Ann Inst Stat Math (2016) 68:1111–1134
DOI 10.1007/s10463-015-0527-8

Spline-based semiparametric estimation of a zero-inflated Poisson regression single-index model

Minggen Lu[1] · Chin-Shang Li[2]

Received: 18 January 2014 / Revised: 6 April 2015 / Published online: 2 July 2015

Abstract When the number of zeros in a count dataset exceeds the accommodation of the probability mass of a regular Poisson distribution at zero, the zero-inflated Poisson (ZIP) distribution is often used. To characterize the potential non-linear effects of covariates and avoid the "curse of dimensionality", we propose a spline-based ZIP regression single-index model. B-splines are employed to estimate the unknown smooth function. A modified Fisher scoring method is proposed to simultaneously estimate the linear coefficients and the regression function. It is shown that the spline estimator of the nonparametric component is uniformly consistent, and achieves the optimal convergence rate under the smooth condition, and that the estimators of regression parameters are asymptotically normal and efficient. The spline-based semiparametric likelihood ratio test is also established. Moreover, a direct and consistent variance estimation method based on least-squares estimation is proposed. Simulations are performed to evaluate the proposed method.

Keywords B-spline · Likelihood estimator · Single-index model · Zero-inflated Poisson regression

1 Introduction

Count data usually contain large numbers of zeros, which can be seen in many disciplines, e.g., biomedical studies, environmental economics, among others. When the

✉ Chin-Shang Li
cssli@ucdavis.edu

Minggen Lu
minggenl@unr.edu

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

[2] Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, CA 95616, USA

number of zeros exceeds the accommodation of the probability mass of a regular Poisson distribution at zero, the ZIP distribution is often used to analyze the type of count data (Singh 1963; Johnson et al. 2005). A ZIP distribution is a mixture of a degenerated distribution at zero and a Poisson distribution, which is expressed as follows:

$$\Pr(Y = y; \pi, \lambda) = \pi I_{\{y=0\}} + (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots,$$
$$= \begin{cases} \pi + (1-\pi)e^{-\lambda}, & \text{if } y = 0, \\ (1-\pi)\frac{e^{-\lambda}\lambda^y}{y!}, & \text{if } y = 1, 2, \ldots \end{cases} \tag{1}$$

Here $\pi \in [0, 1]$ is a mixing weight for the accommodation of extra zeros. λ is the Poisson mean. $I_{\{\cdot\}}$ is the indicator function. It is noted that the ZIP distribution is reduced to a regular Poisson distribution when the mixing weight $\pi = 0$. The ZIP distribution can be thought of as a population that consists of two latent groups: the susceptible group consisting of those who are at risk of an event of interest, and may have the event several times during a specific time period, and the non-susceptible group consisting of those who are not at risk of the event of interest (Dietz and Böhning 1997).

To study covariate effects, Lambert (1992) proposed a parametric ZIP regression model to analyze an example of soldering defects on printed wiring boards. Some researchers also have successfully applied the ZIP regression models to several important clinical studies (e.g., Böhning et al. 1999; Yau and Lee 2001; Cheung 2002; Lu et al. 2004). In these ZIP regression models, the effects of covariates are usually modeled via a linear predictor function. However, it sometimes may not be completely appropriate to assume that the effect of a covariate is linear or some other parametric form. The nonparametric estimation methods have been used to relax the restrictive parametric assumption via modeling the possibly non-linear effect of the covariate by an unspecified smooth function. However, the phenomenon referred to as "curse of dimensionality" (Stone 1985) may arise if the dimension of covariates that have possibly non-linear effects is large. To address the issue, structured nonparametric regression methodology, such as additive models (Stone 1985; Hastie and Tibshirani 1990), has been extensively considered. A single-index model is an alternative way to deal with the curse of dimensionality and meanwhile retain enough flexibility on modeling. Therefore, we propose a ZIP regression single-index model that is described in detail in Sect. 2.

Some authors have widely applied single-index models to a variety of fields, e.g., clinical trials (Huang and Liu 2006; Sun et al. 2008), environmental studies (Yu and Ruppert 2002), and dose-response modeling (Härdle et al. 1993). Estimation methods for single-index models have been extensively discussed, including, e.g., spline estimation (Yu and Ruppert 2002; Huang and Liu 2006; Sun et al. 2008), local linear method (Carroll et al. 1997), average derivative method (Härdle and Stoker 1989; Horowitz and Härdle 1996), and kernel smoothing (Ichimura 1993; Härdle et al. 1993; Delecroix et al. 2003). In contrast to these estimation methods, we consider the spline-based sieve M-estimation method (Lu and Loomis 2013). After the spline basis functions are chosen, the spline coefficients are used to completely describe the approximated unknown smooth function. Therefore, the regression parameters and the spline coefficients can be estimated simultaneously by maximizing the spline likelihood function. Furthermore, the spline estimation can be used to directly estimate the asymptotic variances of the estimators of the regression parameters.

Our estimation method is different from the mentioned spline estimation method for single-index models, in which the number of knots are pre-specified. By allowing for the number of knots to increase at an appropriate rate, it can be shown that the spline estimator of the unknown smooth function achieves the optimal convergence rate. Furthermore, the semiparametric efficient score and the efficiency bound are derived. It is shown that the spline-based sieve semiparametric model can achieve the asymptotic efficiency for the estimators of the regression parameters.

In Sect. 2, we introduce the ZIP regression single-index model and the method of estimating the model parameters. Section 3 states asymptotic results of the spline estimators and likelihood ratio inference. In Sect. 4, we propose an approach to consistently estimate the variances of the estimators of the regression parameters. Simulations are given in Sect. 5. Proofs of the asymptotic results are sketched in the Appendix.

2 Model and method

2.1 ZIP regression single-index model

Let $\mathbf{x} = (x_0, \ldots, x_{p-1})^{\mathrm{T}}$, for $x_0 = 1$, be a $p \times 1$ covariate vector, and $\mathbf{z} = (z_1, \ldots, z_d)^{\mathrm{T}}$ a $d \times 1$ covariate vector. Let $\mathbf{w} = (\mathbf{x}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$. In this work, we are interested in smoothing the effects of $\mathbf{z}$ on the Poisson mean and consider the ZIP regression single-index model as follows:

$$\mathrm{logit}[\pi(\mathbf{w}; \boldsymbol{\alpha})] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}, \tag{2}$$

$$\log[\lambda(\mathbf{w}; \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \psi)] = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 + \psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2). \tag{3}$$

Here $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{p+d-1})^{\mathrm{T}}$. $\boldsymbol{\beta}_1 = (\beta_{10}, \ldots, \beta_{1(p-1)})^{\mathrm{T}}$. $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2d})^{\mathrm{T}}$. ψ is an unknown smooth function. In the ZIP regression single-index model, we assume that the covariates $\mathbf{w}$ have linear effects on the logit of the non-susceptible probability $\pi(\mathbf{w}; \boldsymbol{\alpha})$. The covariates $\mathbf{z}$ have non-linear effects on the logarithm of the Poisson mean, and the effects of $\mathbf{x}$ remain linear. To the best of our knowledge, the ZIP regression single-index model has not been proposed for modeling zero-inflated count data when the dimension of covariates that have possibly non-linear effects is large.

