−0.0002	0.0327	0.0010
	PQ	−1.0001	−0.0001	0.0502	0.0025	−1.0002	−0.0002	0.0329	0.0010
	LQ	−0.8664	0.1336	0.0710	0.0228	−0.8722	0.1278	0.0422	0.0181
σ^2	BQ	1.0068	0.0068	0.1200	0.0144	1.0003	0.0003	0.0831	0.0069
	PQ	1.0092	0.0092	0.1184	0.0141	1.0062	0.0062	0.0836	0.0070
	LQ	1.0539	0.0539	0.0726	0.0081	1.0951	0.0951	0.0511	0.0116

$\boldsymbol{\beta}_0 = (1/2, 1/2, 1/2, -1)^\intercal$ and $\sigma^2 = 1$. Sample mean, bias, standard deviation (sd), and mean squared error (mse) of the estimates from B-spline quasi-likelihood model (BQ), penalized quasi-likelihood model (PQ), and local polynomial quasi-likelihood model (LQ), based on 1000 Monte Carlo samples for $n = 200$ or 400, respectively

Table 2 Bias, standard deviation (sd), and mean squared error (mse) of the estimators of $\psi_0(z) = 2 \times 10^5 z^{11}(1-z)^5 + 10^4 z^2(1-z)^{10}$ from B-spline quasi-likelihood model (BQ), penalized quasi-likelihood model (PQ), and local polynomial quasi-likelihood model (LQ), based on 1000 Monte Carlo samples for sample size $n = 200$ or 400, respectively

	z	BQ			PQ			LQ		
		Bias	sd	mse	Bias	sd	mse	Bias	sd	mse
$n = 200$	0.05	0.0823	0.1730	0.0367	0.0148	0.1357	0.0240	−0.0501	0.1543	0.0263
	0.15	−0.0358	0.0455	0.0033	−0.0191	0.0458	0.0024	−0.2991	0.0810	0.0960
	0.25	0.0332	0.0521	0.0038	0.0228	0.0493	0.0029	−0.1234	0.0712	0.0203
	0.35	−0.0697	0.0952	0.0139	−0.0334	0.0889	0.0091	−0.0701	0.1239	0.0202
	0.45	−0.0158	0.1318	0.0176	0.0092	0.1445	0.0209	−0.0829	0.1980	0.0461
	0.55	0.0461	0.1309	0.0192	−0.0069	0.1580	0.0250	−0.1081	0.2026	0.0527
	0.65	−0.0195	0.1335	0.0182	−0.0154	0.1409	0.0201	−0.1374	0.1805	0.0507
	0.75	−0.0361	0.1395	0.0207	−0.0260	0.1490	0.0228	−0.1412	0.1830	0.0534
	0.85	0.0337	0.1769	0.0324	0.0143	0.1918	0.0371	−0.1044	0.2407	0.0688
	0.95	−0.1017	0.2715	0.0841	−0.0511	0.2912	0.0874	−0.1441	0.2673	0.0922
$n = 400$	0.05	0.0578	0.0996	0.0132	0.0457	0.0902	0.0102	−0.0424	0.1108	0.0140
	0.15	0.0009	0.0285	0.0008	−0.0096	0.0304	0.0010	−0.2896	0.0500	0.0863
	0.25	−0.0015	0.0353	0.0012	0.0176	0.0334	0.0014	−0.1223	0.0466	0.0171
	0.35	−0.0058	0.0679	0.0046	−0.0294	0.0620	0.0047	−0.0748	0.0812	0.0122
	0.45	−0.0062	0.0958	0.0092	0.0154	0.1036	0.0109	−0.0697	0.1223	0.0198
	0.55	0.0043	0.1004	0.0101	−0.0087	0.1157	0.0134	−0.1094	0.1275	0.0282
	0.65	−0.0096	0.0978	0.0096	0.0000	0.1010	0.0102	−0.1211	0.1129	0.0274
	0.75	0.0012	0.1082	0.0117	−0.0148	0.1062	0.0115	−0.1422	0.1199	0.0346
	0.85	−0.0146	0.1503	0.0228	0.0027	0.1364	0.0186	−0.1008	0.1477	0.0320
	0.95	−0.0211	0.2117	0.0452	−0.0253	0.1852	0.0349	−0.1359	0.1724	0.0482

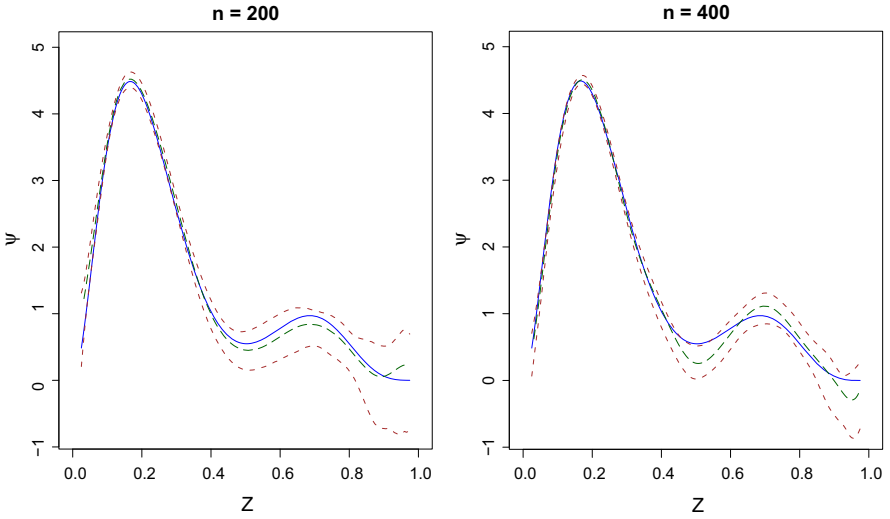

Fig. 1 Curve estimates and the corresponding 95 % confidence bands for $\psi_0(z)$. The *solid curve* is the true mean function; the *long dashed curves* are B-spline fits; and the *dashed curves* are the corresponding 2.5 and 97.5 % quantiles, based on 1000 Monte Carlo samples for $n = 200$ or 400, respectively

demonstrating there is little bias. Moreover, the lower and upper bounds of confidence bands follow the true function ψ_0 pretty closely, indicating little variability of the estimates. When the sample size increases, the variation decreases accordingly. Clearly, the spline estimators accurately capture the nonparametric feature of ψ_0.

The results of assessment for the standard error estimation method based on the conditional expected information given in (9) and the bootstrap method are presented in Table 3. We found that the coverage probabilities of confidence intervals obtained from 1000 replications are close to 5 % nominal level for all sample sizes. Moreover, the averages of estimated standard errors of $\hat{\boldsymbol{\beta}}$ based on the conditional expected information method are close to the corresponding Monte Carlo standard deviations of $\hat{\boldsymbol{\beta}}$ and the asymptotic standard error estimates, indicating that the proposed variance estimation method works reasonably well. These Monte Carlo results provide a numerical justification of asymptotic results presented in Theorems 3 and 4. The proposed variance estimation method is superior to the bootstrap method in terms of the mean and standard deviation of estimated standard errors in our simulation setting. Moreover, we assess the power of the spline Wald test from 2000 Monte Carlo samples for n = 200 or 400, respectively. Under the null hypothesis, $H_0 : \beta_i = \beta_0, i = 1, \ldots, 4$, the spline Wald test statistics $T_i = \left[(\hat{\beta}_i - \beta_0)/se(\hat{\beta}_i)\right]^2$ follow χ^2_1 distribution by Theorem 3. Here the standard errors $se(\hat{\beta}_i)$ are estimated by the conditional expected information. The power is computed as the proportion of hypotheses being rejected in 2000 replications. The power curves are displayed in Fig. 2. As expected, all power curves are symmetric around the true parameters and the power increases as the sample size increases or the effect size increases. Moreover, the sizes of the tests for all sample sizes are close to the nominal level.