Because the regression function ψ can only be identified up to an additive constant, and the scale of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2$ in $\psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)$ can be determined arbitrarily, for purposes of identifiability, impose the restriction $E[\psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)] = 0$, and standardize $\boldsymbol{\beta}_2$ by $\|\boldsymbol{\beta}_2\|_2 = 1$ with the first nonzero element being positive, where $\|\cdot\|_2$ denotes L_2-norm. We handle the constraint $\|\boldsymbol{\beta}_2\|_2 = 1$ with the first nonzero element being positive by reparameterizing the single-index parameter vector $\boldsymbol{\beta}_2$. To do so, let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{d-1})^{\mathrm{T}}$ be a $(d-1)$-dimensional parameter vector, and define $\boldsymbol{\zeta} = \left((1 - \|\boldsymbol{\phi}\|_2^2)^{1/2}, \phi_1, \ldots, \phi_{d-1}\right)^{\mathrm{T}}$, where $\|\boldsymbol{\phi}\|_2^2 = \phi_1^2 + \cdots + \phi_{d-1}^2$. Assume the true parameter vector $\boldsymbol{\phi}_0$ satisfies $\|\boldsymbol{\phi}_0\|_2 < 1$. Thus, $\boldsymbol{\zeta}$ is infinitely differentiable in a neighborhood of $\boldsymbol{\phi}_0$.

Let $\{(y_i, \mathbf{w}_i^{\mathrm{T}})^{\mathrm{T}} : i = 1, \ldots, n\}$ be the observed data set. The kernel of the log-likelihood function for $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}}, \psi)^{\mathrm{T}}$ is

$$\begin{aligned}\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi) = \sum_{i=1}^{n} \Big\{ & I_{\{y_i=0\}} \log \Big\{ \pi(\mathbf{w}_i; \boldsymbol{\alpha}) + [1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] \mathrm{e}^{-\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi)} \Big\} \\ & + I_{\{y_i>0\}} \big\{ \log[1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] - \lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi) \\ & + y_i \log \big[\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi) \big] \big\} \Big\}. \end{aligned} \tag{4}$$

Let $\mathcal{T}_n = \{\xi_i\}_1^{m_n+2l}$ with $a = \xi_1 = \cdots = \xi_l < \xi_{l+1} < \cdots < \xi_{m_n+l} < \xi_{m_n+l+1} = \cdots = \xi_{m_n+2l} = b$ be a sequence of knots that divide the interval $[a, b]$ into $m_n + 1$ subintervals $I_i = [\xi_{l+i}, \xi_{l+1+i}]$, $i = 0, \ldots, m_n$. The spline of order $l \geq 1$ with the knot sequence $\mathcal{T}_n$ is a polynomial of degree $l-1$ within any subinterval $[\xi_{l+i}, \xi_{l+i+1}]$. A spline of order $l = 4$ 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), for any $s \in \mathcal{S}_n(\mathcal{T}_n, l)$, there exists a set of B-spline basis functions $\{b_j : 1 \leq j \leq q_n\}$ such that $s = \sum_{j=1}^{q_n} \gamma_j b_j$, where $q_n = m_n + l$ is the number of basis functions. To ensure $E[\psi(\cdot)] = 0$, we introduce empirically centered B-splines $\mathcal{S}_{0,n} = \big\{ s : s \in \mathcal{S}_n, \frac{1}{n} \sum_{i=1}^{n} s(\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\zeta}) = 0 \big\}$. The basis of $\mathcal{S}_{0,n}$ can be constructed by the set $\big\{ B_j = b_j - \frac{1}{n} \sum_{i=1}^{n} b_j(\mathbf{z}_i^{\mathrm{T}} \boldsymbol{\zeta}),\ j = 1, \ldots, q_n \big\}$.

If ψ is smooth enough, then ψ can be approximated by a B-spline function $\psi_n \in \mathcal{S}_{0,n}$; that is,

$$\psi(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta}) \approx \psi_n(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta}) = \sum_{j=1}^{q_n} \gamma_j B_j(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta}). \tag{5}$$

Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{q_n})^{\mathrm{T}}$. Note that $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ is one-dimension lower than $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. Replacing $\psi(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta})$ by $\psi_n(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta})$ in the kernel of the log-likelihood function (4), we can obtain the log-likelihood function for $\bar{\boldsymbol{\tau}} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ up to an additive constant not containing $\bar{\boldsymbol{\tau}}$ as follows:

$$\begin{aligned}\ell(\bar{\boldsymbol{\tau}}) &= \sum_{i=1}^{n} \ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) \\ &= \sum_{i=1}^{n} \Big\{ I_{\{y_i=0\}} \log \Big\{ \pi(\mathbf{w}_i; \boldsymbol{\alpha}) + [1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] \mathrm{e}^{-\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi_n)} \Big\} \\ &\quad + I_{\{y_i>0\}} \big\{ \log[1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] - \lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi_n) \\ &\quad + y_i \log \big[\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi_n) \big] \big\} \Big\}. \end{aligned} \tag{6}$$

Let $(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\phi}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ be the values that maximize the spline log-likelihood function (6). The spline estimate of $\psi(\mathbf{z}^{\mathrm{T}} \boldsymbol{\zeta})$ is defined as $\sum_{j=1}^{q_n} \hat{\gamma}_j B_j(\mathbf{z}^{\mathrm{T}} \hat{\boldsymbol{\zeta}})$. The advantage of this reparametrization is that we can estimate the regression parameters $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$ and the spline coefficients $\boldsymbol{\gamma}$ simultaneously and, hence, make the computation less demanding.

2.2 Algorithm

Let $\pi_i = \pi(\mathbf{w}_i; \boldsymbol{\alpha})$ and $\lambda_i = \lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi_n)$, $i = 1, \ldots, n$. Denote

$$\begin{aligned}
\dot{\ell}_\pi(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) &= \partial\ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i)/\partial\pi_i,\\
\dot{\ell}_\lambda(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) &= \partial\ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i)/\partial\lambda_i,\\
\ddot{\ell}_{\pi\pi}(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) &= \partial^2\ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i)/\partial\pi_i^2,\\
\ddot{\ell}_{\pi\lambda}(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) &= \partial^2\ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i)/\partial\pi_i\partial\lambda_i, \text{ and}\\
\ddot{\ell}_{\lambda\lambda}(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i) &= \partial^2\ell(\bar{\boldsymbol{\tau}}; y_i, \mathbf{w}_i)/\partial\lambda_i^2.
\end{aligned}$$

Let

$$\begin{aligned}
\dot{\mathbf{l}}_\pi(\bar{\boldsymbol{\tau}}) &= (\dot{\ell}_\pi(\bar{\boldsymbol{\tau}}; y_1, \mathbf{w}_1), \ldots, \dot{\ell}_\pi(\bar{\boldsymbol{\tau}}; y_n, \mathbf{w}_n))^{\mathrm{T}},\\
\dot{\mathbf{l}}_\lambda(\bar{\boldsymbol{\tau}}) &= (\dot{\ell}_\lambda(\bar{\boldsymbol{\tau}}; y_1, \mathbf{w}_1), \ldots, \dot{\ell}_\lambda(\bar{\boldsymbol{\tau}}; y_n, \mathbf{w}_n))^{\mathrm{T}},\\
\ddot{\mathbf{l}}_{\pi\pi}(\bar{\boldsymbol{\tau}}) &= \mathrm{diag}\{E[\ddot{\ell}_{\pi\pi}(\bar{\boldsymbol{\tau}}; y_1, \mathbf{w}_1)|\mathbf{w}_1], \ldots, E[\ddot{\ell}_{\pi\pi}(\bar{\boldsymbol{\tau}}; y_n, \mathbf{w}_n)|\mathbf{w}_n]\},\\
\ddot{\mathbf{l}}_{\pi\lambda}(\bar{\boldsymbol{\tau}}) &= \mathrm{diag}\{E[\ddot{\ell}_{\pi\lambda}(\bar{\boldsymbol{\tau}}; y_1, \mathbf{w}_1)|\mathbf{w}_1], \ldots, E[\ddot{\ell}_{\pi\lambda}(\bar{\boldsymbol{\tau}}; y_n, \mathbf{w}_n)|\mathbf{w}_n]\}, \text{ and}\\
\ddot{\mathbf{l}}_{\lambda\lambda}(\bar{\boldsymbol{\tau}}) &= \mathrm{diag}\{E[\ddot{\ell}_{\lambda\lambda}(\bar{\boldsymbol{\tau}}; y_1, \mathbf{w}_1)|\mathbf{w}_1], \ldots, E[\ddot{\ell}_{\lambda\lambda}(\bar{\boldsymbol{\tau}}; y_n, \mathbf{w}_n)|\mathbf{w}_n]\}.
\end{aligned}$$

Define $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{b}_{\mathbf{z}_i} = (B_1(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\zeta}), \ldots, B_{q_n}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\zeta}))^{\mathrm{T}}$, $\mathbf{B}^{\mathrm{T}} = (\mathbf{b}_{\mathbf{z}_1}, \ldots, \mathbf{b}_{\mathbf{z}_n})$, $\boldsymbol{\xi} = (\xi_{ir})_{n\times(d-1)}$, where $\xi_{ir} = \sum_{j=1}^{q_n} \gamma_j B_j'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\zeta}) \left[\frac{-\phi_r}{(1-\|\boldsymbol{\phi}\|_2^2)^{1/2}} z_{i1} + z_{i(r+1)}\right]$, $\boldsymbol{\pi} = \mathrm{diag}\{\pi_1, \ldots, \pi_n\}$, $\bar{\boldsymbol{\pi}} = \mathrm{diag}\{\bar{\pi}_1, \ldots, \bar{\pi}_n\}$, for $\bar{\pi}_i = \pi_i(1 - \pi_i)$, and $\boldsymbol{\lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Denote $\mathbf{D} = (\mathbf{X}, \boldsymbol{\xi}, \mathbf{B})$ and $\mathbf{W}^{\mathrm{T}} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$. Some derivative calculations yield the score vector

$$\nabla\ell(\bar{\boldsymbol{\tau}}) = \mathbf{H}^{\mathrm{T}} \begin{pmatrix} \dot{\mathbf{l}}_\pi(\bar{\boldsymbol{\tau}}) \\ \dot{\mathbf{l}}_\lambda(\bar{\boldsymbol{\tau}}) \end{pmatrix} \equiv \begin{bmatrix} \mathbf{W}^{\mathrm{T}}\bar{\boldsymbol{\pi}} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{\mathrm{T}}\boldsymbol{\lambda} \end{bmatrix} \begin{pmatrix} \dot{\mathbf{l}}_\pi(\bar{\boldsymbol{\tau}}) \\ \dot{\mathbf{l}}_\lambda(\bar{\boldsymbol{\tau}}) \end{pmatrix}$$

and the expected information matrix

$$\mathbf{E}(\bar{\boldsymbol{\tau}}) = \mathbf{H}^{\mathrm{T}} \begin{bmatrix} \ddot{\mathbf{l}}_{\pi\pi}(\bar{\boldsymbol{\tau}}) & \ddot{\mathbf{l}}_{\pi\lambda}(\bar{\boldsymbol{\tau}}) \\ \ddot{\mathbf{l}}_{\pi\lambda}(\bar{\boldsymbol{\tau}}) & \ddot{\mathbf{l}}_{\lambda\lambda}(\bar{\boldsymbol{\tau}}) \end{bmatrix} \mathbf{H}.$$

We apply the following modified Fisher scoring iterative procedure to calculate the spline estimates $(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \hat{\boldsymbol{\phi}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$.

Step 0. Choose $\beta_{2k}^{(0)} \sim N(0, 1)$, $k = 1, \ldots, d$. Standardize $\boldsymbol{\beta}_2^{(0)} = (\beta_{21}^{(0)}, \ldots, \beta_{2d}^{(0)})^{\mathrm{T}}$ with the first nonzero element being positive to get the initial estimate $\boldsymbol{\phi}^{(0)}$ for $\boldsymbol{\phi}$.

Step 1. Maximize the following log-likelihood function over $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$

$$\ell(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}}) = \sum_{i=1}^{n} \Big\{ I_{\{y_i=0\}} \log \Big\{ \pi(\mathbf{w}_i; \boldsymbol{\alpha}) + [1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] \mathrm{e}^{-\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}^{(0)}, \psi_n)} \Big\} + I_{\{y_i>0\}} \Big\{ \log[1 - \pi(\mathbf{w}_i; \boldsymbol{\alpha})] - \lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}^{(0)}, \psi_n) + y_i \log \Big[\lambda(\mathbf{w}_i; \boldsymbol{\beta}_1, \boldsymbol{\phi}^{(0)}, \psi_n) \Big] \Big\} \Big\}$$

to get $(\boldsymbol{\alpha}^{(0)\mathrm{T}}, \boldsymbol{\beta}_1^{(0)\mathrm{T}}, \boldsymbol{\gamma}^{(0)\mathrm{T}})^{\mathrm{T}}$.

Step 2. Given the initial values $(\boldsymbol{\alpha}^{(0)\mathrm{T}}, \boldsymbol{\beta}_1^{(0)\mathrm{T}}, \boldsymbol{\phi}^{(0)\mathrm{T}}, \boldsymbol{\gamma}^{(0)\mathrm{T}})^{\mathrm{T}}$, use the Fisher-scoring method $\bar{\boldsymbol{\tau}}^{(m)} = \bar{\boldsymbol{\tau}}^{(m-1)} - \mathbf{E}^{-1}(\bar{\boldsymbol{\tau}}^{(m-1)})\nabla\ell(\bar{\boldsymbol{\tau}}^{(m-1)})$ to update $\bar{\boldsymbol{\tau}}^{(m)\mathrm{T}} = (\boldsymbol{\alpha}^{(m)\mathrm{T}}, \boldsymbol{\beta}_1^{(m)\mathrm{T}}, \boldsymbol{\phi}^{(m)\mathrm{T}}, \boldsymbol{\gamma}^{(m)\mathrm{T}})^{\mathrm{T}}$ in the mth iteration. Repeat the iteration until the convergence criterion, e.g., $\|\bar{\boldsymbol{\tau}}^{(m)} - \bar{\boldsymbol{\tau}}^{(m-1)}\| < \varepsilon = 10^{-6}$, is met.

2.3 Comments on selection of knots

In this study, we employ cubic splines to approximate the unknown function ψ. Assume the true function ψ_0 has the rth continuous derivative, $r \geq 1$. According to the optimal rate of convergence for $\hat{\psi}$ given in Theorem 1, we select the number of inner knots, m_n, from a neighborhood of $n^{1/(1+2r)}$, such as $[0.5N_r, \min(4N_r, n^{1/2})]$, where $N_r = n^{1/(1+2r)}$. The optimal number of inner knots, m_n^*, is selected to minimize the Akaike's information criterion (AIC) value

$$\mathrm{AIC}(m_n) = -2\ell(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}}; m_n) + 2(m_n + l + 2d + 2p - 1).$$

After m_n^* is determined, we propose two ways to select the locations of knots. Let $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_{\min}$ and $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_{\max}$ be minimum and maximum values of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}$ respectively. One way is to equally divide $[\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_{\min}, \mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_{\max}]$ into $m_n^* + 1$ subintervals and choose the end points of subintervals as the locations of knots. The alternative one is a data-driven approach first proposed by Rosenbeg (1995) in which the $k/(m_n^* + 1)$ quantiles of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}$, $k = 1, \ldots, m_n^*$, are selected as interior knots. The similar method can be found in Lu and Loomis (2013).