Table 3 Results of variance study

	Parameters	Proposed method				Bootstrap method			
		Mean.se	a.se	sd.se	Cov.prob (%)	Mean.se	a.se	sd.se	Cov.prob (%)
$n = 200$	β_1	0.0233	0.0193	0.0040	94.6	0.0261	0.0193	0.0055	95.9
	β_2	0.0234	0.0193	0.0041	94.7	0.0262	0.0193	0.0056	96.0
	β_3	0.0233	0.0193	0.0040	94.6	0.0262	0.0193	0.0055	96.3
	β_4	0.0490	0.0435	0.0055	95.0	0.0528	0.0435	0.0093	95.2
$n = 400$	β_1	0.0152	0.0136	0.0019	94.8	0.0161	0.0136	0.0028	95.2
	β_2	0.0152	0.0136	0.0019	94.6	0.0160	0.0136	0.0028	95.2
	β_3	0.0152	0.0136	0.0019	94.9	0.0161	0.0136	0.0028	95.6
	β_4	0.0328	0.0307	0.0025	94.8	0.0337	0.0307	0.0046	95.0

Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (sd.se), asymptotic standard error (a.se), and coverage probability (cov.prob) of confidence interval for $\boldsymbol{\beta}$, based on 1000 Monte Carlo samples for $n = 200$ or 400, respectively. The number of bootstrap samples is 100

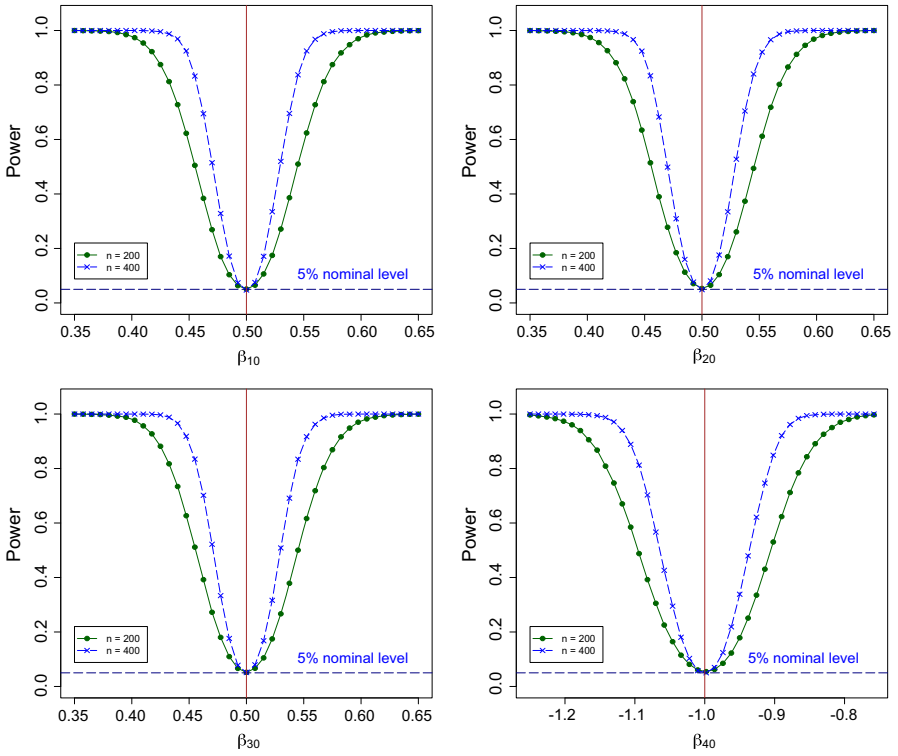

Fig. 2 Power curves of the B-spline Wald test, based on the conditional expected information method with 2000 duplications. The *solid* and *long dashed curves* are estimated powers for $n = 200$ and 400, respectively

In general, the proposed B-spline approach is a sound and practical method for moderate sample sizes.

5.2 Data analysis

The proposed B-spline quasi-likelihood model is applied to a data set from a study of nesting horseshoe crabs (Brockmann 1996; Li 2012). In this study, for each female crab there was a male crab resident in her nest. The aim of the study is to investigate factors affecting whether the female crab had any other males (satellites) resident nearby. The outcome is the number of satellites for each female crab. The covariates include the female crab's color (light median, median, dark median, or dark), spine condition (both good, one worn or broken, or both worn or broken), carapace width in centimeters, and weight in kilograms. A record with an outlier in weight was excluded. The data consist of 172 observations with complete information.

Let y denote the number of satellites. Let $\mathbf{x}_1 = (x_{11}, x_{12}, x_{13})^\intercal$, $\mathbf{x}_2 = (x_{21}, x_{22})^\intercal$, and x_3 be female crab's color, spine condition, and carapace width, respectively, and let z be the weight. We propose the quasi-likelihood model

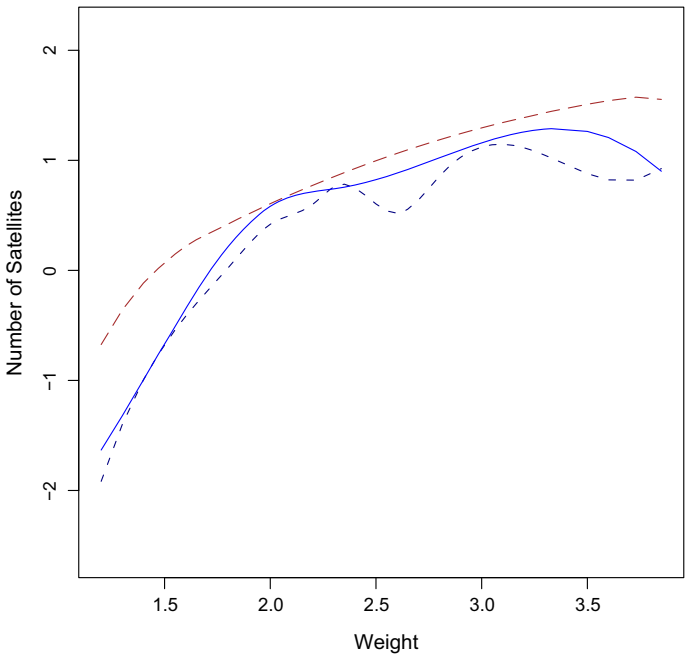

Fig. 3 Comparison of B-spline, penalized, and local polynomial estimators of $\psi(z)$ for crab data: *solid line* is the B-spline quasi-likelihood estimator; *dashed line* is the penalized quasi-likelihood estimator; and *long dashed line* is the local polynomial quasi-likelihood estimator. The numbers of knots for B-spline and penalized estimators are 5 and 9, respectively. The locations of knots are chosen by the quantile method. The smoothing parameter λ for penalized estimation is 0.29, and the bandwidth for local polynomial estimation is chosen as 0.14

$$\mu(\mathbf{x}, z) = E(y|\mathbf{x}, z) = F(\mathbf{x}^\mathsf{T}\boldsymbol{\beta} + \psi(z))$$

with $F(s) = \exp(s)$ and $V(\mu) = \sigma^2\mu$, where $\mathbf{x} = (\mathbf{x}_1^\mathsf{T}, \mathbf{x}_2^\mathsf{T}, x_3)^\mathsf{T}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^\mathsf{T}$. The unknown function ψ is fitted as a cubic B-spline with 5 knots. The place of knots is determined by the quantile method. The Wald test is employed for inference of $\boldsymbol{\beta}$. The standard errors of $\hat{\boldsymbol{\beta}}$ are estimated by the conditional expected information method. We also consider the local polynomial quasi-likelihood method and the penalized quasi-likelihood method for purposes of comparison. Table 4 summaries the results of estimated regression parameters and corresponding standard errors for three methods. The fitted curves from three estimation methods are presented in Fig. 3. The estimated functions are clearly nonlinear and the female crabs with higher weights are more likely to have higher number of satellites.