3 Asymptotic results

Let $\mathcal{R}^{d-1}$, $\mathcal{R}^{p}$, and $\mathcal{R}^{p+d}$ be interiors of some compact sets in $\mathbf{R}^{d-1}$, $\mathbf{R}^{p}$, and $\mathbf{R}^{p+d}$ respectively. Denote $\boldsymbol{\theta} = (\boldsymbol{\phi}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\tau} = (\boldsymbol{\theta}^{\mathrm{T}}, \psi)^{\mathrm{T}}$. Let $\boldsymbol{\tau}_0 = (\boldsymbol{\theta}_0^{\mathrm{T}}, \psi_0)^{\mathrm{T}}$ be the true value of $\boldsymbol{\tau}$. Denote the regression parameter space by $\Theta = \mathcal{R}^{d-1} \times \mathcal{R}^{p+d} \times \mathcal{R}^{p}$, and let

$$\Psi = \{\psi : \text{the third derivative of} \, \psi \, \text{is Lipschitz on a compact subset} \, \Im \, \text{of} \, (0, \infty)\}$$

be the nonparametric space. Define L_2-norm $\|\cdot\|_2$ on $\Theta\times\Psi$ as follows: $\|\boldsymbol{\tau}_1-\boldsymbol{\tau}_2\|_2^2 = \|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2+\|\psi_1-\psi_2\|_2^2$, where $\boldsymbol{\tau}_i=(\boldsymbol{\theta}_i^{\mathrm{T}},\psi_i)^{\mathrm{T}}$, $i=1,2$.

For a single observation $(y,\mathbf{w}^{\mathrm{T}})^{\mathrm{T}}$, its kernel of log density of $\boldsymbol{\tau}$ is

$$\ell(\boldsymbol{\tau};y,\mathbf{w}) = I_{\{y=0\}}\log\left\{\pi(\mathbf{w};\boldsymbol{\alpha})+[1-\pi(\mathbf{w};\boldsymbol{\alpha})]\mathrm{e}^{-\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)}\right\} \\ +I_{\{y>0\}}\left\{\log[1-\pi(\mathbf{w};\boldsymbol{\alpha})]-\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)+y\log\left[\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)\right]\right\}.$$

Let ψ' denote the first derivative of ψ. The score functions for $(\boldsymbol{\phi}^{\mathrm{T}},\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ are

$$\dot{\ell}_{\boldsymbol{\phi}}(\boldsymbol{\tau};y,\mathbf{w}) = \frac{\partial\ell(\boldsymbol{\tau};y,\mathbf{w})}{\partial\boldsymbol{\phi}} = \dot{\ell}_{\lambda}(\boldsymbol{\tau};y,\mathbf{w})\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)\psi'(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta})\mathbf{D}(\boldsymbol{\phi})\mathbf{z},$$

$$\dot{\ell}_{\boldsymbol{\alpha}}(\boldsymbol{\tau};y,\mathbf{w}) = \frac{\partial\ell(\boldsymbol{\tau};y,\mathbf{w})}{\partial\boldsymbol{\alpha}} = \dot{\ell}_{\pi}(\boldsymbol{\tau};y,\mathbf{w})\bar{\pi}(\mathbf{w};\boldsymbol{\alpha})\mathbf{w}, \quad \text{and}$$

$$\dot{\ell}_{\boldsymbol{\beta}_1}(\boldsymbol{\tau};y,\mathbf{w}) = \frac{\partial\ell(\boldsymbol{\tau};y,\mathbf{w})}{\partial\boldsymbol{\beta}_1} = \dot{\ell}_{\lambda}(\boldsymbol{\tau};y,\mathbf{w})\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)\mathbf{x}.$$

Here $\mathbf{D}(\boldsymbol{\phi})=[-(1-\|\boldsymbol{\phi}\|_2^2)^{-1/2}\boldsymbol{\phi},\mathbf{I}_{d-1}]_{(d-1)\times d}$ for $\mathbf{I}_{d-1}$ being a $(d-1)\times(d-1)$ identity matrix. Consider a parametric smooth submodel $(\boldsymbol{\theta},\psi_t)$, where $\psi_t|_{t=0}=\psi$ and $d\psi_t/dt|_{t=0}=h$. Let $\mathcal{H}$ be the class of such h with bounded variation on $\Im$. The score operator for ψ is defined as

$$\dot{\ell}_{\psi}(\boldsymbol{\tau};y,\mathbf{w})[h] = \frac{d\ell(\boldsymbol{\theta},\psi_t;y,\mathbf{w})}{dt}|_{t=0} = \dot{\ell}_{\lambda}(\boldsymbol{\tau};y,\mathbf{w})\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)h.$$

Moreover, denote $\dot{\ell}_{\psi}(\boldsymbol{\tau};y,\mathbf{w})[\mathbf{h}]=(\dot{\ell}_{\psi}(\boldsymbol{\tau};y,\mathbf{w})[h_1],\ldots,\dot{\ell}_{\psi}(\boldsymbol{\tau};y,\mathbf{w})[h_{2p+2d-1}])^{\mathrm{T}}$, where $\mathbf{h}=(h_1,\ldots,h_{2p+2d-1})^{\mathrm{T}}\in\mathcal{H}^{2p+2d-1}$.

The efficient score function for $\boldsymbol{\theta}$ at the true parameter vector $\boldsymbol{\tau}_0$ is

$$\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau}_0;y,\mathbf{w}) = \dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0;y,\mathbf{w})-\dot{\ell}_{\psi}(\boldsymbol{\tau}_0;y,\mathbf{w})[\mathbf{h}^*],$$

where $\mathbf{h}^*$ minimizes $\rho(\mathbf{h})=\|\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0;y,\mathbf{w})-\dot{\ell}_{\psi}(\boldsymbol{\tau}_0;y,\mathbf{w})[\mathbf{h}]\|_2^2$ for $\mathbf{h}\in\mathcal{H}^{2p+2d-1}$. More specifically, $E[\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0;y,\mathbf{w})-\dot{\ell}_{\psi}(\boldsymbol{\tau}_0;y,\mathbf{w})[\mathbf{h}^*]]^{\mathrm{T}}\dot{\ell}_{\psi}(\boldsymbol{\tau}_0;y,\mathbf{w})[\mathbf{h}]=\mathbf{0}$ for any $\mathbf{h}=(\mathbf{h}_1^{\mathrm{T}},\mathbf{h}_2^{\mathrm{T}},\mathbf{h}_3^{\mathrm{T}})^{\mathrm{T}}\in\mathcal{H}^{2p+2d-1}$. Denote $\xi_0=\dot{\ell}_{\pi}(\boldsymbol{\tau}_0;y,\mathbf{w})\bar{\pi}(\mathbf{w};\boldsymbol{\alpha}_0)$, $u_0=\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_0$, and $\eta_0=\dot{\ell}_{\lambda}(\boldsymbol{\tau}_0;y,\mathbf{w})\lambda(\mathbf{w};\boldsymbol{\beta}_{10},\boldsymbol{\phi}_0,\psi_0)$. Some calculations then yield

$$\mathbf{h}^* = \begin{pmatrix}\mathbf{h}_1^*\\ \mathbf{h}_2^*\\ \mathbf{h}_3^*\end{pmatrix} = \frac{1}{E_{(y,\mathbf{w})|u_0}[\eta_0^2|u_0]}\begin{pmatrix}\psi_0'(u_0)E_{(y,\mathbf{w})|u_0}[\eta_0^2\mathbf{D}(\boldsymbol{\phi}_0)\mathbf{z}|u_0]\\ E_{(y,\mathbf{w})|u_0}[\xi_0\eta_0\mathbf{w}|u_0]\\ E_{(y,\mathbf{w})|u_0}[\eta_0^2\mathbf{x}|u_0]\end{pmatrix}.$$

Therefore, the efficient score function for $\boldsymbol{\theta}$ at $\boldsymbol{\tau}_0$ is

$$\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau}_0;y,\mathbf{w}) = \begin{pmatrix}\eta_0[\psi_0'(u_0)\mathbf{D}(\boldsymbol{\phi}_0)\mathbf{z}-\mathbf{h}_1^*]\\ \xi_0\mathbf{w}-\eta_0\mathbf{h}_2^*\\ \eta_0(\mathbf{x}-\mathbf{h}_3^*)\end{pmatrix}.$$

The efficient information $\mathbf{I}(\boldsymbol{\theta}_0)$ takes the form $E\left[\ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0; y, \mathbf{w})\right]^{\otimes 2}$, where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for any vector $\mathbf{a}$.

The following assumptions are needed to derive the asymptotic properties of the spline estimator $\hat{\boldsymbol{\tau}} = (\hat{\boldsymbol{\theta}}^{\mathrm{T}}, \hat{\psi})^{\mathrm{T}}$.

(C1) The maximum spacing of the knots is assumed to be $O(n^{-\nu})$ for $0 < \nu < 1/2$. Moreover, the ratio of maximum and minimum spacings of knots is uniformly bounded.
(C2) The true regression parameters $(\boldsymbol{\theta}_0^{\mathrm{T}}, \psi_0)^{\mathrm{T}}$ is in the interior of $\Theta \times \Psi$.
(C3) (a) The support of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}$ is an interval within $\Im$ for $\boldsymbol{\zeta}$ in a neighborhood of $\boldsymbol{\zeta}_0$; (b) the density of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}$ is positive and bounded away from 0 on $\Im$.
(C4) The fourth moment of $\mathbf{x}$ is finite.
(C5) Write $\varepsilon = y - E(y|\mathbf{w})$. Given $\mathbf{w}$, ε is sub-Gussian.
(C6) For any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}_0$, $\boldsymbol{\zeta} \neq \boldsymbol{\zeta}_0$, and $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{10}$, $\Pr(\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha} \neq \mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}_0) > 0$, $\Pr(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta} \neq \mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_0) > 0$, and $\Pr(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 \neq \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{10}) > 0$.

Remark 1 Condition C1 is a mild assumption on knots and similar to those required by Stone (1986). Condition C2 is the standard assumption in semiparametric estimation. Conditions C3 and C4 are needed for entropy calculations in the proofs of Theorems 1–4. Condition C5 is essential to calculate the bracketing integral with respect to Bernstein norm (van der Vaart and Wellner 1996). Condition C6 is required to establish the identifiability of the model.

Theorem 1 (Uniform convergence and rate of convergence) *Let* $q_n = O(n^{\nu})$ *for* $1/(2r+2) < \nu < 1/(2r)$. *Suppose conditions* C1–C6 *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 nonparametric regression.*

Let $\boldsymbol{\vartheta} = (\boldsymbol{\zeta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}_0 = (\boldsymbol{\zeta}_0^{\mathrm{T}}, \boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}})^{\mathrm{T}}$. The Jacobian matrix of $F : \boldsymbol{\vartheta} \to \boldsymbol{\theta}$ is

$$\mathbf{J}(\boldsymbol{\phi}) = \begin{bmatrix} -(1-\|\boldsymbol{\phi}\|_2^2)^{-1/2}\boldsymbol{\phi}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{I}_{d-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{2p+d} \end{bmatrix}.$$

Note that $\mathbf{J}(\boldsymbol{\phi})$ is a matrix of dimension $(2p+2d) \times (2p+2d-1)$.

Theorem 2 (Asymptotic normality) *Suppose conditions* C1–C6 *hold and* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is nonsingular. Then*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{1/2}\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\sum_{i=1}^{n}\ell^*_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0; y_i, \mathbf{w}_i) + o_p(1) \to N(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0))$$

in distribution as $n \to \infty$. *Furthermore, the vector of constrained spline estimators* $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\zeta}}^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}^{\mathrm{T}}, \hat{\boldsymbol{\beta}}_1^{\mathrm{T}})^{\mathrm{T}}$ *with* $\|\hat{\boldsymbol{\zeta}}\|_2 = 1$ *is asymptotically normally distributed*

Table 1 ($\rho = 0.0$) Summary of parameter estimation for simulation study

	$n = 300$				$n = 500$			
	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE
α_{00}	0.3925	−0.0075	0.7243	0.5247	0.4168	0.0168	0.5582	0.3119
α_{10}	0.5319	0.0319	0.2121	0.0460	0.5101	0.0101	0.1643	0.0271
α_{20}	−0.8510	−0.0510	0.1828	0.0360	−0.8261	−0.0261	0.1320	0.0181
α_{30}	−0.7165	−0.0165	0.3552	0.1264	−0.7123	−0.0123	0.2548	0.0651
α_{40}	0.5364	0.0364	0.3373	0.1151	0.5051	0.0051	0.2492	0.0621
α_{50}	−0.8277	−0.0277	0.3518	0.1245	−0.8209	−0.0209	0.2648	0.0705
β_{110}	1.0013	0.0013	0.0179	0.0003	1.0004	0.0004	0.0124	0.0002
β_{120}	0.4001	0.0001	0.0051	0.0000	0.4000	−0.0000	0.0035	0.0000
β_{210}	0.5768	−0.0005	0.0141	0.0002	0.5776	0.0003	0.0102	0.0001
β_{220}	0.5779	0.0006	0.0128	0.0002	0.5774	0.0001	0.0099	0.0001
β_{230}	0.5768	−0.0005	0.0135	0.0002	0.5768	−0.0006	0.0102	0.0001
π_{10}	0.6050	0.0051	0.0662	0.0044	0.5996	−0.0003	0.0511	0.0026
π_{20}	0.8087	0.0034	0.0799	0.0064	0.8010	−0.0043	0.0643	0.0042
λ_{10}	12.3285	0.0303	0.5182	0.2694	12.2999	0.0017	0.3795	0.1440
λ_{20}	48.0766	−0.0259	2.4502	6.0043	48.0216	−0.0810	1.9838	3.9421

$\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \alpha_{30}, \alpha_{40}, \alpha_{50})^{\mathrm{T}} = (0.4, 0.5, -0.8, -0.7, 0.5, -0.8)^{\mathrm{T}}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{\beta}_{10} = (\beta_{110}, \beta_{120})^{\mathrm{T}} = (1, 0.4)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (\beta_{210}, \beta_{220}, \beta_{230})^{\mathrm{T}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$. $\pi_{10} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.5998884$ and $\lambda_{10} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{10}) = 12.298234$ for $\psi_{10} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 5\sin(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) - 15\sqrt{3}/8\left[3\cos(2/\sqrt{3}) - 3\cos(4/\sqrt{3}) - 1 + \cos(6/\sqrt{3})\right] = 0.7694557$ at $\mathbf{w} = (1.0, 1.5, 0.6, 0.7, 1.25, 0.5)^{\mathrm{T}}$. $\pi_{20} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.8053384$ and $\lambda_{20} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{20}) = 48.102577$ for $\psi_{20} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.5533357$ at $\mathbf{w} = (1.0, 3.2, 0.3, 0.9, 1.70, 0.7)^{\mathrm{T}}$. Sample mean (mean), bias, standard deviation (SD), and mean squared error (MSE) of the partial linear zero-inflated Poisson single-index model by B-splines, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

$$n^{1/2}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \to N\left(\mathbf{0}, \mathbf{J}(\boldsymbol{\phi}_0)\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{J}^{\mathrm{T}}(\boldsymbol{\phi}_0)\right).$$

Remark 2 Theorem 1 shows $\hat{\psi}$ is a uniformly consistent estimator of ψ_0. Although the overall rate of convergence for spline estimators $\hat{\boldsymbol{\tau}}$ is $n^{r/(1+2r)} < n^{1/2}$, the rate of convergence for $\hat{\boldsymbol{\theta}}$ is still $n^{1/2}$. Theorem 2 shows the vector of spline estimators $\hat{\boldsymbol{\theta}}$ achieves the information bound, and is therefore efficient in the semiparametric sense.