6 Proofs of asymptotic results

6.1 Notations and technical lemmas

Let P be a probability distribution. Define $L_2(P) = \{f : \int f^2 dP < \infty\}$. Given two functions f_L and f_R, an $\varepsilon-$bracket $[f_L, f_R]$ in $L_2(P)$ is the set of all functions f with

Table 4 The estimates and the corresponding standard errors (se) and p values from B-spline quasi-likelihood model (BQ), penalized quasi-likelihood model (PQ) and local polynomial quasi-likelihood model (LQ) for crab study

	BQ			PQ			LQ		
	$\hat{\boldsymbol{\beta}}$	se	p	$\hat{\boldsymbol{\beta}}$	se	p	$\hat{\boldsymbol{\beta}}$	se	p
Crab's color									
Dark (referent)	0.000			0.000			0.000		
Light median	0.359	0.235	0.127	0.314	0.237	0.184	0.399	0.228	0.080
Median	0.235	0.166	0.155	0.189	0.166	0.257	0.231	0.165	0.162
Dark median	0.013	0.182	0.941	0.003	0.182	0.989	−0.002	0.180	0.990
Spine condition									
Both worn or broken (referent)	0.000			0.000			0.000		
Both good	0.053	0.124	0.667	0.095	0.126	0.449	−0.006	0.117	0.956
One worn or broken	−0.175	0.204	0.392	−0.152	0.204	0.459	−0.182	0.192	0.342
Carapace width	0.003	0.056	0.956	0.009	0.056	0.868	−0.002	0.006	0.741
σ	1.803			1.792			1.775		

$f_L \le f \le f_R$ and $P(f_R - f_L)^2 < \varepsilon^2$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. Define the bracketing integral $J_{[]}(\eta, \mathcal{F}, L_2(P)) = \int_0^\eta \left[1 + N_{[]}(\varepsilon, \mathcal{F}, L_2(P))\right]^{1/2} d\varepsilon$. In the following, C represents a positive constant that may vary from place to place.

Lemma 1 *For any $\eta > 0$ and $0 < \varepsilon \le \eta$,*

$$\log N_{[]}(\varepsilon, \{Q(\boldsymbol{\tau}; \mathbf{v}) : \boldsymbol{\tau} \in \Theta \times \mathcal{S}_n, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta\}, \|\cdot\|_{P,B}) \le C q_n \log(\eta/\varepsilon),$$

and consequently,

$$J_{[]}(\eta, \{Q(\boldsymbol{\tau}; \mathbf{v}) : \boldsymbol{\tau} \in \Theta \times \mathcal{S}_n : \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta\}, \|\cdot\|_{P,B}) \le C q_n^{1/2} \eta,$$

where $\|\cdot\|_{P,B}$ is the Bernstein norm defined as $\|f\|_{P,B}^2 = 2P\left(e^{|f|} - |f| - 1\right)$ in van der Vaart and Wellner (1996) and $q_n = m_n + l$ is the number of spline basis functions.

Lemma 2 *If conditions C1–C8 hold, then there exist $0 < C_1 < C_2$ such that*

$$C_1 \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2 \le P[Q(\boldsymbol{\tau}_0; \mathbf{v}) - Q(\boldsymbol{\tau}; \mathbf{v})] \le C_2 \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2,$$

for $\boldsymbol{\tau}$ in a neighborhood of $\boldsymbol{\tau}_0$.

Lemma 3 (Consistency) *If conditions C1–C8 hold, then $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2 = o_p(1)$.*

Let

$$\begin{aligned}
S_1(\boldsymbol{\tau}) &= Pm_1(\boldsymbol{\tau}; \mathbf{v}), \quad S_2(\boldsymbol{\tau})[h] = Pm_2(\boldsymbol{\tau}; \mathbf{v})[h], \\
S_{1n}(\boldsymbol{\tau}) &= \mathbb{P}_n m_1(\boldsymbol{\tau}; \mathbf{v}), \quad S_{2n}(\boldsymbol{\tau})[h] = \mathbb{P}_n m_2(\boldsymbol{\tau}; \mathbf{v})[h], \\
\dot{S}_{11}(\boldsymbol{\tau}) &= Pm_{11}(\boldsymbol{\tau}; \mathbf{v}), \quad \dot{S}_{12}(\boldsymbol{\tau})[h] = Pm_{12}(\boldsymbol{\tau}; \mathbf{v})[h], \\
\dot{S}_{21}(\boldsymbol{\tau})[h] &= \dot{S}_{12}^{\mathsf{T}}(\boldsymbol{\tau})[h], \quad \dot{S}_{22}(\boldsymbol{\tau})[h_1, h_2] = Pm_{22}(\boldsymbol{\tau}; \mathbf{v})[h_1, h_2],
\end{aligned}$$

and

$$\begin{aligned}
S_2(\boldsymbol{\tau})[\mathbf{h}] &= Pm_2(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}], \quad S_{2n}(\boldsymbol{\tau})[\mathbf{h}] = \mathbb{P}_n m_2(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}], \\
\dot{S}_{12}(\boldsymbol{\tau})[\mathbf{h}] = \dot{S}_{21}^{\mathsf{T}}(\boldsymbol{\tau})[\mathbf{h}] &= Pm_{12}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}], \quad \dot{S}_{22}(\boldsymbol{\tau})[\mathbf{h}, h] = Pm_{22}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}, h].
\end{aligned}$$

Lemma 4 *Suppose that the following assumptions hold*

B1. *$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(1)$ and $\|\hat{\psi} - \psi_0\|_2 = O_p(n^{-\gamma})$, for some $\gamma > 0$.*
B2. *$S_1(\boldsymbol{\tau}_0) = 0$ and $S_2(\boldsymbol{\tau}_0)[h] = 0$, for all $h \in \mathcal{H}$.*
B3. *There exists an $\mathbf{h}^* = (h_1^*, \ldots, h_d^*)^{\mathsf{T}} \in \mathcal{H}^d$ such that $\dot{S}_{12}(\boldsymbol{\tau}_0)[h] - \dot{S}_{22}(\boldsymbol{\tau}_0)[\mathbf{h}^*, h] = 0$, for all $h \in \mathcal{H}$. Moreover, the matrix $\mathbf{A}_0 = -\dot{S}_{11}(\boldsymbol{\tau}_0) + \dot{S}_{21}(\boldsymbol{\tau}_0)[\mathbf{h}^*]$ is nonsingular.*
B4. *The estimator $\hat{\boldsymbol{\tau}}$ satisfy $S_{1n}(\hat{\boldsymbol{\tau}}) = o_p(n^{-1/2})$ and $S_{2n}(\hat{\boldsymbol{\tau}})[\mathbf{h}^*] = o_p(n^{-1/2})$.*

B5. *For any* $\delta_n \downarrow 0$,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\delta_n, \|\psi-\psi_0\|_2\le Cn^{-\gamma}} |\sqrt{n}(S_{1n}-S_1)(\boldsymbol{\tau})-\sqrt{n}(S_{1n}-S_1)(\boldsymbol{\tau}_0)| = o_p(1)$$

and

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\delta_n, \|\psi-\psi_0\|_2\le Cn^{-\gamma}} |\sqrt{n}(S_{2n}-S_2)(\boldsymbol{\tau})[\mathbf{h}^*]-\sqrt{n}(S_{2n}-S_2)(\boldsymbol{\tau}_0)[\mathbf{h}^*]| = o_p(1).$$

B6. *For some* $\alpha > 1$ *satisfying* $\alpha\gamma > 1/2$, *and for* $\boldsymbol{\tau}$ *with* $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \delta_n$ *and* $\|\psi-\psi_0\|_2 \le Cn^{-\gamma}$,

$$\begin{aligned}&|S_1(\boldsymbol{\tau})-S_1(\boldsymbol{\tau}_0)-\dot{S}_{11}(\boldsymbol{\tau}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\dot{S}_{12}(\boldsymbol{\tau}_0)[\psi-\psi_0]|\\ &\quad = o(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|)+O(\|\psi-\psi_0\|_2^{\alpha})\end{aligned}$$

and

$$\begin{aligned}&|S_2(\boldsymbol{\tau})[\mathbf{h}^*]-S_2(\boldsymbol{\tau}_0)[\mathbf{h}^*]-\dot{S}_{21}(\boldsymbol{\tau}_0)[\mathbf{h}^*](\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\dot{S}_{22}(\boldsymbol{\tau}_0)[\mathbf{h}^*,\psi-\psi_0]|\\ &\quad = o(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|)+O(\|\psi-\psi_0\|_2^{\alpha}).\end{aligned}$$

Then

$$n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=\mathbf{A}_0^{-1}n^{1/2}\mathbb{P}_n[m_1(\boldsymbol{\tau}_0;\mathbf{v})-m_2(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}^*]]+o_p(1)\to(0,\mathbf{A}_0^{-1}\mathbf{B}_0\mathbf{A}_0^{-1}),$$

in distribution, as $n\to\infty$, *where* $\mathbf{B}_0 = P[m_1(\boldsymbol{\tau}_0;\mathbf{v})-m_2(\boldsymbol{\tau}_0;\mathbf{v})[\mathbf{h}^*]]^{\otimes 2}$ *and* $\mathbf{A}_0$ *is given in assumption* B3.

Remark 8 Lemma 1 is used to derive the consistency of $\hat{\boldsymbol{\tau}}$. The similar entropy calculations are also used to prove Theorems 1–4. Lemma 2 is a key result to derive the consistency and the rate of convergence of $\hat{\boldsymbol{\tau}}$. Lemma 3 shows $\hat{\boldsymbol{\tau}}$ is asymptotically consistent to $\boldsymbol{\tau}_0$. Lemma 4, which is Theorem 6.1 of Wellner and Zhang (2007), is used to develop the asymptotic normality of $\hat{\boldsymbol{\beta}}$. This theorem generalizes the theorem for the asymptotic normality of semiparametric M-estimators developed by Huang (1996) to accommodate the quasi-likelihood estimation.

6.2 Proof of Lemma 1

According to the bracketing number calculation in Shen and Wong (1994), for any $\eta > 0$ and $\varepsilon \le \eta$, the logarithm of bracketing number of $\mathcal{S}_n$, computed with $L_2(P)$, is bounded by $Cq_n \log(\eta/\varepsilon)$. Note that the neighborhood $\mathbf{B}(\eta) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \eta\}$ can be covered by $C(\eta/\varepsilon)^d$ balls with radius ε. By the Cauchy–Schwarz inequality, $|\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_t - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}_s| \le \|\mathbf{x}\|\|\boldsymbol{\beta}_t-\boldsymbol{\beta}_s\|$. Theorem 9.23 of Kosorok (2008) yields the bracketing number of the class $\{\mathbf{x}^\mathsf{T}\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le \eta\}$ is $(\eta/\varepsilon)^d$, up to a constant. Thus, for any $\mathbf{x}^\mathsf{T}\boldsymbol{\beta}+\psi(z)$, there exist some $\boldsymbol{\beta}_r$, for $r = 1,\ldots,C(\eta/\varepsilon)^d$, and brackets $\left[\psi_s^L, \psi_s^R\right]$,

for $s = 1, \ldots, (\eta/\varepsilon)^{Cq_n}$, such that $\mathbf{x}^\intercal\boldsymbol{\beta}_r - \varepsilon \le \mathbf{x}^\intercal\boldsymbol{\beta} \le \mathbf{x}^\intercal\boldsymbol{\beta}_r + \varepsilon$ and $\psi_s^L \le \psi \le \psi_s^R$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_r\| \le C\varepsilon$ and $\|\psi_s^R - \psi_s^L\|_2 \le C\varepsilon$, and hence

$$F(\mathbf{x}^\intercal\boldsymbol{\beta}_r - \varepsilon + \psi_s^L) \le \mu \le F(\mathbf{x}^\intercal\boldsymbol{\beta}_r + \varepsilon + \psi_s^R),$$

for a monotone and smooth link function F. By the mean value theorem and the Cauchy–Schwarz inequality as well as the boundedness of first derivative of F, $\|F(\mathbf{x}^\intercal\boldsymbol{\beta}_r + \varepsilon + \psi_s^R) - F(\mathbf{x}^\intercal\boldsymbol{\beta}_r - \varepsilon + \psi_s^L)\|_2$ can be bounded by $C\varepsilon$. Because $Q(\boldsymbol{\tau}; \mathbf{v})$ is monotone increasing for $\mu \le y$, and monotone decreasing for $\mu > y$, the class of functions $Q(\boldsymbol{\tau}; \mathbf{v})$ with $\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta$ can be covered by $[A_{r,s}^L, A_{r,s}^R]$, where $A_{r,s}^L = \int_y^{\mu_{r,s}^L} \frac{y-s}{\sigma^2 V(s)} ds$ and $A_{r,s}^R = \int_y^{\mu_{r,s}^R} \frac{y-s}{\sigma^2 V(s)} ds$ with

$$\begin{aligned}
\mu_{r,s}^L &= F(\mathbf{x}^\intercal\boldsymbol{\beta}_r - \varepsilon + \psi_s^L)1[\mu \le y] + F(\mathbf{x}^\intercal\boldsymbol{\beta}_r + \varepsilon + \psi_s^R)1[\mu > y], \\
\mu_{r,s}^R &= F(\mathbf{x}^\intercal\boldsymbol{\beta}_r + \varepsilon + \psi_s^R)1[\mu \le y] + F(\mathbf{x}^\intercal\boldsymbol{\beta}_r - \varepsilon + \psi_s^L)1[\mu > y].
\end{aligned}$$

Therefore, a Taylor expansion and the Cauchy–Schwarz inequality yield $P(A_{r,s}^R - A_{r,s}^L)^2 \le C\varepsilon^2$. According to the inequality $2(e^{|x|} - |x| - 1) \le x^2 e^{|x|}$ and conditions C3–C7, we have $\|A_{r,s}^R - A_{r,s}^L\|_{P,B}^2 \le C P(A_{r,s}^R - A_{r,s}^L)^2 \le C\varepsilon^2$. This implies Lemma 1. □

6.3 Proof of Lemma 2

Let $\mathbb{M}(\boldsymbol{\tau}) = P\,Q(\boldsymbol{\tau}; \mathbf{v})$ and $\mathbb{M}_n(\boldsymbol{\tau}) = \mathbb{P}_n Q(\boldsymbol{\tau}; \mathbf{v})$. A Taylor expansion yields $\mathbb{M}(\boldsymbol{\tau}_0) - \mathbb{M}(\boldsymbol{\tau}) = \sigma^{-2}(1 - \xi_1)P\left[(\mu - \mu_0)^2 V^{-1}(\mu + \xi_1(\mu_0 - \mu))\right]$, for some $0 < \xi_1 < 1$. Because the variance function $V(\cdot)$ and the first derivative of $F(\cdot)$ are assumed to be bounded, there exist $0 < C_1 < C_2$ such that

$$C_1 d_2^2(\boldsymbol{\tau}, \boldsymbol{\tau}_0) \le \mathbb{M}(\boldsymbol{\tau}_0) - \mathbb{M}(\boldsymbol{\tau}) \le C_2 d_2^2(\boldsymbol{\tau}, \boldsymbol{\tau}_0),$$

where $d_2^2(\boldsymbol{\tau}, \boldsymbol{\tau}_0) = \|\mathbf{x}^\intercal(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \psi(z) - \psi_0(z)\|_2^2$. Let $g_1(\mathbf{x}) = \mathbf{x}^\intercal(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and $g_2(z) = \psi(z) - \psi_0(z)$. Cauchy–Schwarz inequality and law of total expectation yield $\{E[g_1(\mathbf{x})g_2(z)]\}^2 \le E_z[g_2^2(z)]E_z[\{E_{\mathbf{x}|z}[g_1(\mathbf{x})|z]\}^2]$. By the orthogonality of a conditional expectation, there exists $0 < \xi_2 < 1$ such that $[Eg_1(\mathbf{x})g_2(z)]^2 \le \xi_2 E[g_1^2(\mathbf{x})]E[g_2^2(z)]$. In view of Lemma 25.86 of van der Vaart (2000) and the finite second moment of $\mathbf{x}$, there exist $0 < C_3 < C_4$ such that

$$C_3\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2 \le \mathbb{M}(\boldsymbol{\tau}_0) - \mathbb{M}(\boldsymbol{\tau}) \le C_4\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2.$$