Let $\hat{\psi}_{\boldsymbol{\theta}}$ be the spline estimator of ψ for any $\boldsymbol{\theta}$ in the neighborhood of $\hat{\boldsymbol{\theta}}$. The profile log-likelihood for $\boldsymbol{\theta}$ is defined as $pl(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, \hat{\psi}_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \hat{\psi}_{\boldsymbol{\theta}}; y_i, \mathbf{w}_i)$. The likelihood ratio test statistic for testing $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is given by $lrt(\boldsymbol{\theta}_0) = 2pl(\hat{\boldsymbol{\theta}}) - 2pl(\boldsymbol{\theta}_0)$.

Theorem 3 (Likelihood ratio inference) *Suppose conditions* C1–C6 *hold. Then under* $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, $lrt(\boldsymbol{\theta}_0) \to \chi^2_{2p+2d-1}$ *in distribution as* $n \to \infty$.

Table 2 ($\rho = 0.3$) Summary of parameter estimation for simulation study

	$n = 300$				$n = 500$			
	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE
α_{00}	0.4466	0.0466	0.7151	0.5135	0.3967	−0.0033	0.5518	0.3044
α_{10}	0.5142	0.0142	0.2162	0.0469	0.5228	0.0228	0.1599	0.0261
α_{20}	−0.8520	−0.0520	0.1840	0.0366	−0.8243	−0.0243	0.1378	0.0196
α_{30}	−0.7362	−0.0362	0.3679	0.1366	−0.7125	−0.0125	0.2671	0.0715
α_{40}	0.5187	0.0187	0.4578	0.2100	0.5071	0.0071	0.3270	0.1070
α_{50}	−0.8329	−0.0329	0.5039	0.2550	−0.8218	−0.0218	0.3913	0.1536
β_{110}	1.0010	0.0010	0.0176	0.0003	0.9996	−0.0004	0.0133	0.0002
β_{120}	0.3999	−0.0001	0.0056	0.0000	0.4002	0.0002	0.0037	0.0000
β_{210}	0.5772	−0.0001	0.0188	0.0004	0.5773	−0.0001	0.0136	0.0002
β_{220}	0.5778	0.0005	0.0205	0.0004	0.5767	−0.0006	0.0148	0.0002
β_{230}	0.5759	−0.0014	0.0221	0.0005	0.5775	0.0002	0.0157	0.0002
π_{10}	0.6012	0.0013	0.0847	0.0072	0.5995	−0.0004	0.0639	0.0041
π_{20}	0.7960	−0.0093	0.0955	0.0092	0.8034	−0.0019	0.0698	0.0049
λ_{10}	12.3171	0.0188	0.5187	0.2694	12.2894	−0.0088	0.3869	0.1498
λ_{20}	48.2204	0.1179	2.8445	8.1050	48.0024	−0.1002	2.1548	4.6532

$\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \alpha_{30}, \alpha_{40}, \alpha_{50})^{\mathrm{T}} = (0.4, 0.5, -0.8, -0.7, 0.5, -0.8)^{\mathrm{T}}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{\beta}_{10} = (\beta_{110}, \beta_{120})^{\mathrm{T}} = (1, 0.4)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (\beta_{210}, \beta_{220}, \beta_{230})^{\mathrm{T}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$. $\pi_{10} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.5998884$ and $\lambda_{10} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{10}) = 12.298234$ for $\psi_{10} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 5\sin(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) - 15\sqrt{3}/8\left[3\cos(2/\sqrt{3}) - 3\cos(4/\sqrt{3}) - 1 + \cos(6/\sqrt{3})\right] = 0.7694557$ at $\mathbf{w} = (1.0, 1.5, 0.6, 0.7, 1.25, 0.5)^{\mathrm{T}}$. $\pi_{20} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.8053384$ and $\lambda_{20} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{20}) = 48.102577$ for $\psi_{20} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.5533357$ at $\mathbf{w} = (1.0, 3.2, 0.3, 0.9, 1.70, 0.7)^{\mathrm{T}}$. Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the partial linear zero-inflated Poisson single-index model by B-splines, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

4 Variance estimation

Although the efficient information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$ has an explicit expression, it is not trivial to estimate $\mathbf{I}(\boldsymbol{\theta}_0)$ directly. There are at least two methods for estimation of asymptotic variances for semiparametric models. One method is to use the second derivative of the profile likelihood to estimate $\mathbf{I}(\boldsymbol{\theta}_0)$. Because the profile likelihood may not be differentiated directly, instead the discretized version of the observed profile information proposed by Nielsen et al. (1992) is often used as an estimator in practice. Murphy and van der Vaart (1999) showed that $\mathbf{I}(\boldsymbol{\theta}_0)$ can be consistently estimated by a discretized version of the negative second derivative of the profile likelihood function. In this study, we adopt the alternative estimation method proposed by Huang et al. (2008). This approach is based on the orthogonal projection of the efficient score function for $\boldsymbol{\theta}$ onto the tangent space for ψ.

Estimate $\mathbf{I}(\boldsymbol{\theta}_0)$ by its empirical version

$$\mathbb{P}_n[\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w}) - \dot{\ell}_{\psi}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[\mathbf{h}^*]]^{\otimes 2},$$

Table 3 ($\rho = 0.55$) Summary of parameter estimation for simulation study

	$n = 300$				$n = 500$			
	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE
α_{00}	0.4357	0.0357	0.7298	0.5339	0.4015	0.0015	0.5921	0.3506
α_{10}	0.5391	0.0391	0.2219	0.0508	0.5177	0.0177	0.1642	0.0273
α_{20}	−0.8628	−0.0628	0.1951	0.0420	−0.8369	−0.0369	0.1437	0.0220
α_{30}	−0.7158	−0.0158	0.4075	0.1663	−0.7205	−0.0205	0.3095	0.0962
α_{40}	0.5493	0.0493	0.5716	0.3292	0.5072	0.0072	0.4490	0.2016
α_{50}	−0.8895	−0.0895	0.6674	0.4535	−0.8059	−0.0059	0.5368	0.2882
β_{110}	0.9997	−0.0003	0.0194	0.0004	1.0008	0.0008	0.0137	0.0002
β_{120}	0.3999	−0.0001	0.0059	0.0000	0.4003	0.0003	0.0037	0.0000
β_{210}	0.5778	0.0004	0.0245	0.0006	0.5771	−0.0002	0.0177	0.0003
β_{220}	0.5757	−0.0016	0.0264	0.0007	0.5764	−0.0009	0.0201	0.0004
β_{230}	0.5767	−0.0007	0.0292	0.0009	0.5775	0.0001	0.0214	0.0005
π_{10}	0.6102	0.0103	0.1001	0.0101	0.5966	−0.0033	0.0802	0.0064
π_{20}	0.8066	0.0013	0.1046	0.0109	0.7979	−0.0075	0.0838	0.0071
λ_{10}	12.2913	−0.0070	0.5563	0.3096	12.2451	−0.0531	0.3967	0.1602
λ_{20}	48.1524	0.0498	3.1996	10.2498	48.1175	0.0149	2.4536	6.0205