This completes the proof of Lemma 2. □

6.4 Proof of Lemma 3

We verify the conditions of Theorem 5.7 in van der Vaart (2000) to prove the consistency of $\hat{\boldsymbol{\tau}}$. Lemma 1 implies that the class of functions $Q(\boldsymbol{\tau}; \mathbf{v})$ with $\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta$

is a Glivenko–Cantelli class. Thus, $\sup|\mathbb{M}_n(\boldsymbol{\tau}) - \mathbb{M}(\boldsymbol{\tau})| = o_p(1)$, for $\boldsymbol{\tau}$ in a neighborhood of $\boldsymbol{\tau}_0$. The first condition of the theorem holds. It follows from Lemma 2 that

$$\sup_{\|\boldsymbol{\tau}-\boldsymbol{\tau}_0\|_2\geq\varepsilon} \mathbb{M}(\boldsymbol{\tau}) \leq \mathbb{M}(\boldsymbol{\tau}_0) - C\varepsilon^2 < \mathbb{M}(\boldsymbol{\tau}_0).$$

The second condition of the theorem is verified.

According to Jackson's theorem for polynomials (de Boor 2001), there exists a spline $\psi_{0,n} \in \mathcal{S}_n$ of order $l \geq 2$ such that $\|\psi_{0,n} - \psi_0\|_\infty = O(n^{-r\nu})$, for $1/(2r+2) < \nu < 1/(2r)$. Let $\boldsymbol{\tau}_{0,n} = (\boldsymbol{\beta}_0^\mathsf{T}, \psi_{0,n})^\mathsf{T}$ and $\mu_{0,n} = F(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0 + \psi_{0,n})$. Observe that

$$\mathbb{M}_n(\hat{\boldsymbol{\tau}}) - \mathbb{M}_n(\boldsymbol{\tau}_0) \geq \mathbb{M}_n(\boldsymbol{\tau}_{0,n}) - \mathbb{M}_n(\boldsymbol{\tau}_0) = I_{n1} + I_{n2},$$

where $I_{n1} = (\mathbb{P}_n - P)[Q(\boldsymbol{\tau}_{0,n}; \mathbf{v}) - Q(\boldsymbol{\tau}_0; \mathbf{v})]$ and $I_{n2} = P[Q(\boldsymbol{\tau}_{0,n}; \mathbf{v}) - Q(\boldsymbol{\tau}_0; \mathbf{v})]$. Write $I_{n1} = n^{-r\nu+\varepsilon}(\mathbb{P}_n - P)\{[Q(\boldsymbol{\tau}_{0,n}; \mathbf{v}) - Q(\boldsymbol{\tau}_0; \mathbf{v})]/n^{-r\nu+\varepsilon}\}$, for $0 < \varepsilon < 1/2 - n\nu$. As shown in the proof of Lemma 1, $\mathcal{S}_n$ is a Donsker class, and so is the class of functions $\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0 + \psi$, for $\psi \in \mathcal{S}_n$ and $\|\psi - \psi_0\|_2 \leq \eta$. Because the function $m \mapsto F(m)$ has the bounded first derivative on $\mathcal{M}$, the preservation theorem of Donsker class yields that the class of functions $F(\mathbf{x}^\mathsf{T}\boldsymbol{\beta}_0 + \psi)$ is also a Donsker class. Moreover, because the function $Q(\boldsymbol{\tau}; \mathbf{v}) = \int_y^\mu \frac{y-s}{\sigma^2 V(s)} ds$ is Lipschitz with respect to μ on $\mathcal{F}$, it implies that the class of functions $Q(\boldsymbol{\beta}_0, \psi; \mathbf{v}) - Q(\boldsymbol{\beta}_0, \psi_0; \mathbf{v})$ with $\psi \in \mathcal{S}_n$ and $\|\psi - \psi_0\|_2 \leq \eta$ is a Donsker class, and $P[Q(\boldsymbol{\tau}_{0,n}; \mathbf{v}) - Q(\boldsymbol{\tau}_0; \mathbf{v})]^2/n^{-2n\nu+2\varepsilon} \to 0$, as $n \to \infty$. In view of Lemma 19.24 of van der Vaart (2000), $I_{n1} = o_p(n^{-r\nu+\varepsilon}n^{-1/2}) = o_p(n^{-2r\nu})$. Lemma 2 implies $I_{n2} \geq -C\|\psi_{0,n} - \psi_0\|_\infty^2 = -O(n^{-2r\nu})$. It follows that

$$\mathbb{M}_n(\hat{\boldsymbol{\tau}}) - \mathbb{M}_n(\boldsymbol{\tau}_0) > o_p(n^{-2r\nu}) - O(n^{-2r\nu}) = -o_p(1).$$

We conclude that Theorem 5.7 of van der Vaart (2000) applies and yields $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2 = o_p(1)$. □

6.5 Proof of efficient score and information bound

We derive the efficient score and information bound following Bickel et al. (1993) and Huang et al. (2007). Consider the model

$$E(y|\mathbf{w}) = g_{\boldsymbol{\nu}}(\mathbf{w}), \tag{10}$$

where $g_{\boldsymbol{\nu}}(\mathbf{w})$ is a known function indexed by $\nu \in \mathbb{R}^d$. Let $f(\mathbf{v})$ be the joint density of $\boldsymbol{\nu}$. It is assumed that $f(\mathbf{v})$ is smooth and bounded on the support of $\mathbf{v}$. Following the same arguments as those in Lemma 4 of Huang et al. (2007), we can show that the efficient score for $\boldsymbol{\nu}$ is given by

$$\ell^*_{\boldsymbol{\nu},f}(\mathbf{w}, \varepsilon) = \frac{\partial g_{\boldsymbol{\nu}}(\mathbf{w})}{\partial \boldsymbol{\nu}}[E(\varepsilon^2|\mathbf{w})]^{-1}\varepsilon, \tag{11}$$

where $\varepsilon = y - E(y|\mathbf{w})$. Let $\mathcal{P} = \{P_{(\boldsymbol{\beta},\psi,f)} : \boldsymbol{\beta} \in \boldsymbol{\Theta}, \psi \in \boldsymbol{\Psi}, f \in L_2(P)\}$ denote the model specified by (1). We follow the notations in Bickel et al. (1993) and define

$$\mathcal{P}_1 = \{P_{(\boldsymbol{\beta},\psi_0,f_0)} : \boldsymbol{\beta} \in \boldsymbol{\Theta}\}, \quad \mathcal{P}_2 = \{P_{(\boldsymbol{\beta}_0,\psi,f_0)} : \psi \in \boldsymbol{\Psi}\},$$
$$\mathcal{P}_3 = \{P_{(\boldsymbol{\beta}_0,\psi_0,f)} : f \in L_2(P)\}, \quad \mathcal{P}_{13} = \{P_{(\boldsymbol{\beta},\psi_0,f)} : \boldsymbol{\beta} \in \boldsymbol{\Theta}, f \in L_2(P)\}.$$

Let $\dot{\mathcal{P}}_1$, $\dot{\mathcal{P}}_2$, and $\dot{\mathcal{P}}_3$ be the tangent spaces of $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$, respectively, and let $\dot{\ell}_{\boldsymbol{\beta}}$ be the score function for $\boldsymbol{\beta}$ in model $\mathcal{P}_1$. According to the project properties [see section 3.4 of Bickel et al. (1993) and appendix A6 of Huang et al. (2007)], the efficient score for $\boldsymbol{\beta}$ in model $\mathcal{P}$ is given by

$$\ell^*_{\boldsymbol{\beta}} = \dot{\ell}_{\boldsymbol{\beta}} - \Pi(\dot{\ell}_{\boldsymbol{\beta}}|\dot{\mathcal{P}}_2 + \dot{\mathcal{P}}_3) = \Pi(\dot{\ell}_{\boldsymbol{\beta}}|\dot{\mathcal{P}}_3{}^{\perp}) - \Pi[\Pi(\dot{\ell}_{\boldsymbol{\beta}}|\dot{\mathcal{P}}_3^{\perp})|\Pi_{\dot{\mathcal{P}}_3^{\perp}}\dot{\mathcal{P}}_2],$$

where Π is a projection operator. Because $\Pi(\dot{\ell}_{\boldsymbol{\beta}}|\dot{\mathcal{P}}_3{}^{\perp})$ is the efficient score function for $\boldsymbol{\beta}$ in model $\mathcal{P}_{13}$, (11) applies and yields

$$\Pi(\dot{\ell}_{\boldsymbol{\beta}}|\dot{\mathcal{P}}_3{}^{\perp}) = \mathbf{x}\Delta_0\Sigma_0^{-1}(y - \mu_0).$$

For one-dimensional parametric submodel $(\boldsymbol{\beta}, \psi_\eta)$ with $\psi_\eta|_{\eta=0} = \psi$ and $\partial\psi_\eta/\partial\eta|_{\eta=0} = h \in \mathcal{H}$, replacing ψ by ψ_η and applying (11), we can show that

$$\Pi_{\dot{\mathcal{P}}_3^{\perp}}\dot{\mathcal{P}}_2 = \{\Delta_0\Sigma_0^{-1}(y - \mu_0)\mathbf{h} : \mathbf{h} \in \mathcal{H}^d\}.$$

It follows that

$$\ell^*_{\boldsymbol{\beta}} = (\mathbf{x} - \boldsymbol{\phi}^*)\Delta_0\Sigma_0^{-1}(y - \mu_0),$$

where $\boldsymbol{\phi}^*$ satisfies

$$E[(\mathbf{x} - \boldsymbol{\phi}^*)\Delta_0^2\Sigma_0^{-1}h] = 0,$$

for any $h \in \mathcal{H}$. By the law of total expectation, the least favorable direction is given by

$$\boldsymbol{\phi}^*(z) = \frac{E_{\mathbf{x}|z}[\mathbf{x}\Delta_0^2\Sigma_0^{-1}|z]}{E_{\mathbf{x}|z}[\Delta_0^2\Sigma_0^{-1}|z]}.$$

□

6.6 Proof of rate of convergence

We apply Theorem 3.4.1 of van der Vaart and Wellner (1996) to prove the rate of convergence of $\hat{\boldsymbol{\tau}}$. Let $\theta = \mathbf{x}^\intercal\boldsymbol{\beta} + \psi(z)$. Denote $\theta_0 = \mathbf{x}^\intercal\boldsymbol{\beta}_0 + \psi_0(z)$ and $\theta_n = \mathbf{x}^\intercal\boldsymbol{\beta}_0 + \psi_{0,n}(z)$. Also denote by $\hat{\theta} = \mathbf{x}^\intercal\hat{\boldsymbol{\beta}} + \hat{\psi}(z)$ the estimate of θ_0. Define $l(\theta) =$

$\sigma^{-2}\int_y^{F(\theta)}(y-s)/V(s)ds$ and $\mathbb{M}(\theta) = Pl(\theta)$. By the similar entropy calculation to that in Lemma 1, we can show that, for any $\eta > 0$,

$$J_{[]}(\eta, \{l(\theta) - l(\theta_n) : \psi \in \mathcal{S}_n, \|\theta - \theta_n\|_2 \le \eta\}, \|\cdot\|_{P,B}) \le Cq_n^{1/2}\eta.$$

Moreover, for any $l(\theta) - l(\theta_n)$ with $\psi \in \mathcal{S}_n$ and $\|\theta - \theta_n\|_2 \le \eta$, the inequality $2(e^{|x|} - |x| - 1) \le x^2 e^{|x|}$ and conditions C3–C7 as well as Cauchy–Schwarz inequality yield $\|l(\theta) - l(\theta_n)\|_{P,B}^2 \le C\eta^2$. In view of Lemma 3.4.3 of van der Vaart and Wellner (1996),

$$P\left[\sup_{\eta/2\le\|\theta-\theta_n\|_2\le\eta} |\mathbb{G}_n l(\theta) - \mathbb{G}_n l(\theta_n)|\right] \le C\phi_n(\eta)$$

with $\phi_n(\eta) = q_n^{1/2}\eta + q_n/n^{1/2}$. Obviously, $\phi_n(\eta)/\eta$ is decreasing in η. Moreover, because $l(\theta)$ is Lipschitz with respect to θ, the consistency of $\hat{\boldsymbol{\tau}}$ and $\|\psi_{0,n} - \psi_0\|_\infty = O(n^{-r\nu})$ yield that $\mathbb{M}(\theta_n) - \mathbb{M}(\hat{\theta}) = o_p(1)$. Therefore, by choosing the distance d_n defined in Theorem 3.4.1 of van der Vaart and Wellner (1996) to be $d_n^2(\theta_n, \hat{\theta}) = \mathbb{M}(\theta_n) - \mathbb{M}(\hat{\theta})$, we obtain $r_n^2[\mathbb{M}(\theta_n) - \mathbb{M}(\hat{\theta})] = O_p(1)$, where r_n satisfies $r_n^2\phi_n(1/r_n) \le n^{1/2}$ for every n. It follows that $r_n = n^{\min(r\nu,(1-\nu)/2)}$. Lemma 2 implies that $\mathbb{M}(\theta_0) - \mathbb{M}(\hat{\theta}) \ge C\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2^2$ and $\mathbb{M}(\theta_n) - \mathbb{M}(\theta_0) \ge -C\|\psi_{0,n} - \psi_0\|_\infty^2$. Observe that $\mathbb{M}(\theta_0) - \mathbb{M}(\hat{\theta}) = \mathbb{M}(\theta_0) - \mathbb{M}(\theta_n) + \mathbb{M}(\theta_n) - \mathbb{M}(\hat{\theta})$. It follows that

$$C\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2^2 \le \mathbb{M}(\theta_0) - \mathbb{M}(\hat{\theta}) \le O(n^{-2r\nu}) + O_p(r_n^{-2}) = O_p(r_n^{-2}).$$

This yields the rate of convergence of $\hat{\boldsymbol{\tau}}$. □

6.7 Proof of asymptotic normality

We verify conditions of Lemma 4 to show the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Condition B1 is valid because of the rate of convergence of $\hat{\boldsymbol{\tau}}$ with $\gamma = r/(1+2r)$, for $r \ge 3$. Condition B2 holds due to the model assumption (1). For condition B3, by applying for law of total expectation, we have $E_{\mathbf{x}}\left\{E_{\mathbf{x}|z}[\Delta_0^2 V^{-1}(\mu_0)(\mathbf{x} - \mathbf{h}^*)]|z\right\}h(z) = 0$. It follows that

$$\mathbf{h}^* = \frac{E_{\mathbf{x}|z}[\mathbf{x}\Delta_0^2 V^{-1}(\mu_0)|z]}{E_{\mathbf{x}|z}[\Delta_0^2 V^{-1}(\mu_0)|z]}.$$