$\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \alpha_{30}, \alpha_{40}, \alpha_{50})^{\mathrm{T}} = (0.4, 0.5, -0.8, -0.7, 0.5, -0.8)^{\mathrm{T}}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{\beta}_{10} = (\beta_{110}, \beta_{120})^{\mathrm{T}} = (1, 0.4)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (\beta_{210}, \beta_{220}, \beta_{230})^{\mathrm{T}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$. $\pi_{10} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.5998884$ and $\lambda_{10} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{10}) = 12.298234$ for $\psi_{10} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 5\sin(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) - 15\sqrt{3}/8\left[3\cos(2/\sqrt{3}) - 3\cos(4/\sqrt{3}) - 1 + \cos(6/\sqrt{3})\right] = 0.7694557$ at $\mathbf{w} = (1.0, 1.5, 0.6, 0.7, 1.25, 0.5)^{\mathrm{T}}$. $\pi_{20} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.8053384$ and $\lambda_{20} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{20}) = 48.102577$ for $\psi_{20} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.5533357$ at $\mathbf{w} = (1.0, 3.2, 0.3, 0.9, 1.70, 0.7)^{\mathrm{T}}$. Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the partial linear zero-inflated Poisson single-index model by B-splines, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

where $\mathbb{P}_n$ is the empirical measure, and $\mathbf{h}^* = (h_1^*, \ldots, h_{2p+2d-1}^*)^{\mathrm{T}}$ minimizes $\rho(\mathbf{h}) = \|\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{\tau}_0; y, \mathbf{w}) - \dot{\ell}_{\psi}(\boldsymbol{\tau}_0; y, \mathbf{w})[\mathbf{h}]\|_2^2$ over $\mathcal{H}^{2p+2d-1}$. Approximate h_s^* by a B-spline function $h_{n,s}^* = \sum_{j=1}^{q_n} \gamma_{j,s} B_j$, $s = 1, \ldots, 2p + 2d - 1$. The coefficient vectors $\boldsymbol{\gamma}_s = (\gamma_{1,s}, \ldots, \gamma_{q_n,s})^{\mathrm{T}}$ can be estimated by minimizing

$$\mathbb{P}_n \sum_{s=1}^{2p+2d-1} \left[\dot{\ell}_{\boldsymbol{\theta},s}(\hat{\boldsymbol{\tau}}; y, \mathbf{w}) - \sum_{j=1}^{q_n} \gamma_{j,s} \dot{\ell}_{\psi}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[B_j] \right]^2,$$

where $\dot{\ell}_{\boldsymbol{\theta},s}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})$ is the sth element of $\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})$. Therefore, the variance estimation is essentially a least-squares estimation problem. The estimator of h_s^* is defined as $\hat{h}_s^* = \sum_{j=1}^{q_n} \hat{\gamma}_{j,s} B_j$. Let $\hat{\mathbf{h}}^* = (\hat{h}_1^*, \ldots, \hat{h}_{2p+2d-1}^*)^{\mathrm{T}}$ and $\mathcal{B}_n = (B_1, \ldots, B_{q_n})^{\mathrm{T}}$. By standard least-squares calculation, it follows that

$$\mathbb{P}_n[\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w}) - \dot{\ell}_{\psi}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[\hat{\mathbf{h}}^*]]^{\otimes 2} = \hat{A}_{\boldsymbol{\theta}\boldsymbol{\theta}} - \hat{A}_{\boldsymbol{\theta}\psi} \hat{A}_{\psi\psi}^{-1} \hat{A}_{\psi\boldsymbol{\theta}},$$

Table 4 ($\rho = 0.8$) Summary of parameter estimation for simulation study

	$n = 300$				$n = 500$			
	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE
α_{00}	0.4316	0.0316	0.7786	0.6072	0.4368	0.0368	0.5783	0.3357
α_{10}	0.5304	0.0304	0.2187	0.0488	0.5072	0.0072	0.1703	0.0291
α_{20}	−0.8635	−0.0635	0.2004	0.0442	−0.8272	−0.0272	0.1465	0.0222
α_{30}	−0.7521	−0.0521	0.6080	0.3724	−0.7201	−0.0201	0.4335	0.1883
α_{40}	0.5204	0.0204	0.8671	0.7522	0.5021	0.0021	0.6393	0.4087
α_{50}	−0.8149	−0.0149	0.9947	0.9896	−0.8213	−0.0213	0.7708	0.5947
β_{110}	1.0002	0.0002	0.0198	0.0004	1.0002	0.0002	0.0144	0.0002
β_{120}	0.3998	−0.0002	0.0061	0.0000	0.4002	0.0002	0.0041	0.0000
β_{210}	0.5761	−0.0012	0.0346	0.0012	0.5759	−0.0014	0.0255	0.0007
β_{220}	0.5762	−0.0011	0.0388	0.0015	0.5773	−0.0001	0.0308	0.0010
β_{230}	0.5758	−0.0016	0.0425	0.0018	0.5765	−0.0008	0.0333	0.0011
π_{10}	0.5963	−0.0036	0.1402	0.0197	0.5968	−0.0031	0.1109	0.0123
π_{20}	0.7841	−0.0212	0.1406	0.0202	0.7873	−0.0180	0.1111	0.0127
λ_{10}	12.2993	0.0011	0.5994	0.3593	12.2737	−0.0245	0.4802	0.2312
λ_{20}	48.0720	−0.0306	3.8491	14.8167	47.9548	−0.1478	2.9917	8.9723

$\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \alpha_{30}, \alpha_{40}, \alpha_{50})^{\mathrm{T}} = (0.4, 0.5, -0.8, -0.7, 0.5, -0.8)^{\mathrm{T}}$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{\beta}_{10} = (\beta_{110}, \beta_{120})^{\mathrm{T}} = (1, 0.4)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (\beta_{210}, \beta_{220}, \beta_{230})^{\mathrm{T}} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$. $\pi_{10} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.5998884$ and $\lambda_{10} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{10}) = 12.298234$ for $\psi_{10} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 5\sin(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) - 15\sqrt{3}/8\left[3\cos(2/\sqrt{3}) - 3\cos(4/\sqrt{3}) - 1 + \cos(6/\sqrt{3})\right] = 0.7694557$ at $\mathbf{w} = (1.0, 1.5, 0.6, 0.7, 1.25, 0.5)^{\mathrm{T}}$. $\pi_{20} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.8053384$ and $\lambda_{20} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{20}) = 48.102577$ for $\psi_{20} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.5533357$ at $\mathbf{w} = (1.0, 3.2, 0.3, 0.9, 1.70, 0.7)^{\mathrm{T}}$. Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the partial linear zero-inflated Poisson single-index model by B-splines, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

where $\hat{A}_{\boldsymbol{\theta\theta}} = \mathbb{P}_n[\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})]^{\otimes 2}$, $\hat{A}_{\boldsymbol{\theta}\psi} = \mathbb{P}_n[\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})\dot{\ell}_{\psi}^{\mathrm{T}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[\mathcal{B}_n]]$, $\hat{A}_{\psi\boldsymbol{\theta}} = \hat{A}_{\boldsymbol{\theta}\psi}^{\mathrm{T}}$, and $\hat{A}_{\psi\psi} = \mathbb{P}_n[\dot{\ell}_{\psi}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[\mathcal{B}_n]]^{\otimes 2}$. Denote $\mathcal{O}_n = \hat{A}_{\boldsymbol{\theta\theta}} - \hat{A}_{\boldsymbol{\theta}\psi}\hat{A}_{\psi\psi}^{-1}\hat{A}_{\psi\boldsymbol{\theta}}$. Huang et al. (2008) applied the observed information $\mathcal{O}_n$ to estimate the efficient information. We instead employ the Fisher information to estimate $\mathbf{I}(\boldsymbol{\theta}_0)$. Denote $\hat{E}_{\boldsymbol{\theta\theta}} = E[\hat{A}_{\boldsymbol{\theta\theta}}|\mathbf{w}]$, $\hat{E}_{\boldsymbol{\theta}\psi} = E[\hat{A}_{\boldsymbol{\theta}\psi}|\mathbf{w}]$, $\hat{E}_{\psi\boldsymbol{\theta}} = \hat{E}_{\boldsymbol{\theta}\psi}^{\mathrm{T}}$, and $\hat{E}_{\psi\psi} = E[\hat{A}_{\psi\psi}|\mathbf{w}]$. Theorem 4 shows that the conditional expected information $\hat{\mathcal{E}}_n = \hat{E}_{\boldsymbol{\theta\theta}} - \hat{E}_{\boldsymbol{\theta}\psi}\hat{E}_{\psi\psi}^{-1}\hat{E}_{\psi\boldsymbol{\theta}}$ is a consistent estimator of $\mathbf{I}(\boldsymbol{\theta}_0)$. More specifically, with the notations defined in Sect. 2, $\mathbf{I}(\boldsymbol{\theta}_0)$ can be consistently estimated by