The first part of B4 automatically holds because $\hat{\boldsymbol{\beta}}$ satisfies the quasi-score function $S_{1n}(\hat{\boldsymbol{\tau}}) = 0$. We only need to verify that $S_{2n}(\hat{\boldsymbol{\tau}})[\mathbf{h}^*] = o_p(n^{-1/2})$. According to Jackson's theorem for polynomials (de Boor 2001), there exist $h_{n,s}^* \in \mathcal{S}_n$ of order $l \ge 2$ such that $\|h_s^* - h_{n,s}^*\|_\infty = O(n^{-r\nu})$, for $1/(2r+2) < \nu < 1/(2r)$, $s = 1, \ldots, d$. Because $(\hat{\boldsymbol{\beta}}, \hat{\psi})$ maximizes $\mathbb{P}_n Q(\boldsymbol{\beta}, \psi; \mathbf{v})$ over the region $(\hat{\boldsymbol{\beta}}, \hat{\psi} + \varepsilon h)$, for any $h \in \mathcal{S}_n$, we have $\lim_{\varepsilon\downarrow 0}\frac{d}{d\varepsilon}\mathbb{P}_n Q(\hat{\boldsymbol{\beta}}, \hat{\psi} + \varepsilon h; \mathbf{v}) = \mathbb{P}_n m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h] = 0$. Therefore, to show

$\mathbb{P}_n m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^*] = o_p(n^{-1/2})$, it is equivalent to showing $\mathbb{P}_n m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^* - h_{n,s}^*] = o_p(n^{-1/2})$. Write $\mathbb{P}_n m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^* - h_{n,s}^*] = I_{n3} + I_{n4}$ with $I_{n3} = (\mathbb{P}_n - P)m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^* - h_{n,s}^*]$ and $I_{n4} = Pm_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^* - h_{n,s}^*]$. Some entropy calculation yields, for any $\eta > 0$,

$$J_{[]}(\eta, \{m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^* - h] : h \in \mathcal{S}_n, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta, \|h_s^* - h\|_\infty \le \eta\}, \|\cdot\|_{P,B}) \le C q_n^{1/2} \eta.$$

Moreover, in view of conditions C3–C7 and the inequality $2(e^{|x|} - |x| - 1) \le x^2 e^{|x|}$, we can show that $\|m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^* - h]\|_{P,B}^2 \le C\|h_s^* - h\|_\infty^2 \le C\eta^2$, for $\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta$ and $\|h - h_s^*\|_\infty \le \eta$. Therefore, Lemma 3.4.3 of van der Vaart and Wellner (1996) applies and yields

$$E\left[\sup_{\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta, \|h_s^* - h\|_\infty \le \eta} |(\mathbb{P}_n - P)m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^* - h]|\right] = o(n^{-1/2}).$$

It follows that $I_{n3} = o_p(n^{-1/2})$. Moreover, Cauchy–Schwarz inequality and conditions C3–C7 yield

$$I_{n4}^2 \le C\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2^2 \|h_s^* - h_{n,s}^*\|_\infty^2 = O_p(n^{-2r/(1+2r)}) O_p(n^{-2r/(1+2r)}).$$

The last equality holds because of the rate of convergence of $\hat{\boldsymbol{\tau}}$. Thus, condition B4 holds. To verify condition B5 is equivalent to showing $\mathbb{G}_n[m_1(\hat{\boldsymbol{\tau}}; \mathbf{v}) - m_1(\boldsymbol{\tau}_0; \mathbf{v})] = o_p(1)$ and $\mathbb{G}_n[m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[\mathbf{h}^*] - m_2(\boldsymbol{\tau}_0; \mathbf{v})[\mathbf{h}^*]] = o_p(1)$. We only show the second equation because the proof of the first equation is similar. Using the similar arguments to those in the proof of condition B4, we can show that, for any $\eta > 0$,

$$J_{[]}(\eta, \{m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^*] - m_2(\boldsymbol{\tau}_0; \mathbf{v})[h_s^*] : \boldsymbol{\tau} \in \Theta \times \mathcal{F}, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta\}, \|\cdot\|_{P,B}) \le C q_n^{1/2} \eta,$$

$s = 1, \ldots, d$. Furthermore, for any $m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^*] - m_2(\boldsymbol{\tau}_0; \mathbf{v})[h_s^*]$ with $\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta$, mean value theorem and Cauchy–Schwarz inequality yield

$$P\left[m_2(\boldsymbol{\tau}; \mathbf{v})[h_s^*] - m_2(\boldsymbol{\tau}_0; \mathbf{v})[h_s^*]\right]^2 \le C\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2 \|h_s^*\|_\infty^2 \le C\eta^2.$$

Therefore, in view of Lemma 3.4.3 of van der Vaart and Wellner (1996), $(\mathbb{P}_n - P)[m_2(\hat{\boldsymbol{\tau}}; \mathbf{v})[h_s^*] - m_2(\boldsymbol{\tau}_0; \mathbf{v})[h_s^*]] = o_p(n^{-1/2})$, and hence condition B5 holds. Finally, condition B6 holds with $\alpha = 2$ by applying for a Taylor expansion and the Cauchy–Schwarz inequality as well as conditions C3–C7. We conclude that Lemma 4 applies and yields the asymptotic normality of $\hat{\boldsymbol{\beta}}$. □

6.8 Proof of variance estimation

Denote $\rho_s(\boldsymbol{\tau}, h) = [m_{1,s}(\boldsymbol{\tau}; \mathbf{v}) - m_2(\boldsymbol{\tau}; \mathbf{v})[h]]^2$, $s = 1, \ldots, d$. We first show that $\|\hat{h}_s^* - h_s^*\|_2 \to 0$, in probability, as $n \to \infty$. According to Jackson's theorem for polynomials (de Boor 2001), there exist $h_{n,s}^* \in \mathcal{S}_n$ with order of $l \ge 2$ such that

$\|h_s^* - h_{n,s}^*\|_\infty = O(n^{-r\nu})$, for $1/(2r+2) < \nu < 1/(2r)$. Because $\hat{h}_s^*$ minimize $\rho_s(\hat{\boldsymbol{\tau}}, h)$, for all $h \in \mathcal{S}_n$, we have $\mathbb{P}_n \rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) \le \mathbb{P}_n \rho_s(\hat{\boldsymbol{\tau}}, h_{n,s}^*)$. It implies that

$$\begin{aligned}&\mathbb{P}_n[\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)]\\ &\quad \le (\mathbb{P}_n - P)[\rho_s(\hat{\boldsymbol{\tau}}, h_{n,s}^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)] + P[\rho_s(\hat{\boldsymbol{\tau}}, h_{n,s}^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)].\end{aligned}$$

By some entropy calculation, it is readily to show that, for any $\eta > 0$,

$$\mathcal{L}_s = \{\rho_s(\boldsymbol{\tau}, h) - \rho_s(\boldsymbol{\tau}, h_s^*) : \boldsymbol{\beta} \in \Theta, \psi, h \in \mathcal{S}_n, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta, \|h - h_s^*\|_2 \le \eta\}$$

are Glivenko–Cantelli classes. It implies that $(\mathbb{P}_n - P)[\rho_s(\hat{\boldsymbol{\tau}}, h_{n,s}^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)] = o_p(1)$. Moreover, continuous mapping theorem and dominated convergence theorem yield $P[\rho_s(\hat{\boldsymbol{\tau}}, h_{n,s}^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)] = o_p(1)$. It follows that $\mathbb{P}_n[\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)] \le o_p(1)$. In view of Glivenko–Cantelli theorem, we have $(\mathbb{P}_n - P)\rho_s(\hat{\boldsymbol{\tau}}, h_s^*) = o_p(1)$. Hence,

$$\begin{aligned}\mathbb{P}_n \rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) &\le (\mathbb{P}_n - P)\rho_s(\hat{\boldsymbol{\tau}}, h_s^*) + P\rho_s(\hat{\boldsymbol{\tau}}, h_s^*) + o_p(1)\\ &= P\rho_s(\hat{\boldsymbol{\tau}}, h_s^*) + o_p(1). \qquad (12)\end{aligned}$$

Continuous mapping theorem and dominated convergence theorem as well as the consistency of $\hat{\boldsymbol{\tau}}$ yield $P[\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) - \rho_s(\boldsymbol{\tau}_0, \hat{h}_s^*)] \to 0$ and $P[\rho_s(\hat{\boldsymbol{\tau}}, h_s^*) - \rho_s(\boldsymbol{\tau}_0, h_s^*)] \to 0$. It follows that

$$\begin{aligned}0 \le P[\rho_s(\boldsymbol{\tau}_0, \hat{h}_s^*) - \rho_s(\boldsymbol{\tau}_0, h_s^*)] &= o_p(1) + P[\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) - \rho_s(\hat{\boldsymbol{\tau}}, h_s^*)]\\ &\le o_p(1) - (\mathbb{P}_n - P)\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*).\end{aligned}$$

The last inequality holds due to (12). In view of Glivenko–Cantelli theorem, $(\mathbb{P}_n - P)\rho_s(\hat{\boldsymbol{\tau}}, \hat{h}_s^*) = o_p(1)$. It follows that $P[\rho_s(\boldsymbol{\tau}_0, \hat{h}_s^*) - \rho_s(\boldsymbol{\tau}_0, h_s^*)] = o_p(1)$. By the uniqueness of h_s^*, the event $\|\hat{h}_s^* - h_s^*\|_2 \ge \varepsilon$ is the subset of the event $P[\rho_s(\boldsymbol{\tau}_0, \hat{h}_s^*) - \rho_s(\boldsymbol{\tau}_0, h_s^*)] > 0$ and the latter approaches to 0, in probability, as $n \to \infty$. This implies that $\|h_s^* - \hat{h}_s^*\|_2 = o_p(1)$.

Next we show the consistency of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. Denote $\rho_1(\boldsymbol{\tau}, \mathbf{h}) = [m_1(\boldsymbol{\tau}; \mathbf{v}) - m_2(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}]]^{\otimes 2}$ and $\rho_2(\boldsymbol{\tau}, \mathbf{h}) = m_{11}(\boldsymbol{\tau}; \mathbf{v}) - m_{21}(\boldsymbol{\tau}; \mathbf{v})[\mathbf{h}]$. By some entropy calculation, we can similarly show that

$$\Re_i = \{\rho_i(\boldsymbol{\tau}, \mathbf{h}) : \boldsymbol{\beta} \in \boldsymbol{\Theta}, \psi \in \mathcal{S}_n, \mathbf{h} \in \mathcal{S}_n^d, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \le \eta, \|\mathbf{h} - \mathbf{h}^*\|_2 \le \eta\}$$

are Glivenko–Cantelli classes, i = 1,2. In view of Glivenko–Cantelli theorem, $(\mathbb{P}_n - P)\rho_i(\hat{\boldsymbol{\tau}}, \hat{\mathbf{h}}^*) = o_p(1)$. Hence, continuous mapping theorem and dominated convergence theorem yield

$$\begin{aligned}\hat{\mathbf{B}} - \mathbf{B}_0 &= (\mathbb{P}_n - P)\rho_1(\hat{\boldsymbol{\tau}}, \hat{\mathbf{h}}^*) + P[\rho_1(\hat{\boldsymbol{\tau}}, \hat{\mathbf{h}}^*) - \rho_1(\boldsymbol{\tau}_0, \mathbf{h}^*)] = o_p(1),\\ \hat{\mathbf{A}} - \mathbf{A}_0 &= (\mathbb{P}_n - P)\rho_2(\hat{\boldsymbol{\tau}}, \hat{\mathbf{h}}^*) + P[\rho_2(\hat{\boldsymbol{\tau}}, \hat{\mathbf{h}}^*) - \rho_2(\boldsymbol{\tau}_0, \mathbf{h}^*)] = o_p(1).\end{aligned}$$

This completes the proof of Theorem 4. □

6.9 Proof of Corollary 1

Denote $\hat{\mathcal{O}} = \hat{\mathcal{O}}_{11} - \hat{\mathcal{O}}_{12}\hat{\mathcal{O}}_{22}^{-1}\hat{\mathcal{O}}_{21}$. Using the similar arguments to those in the proof of Theorem 4, we can show that

$$\hat{\mathcal{O}} - \mathbf{B}_0 = (\mathbb{P}_n - P)\rho_1(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\phi}}^*) + P[\rho_1(\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\phi}}^*) - \rho_1(\boldsymbol{\tau}_0, \boldsymbol{\phi}^*)] = o_p(1).$$

Some entropy calculation and law of large numbers yield that $\hat{\mathcal{E}} \to \mathbf{B}_0$, in probability, as $n \to \infty$. □

Acknowledgements The author is grateful to the editor and two reviewers for their useful comments and constructive suggestions which led to significant improvement in the presentation. The author also expresses his thanks to Dr. Chin-Shang Li for kindly providing the crab data.

References

Bickel, P. J., Klaassen, C. A., Ritov, Y., Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press.

Brockmann, J. (1996). Satellite male groups in horseshoe crabs, *Limulus polyphemus*. *Ethology*, *102*, 1–21.

Chen, J., Fan, J., Li, K., Zhou, H. (2006). Local quasi-likelihood estimation with data missing at random. *Statistica Sinica*, *16*, 1071–1100.

de Boor, C. (2001). *A practical guide to splines*. New York: Spring-Verlag.

Fan, J., Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*, 927–943.

Fan, J., Heckman, N. E., 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.

Härdle, W., Mammen, E., Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, *93*, 1461–1474.

Hua, L., Zhang, Y. (2012). Spline-based semiparametric projected generalized estimating equation method for panel count data. *Biostatistics*, *13*, 440–454.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censored data. *Annals of Statistics*, *24*, 540–568.

Huang, J. Z., Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, *62*, 793–802.

Huang, J. Z., Zhang, L., Zhou, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scandinavian Journal of Statistics*, *34*, 451–477.

Kooperberg, C., Stone, C. J., Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, *90*, 78–94.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Dordrecht: Springer.

Li, C. S. (2012). Lack-of-fit tests for generalized linear models via splines. *Communications in Statistics-Theory and Methods*, *41*, 4240–4250.

Lu, M., Loomis, D. (2013). Spline-based semiparametric estimation of partially linear poisson regression with single-index models. *Journal of Nonparametric Statistics*, *25*, 905–922.

Lu, M., Zhang, Y., Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association*, *104*, 1060–1070.

Mammen, E., van de Geer, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *The Annals of Statistics*, *90*, 1014–1035.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, *11*, 59–67.

McCullagh, P., Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, *51*, 874–887.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, *11*, 735–757.

Schumaker, L. (1981). *Spline functions: Basic theory*. New York: Wiley & Son.

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

Shen, X., Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics*, *22*, 580–615.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics*, *14*, 590–606.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture (with discussion). *The Annals of Statistics*, *25*, 1371–1470.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer-Verlag.

Wahba, G. (1990). *Spline models for observational data* (Vol. 59). Siam.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the GaussNewton method. *Biometrika*, *61*, 439–447.

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

Xue, L., Liang, H. (2010). Polynomial spline estimation for a generalized additive coefficient model. *Scandinavian Journal of Statistics*, *37*, 26–46.