$$\hat{\mathcal{E}}_n = \frac{1}{n}\left\{\begin{bmatrix} \hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\boldsymbol{\xi}} & \hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\mathbf{W} & \hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\mathbf{X} \\ \mathbf{W}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\hat{\boldsymbol{\xi}} & \mathbf{W}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\pi}\mathbf{W} & \mathbf{W}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\mathbf{X} \\ \mathbf{X}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\boldsymbol{\xi}} & \mathbf{X}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\mathbf{W} & \mathbf{X}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\mathbf{X} \end{bmatrix} - \begin{bmatrix} \hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\mathbf{B}} \\ \mathbf{W}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\hat{\mathbf{B}} \\ \mathbf{X}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\mathbf{B}} \end{bmatrix} (\hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\mathbf{B}})^{-1} \begin{bmatrix} \hat{\boldsymbol{\xi}}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\mathbf{B}} \\ \mathbf{W}^{\mathrm{T}}\hat{\mathbf{L}}_{\pi\lambda}\hat{\mathbf{B}} \\ \mathbf{X}^{\mathrm{T}}\hat{\mathbf{L}}_{\lambda\lambda}\hat{\mathbf{B}} \end{bmatrix}^{\mathrm{T}}\right\}. \tag{7}$$

Table 5 ($\rho = 0.0$) Results of variance study

	$n = 300$			$n = 500$		
	Mean.se	SD.se	CP (%)	Mean.se	SD.se	CP (%)
α_{00}	0.7059	0.0560	95.1	0.5388	0.0324	95.8
α_{10}	0.2075	0.0212	94.3	0.1570	0.0126	94.5
α_{20}	0.1782	0.0284	95.1	0.1345	0.0155	96.8
α_{30}	0.3331	0.0275	93.8	0.2538	0.0157	94.9
α_{40}	0.3233	0.0240	94.1	0.2457	0.0127	95.9
α_{50}	0.3362	0.0282	93.9	0.2562	0.0159	94.8
β_{110}	0.0174	0.0029	93.7	0.0125	0.0018	95.7
β_{120}	0.0050	0.0014	95.0	0.0034	0.0009	94.9
β_{210}	0.0131	0.0019	93.3	0.0095	0.0012	92.7
β_{220}	0.0131	0.0019	95.2	0.0095	0.0012	93.7
β_{230}	0.0131	0.0019	94.3	0.0095	0.0012	93.3

Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95 % confidence interval (CP) of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

Table 6 ($\rho = 0.3$) Results of variance study

	$n = 300$			$n = 500$		
	Mean.se	SD.se	CP(%)	Mean.se	SD.se	CP (%)
α_{00}	0.7148	0.0558	94.6	0.5412	0.0320	95.5
α_{10}	0.2082	0.0232	95.1	0.1579	0.0123	95.7
α_{20}	0.1808	0.0287	95.4	0.1362	0.0168	95.0
α_{30}	0.3450	0.0305	93.7	0.2615	0.0162	95.1
α_{40}	0.4271	0.0316	93.9	0.3237	0.0173	95.8
α_{50}	0.5023	0.0386	95.7	0.3817	0.0212	95.2
β_{110}	0.0178	0.0030	94.7	0.0130	0.0019	94.6
β_{120}	0.0052	0.0015	93.8	0.0036	0.0009	96.5
β_{210}	0.0178	0.0026	94.3	0.0132	0.0016	95.0
β_{220}	0.0189	0.0026	92.4	0.0139	0.0017	93.4
β_{230}	0.0203	0.0028	93.8	0.0150	0.0017	93.4

Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95 % confidence interval (CP) of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

Here $\mathbf{L}_{\pi\pi} = \ddot{\mathbf{l}}_{\pi\pi}(\bar{\boldsymbol{\tau}})\bar{\boldsymbol{\pi}}^2$. $\mathbf{L}_{\pi\lambda} = \ddot{\mathbf{l}}_{\pi\lambda}(\bar{\boldsymbol{\tau}})\bar{\boldsymbol{\pi}}\boldsymbol{\lambda}$. $\mathbf{L}_{\lambda\lambda} = \ddot{\mathbf{l}}_{\lambda\lambda}(\bar{\boldsymbol{\tau}})\boldsymbol{\lambda}^2$. $\hat{\boldsymbol{\xi}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{L}}_{\pi\pi}$, $\hat{\mathbf{L}}_{\pi\lambda}$, and $\hat{\mathbf{L}}_{\lambda\lambda}$ represent $\boldsymbol{\xi}$, $\mathbf{B}$, $\mathbf{L}_{\pi\pi}$, $\mathbf{L}_{\pi\lambda}$, and $\mathbf{L}_{\lambda\lambda}$ evaluated at $\bar{\boldsymbol{\tau}} = \hat{\bar{\boldsymbol{\tau}}}$ respectively.

Theorem 4 (Variance estimation) *Under the same conditions assumed in Theorem 2, $\hat{\mathcal{E}}_n$ is asymptotically consistent to $\mathbf{I}(\boldsymbol{\theta}_0)$. Furthermore, $\mathbf{J}(\hat{\boldsymbol{\phi}})\hat{\mathcal{E}}_n^{-1}\mathbf{J}^{\mathrm{T}}(\hat{\boldsymbol{\phi}})$ is asymptotically consistent to $\mathbf{J}(\boldsymbol{\phi}_0)\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\mathbf{J}^{\mathrm{T}}(\boldsymbol{\phi}_0)$.*

Table 7 ($\rho = 0.55$) Results of variance study

	$n = 300$			$n = 500$		
	Mean.se	SD.se	CP (%)	Mean.se	SD.se	CP (%)
α_{00}	0.7276	0.0576	95.6	0.5538	0.0351	94.4
α_{10}	0.2128	0.0225	94.8	0.1606	0.0131	94.9
α_{20}	0.1854	0.0317	95.5	0.1398	0.0176	95.4
α_{30}	0.3990	0.0347	95.0	0.3030	0.0194	94.5
α_{40}	0.5638	0.0431	94.8	0.4279	0.0250	93.9
α_{50}	0.6700	0.0508	95.6	0.5089	0.0301	93.6
β_{110}	0.0189	0.0032	94.1	0.0134	0.0020	94.6
β_{120}	0.0054	0.0016	93.7	0.0037	0.0009	95.2
β_{210}	0.0228	0.0033	93.5	0.0167	0.0020	93.7
β_{220}	0.0256	0.0036	94.2	0.0187	0.0022	93.2
β_{230}	0.0280	0.0038	94.5	0.0203	0.0022	94.5

Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95 % confidence interval (CP) of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

Table 8 ($\rho = 0.8$) Results of variance study

	$n = 300$			$n = 500$		
	Mean.se	SD.se	CP	Mean.se	SD.se	CP
α_{00}	0.7551	0.0663	95.7	0.5714	0.0362	94.5
α_{10}	0.2192	0.0243	95.7	0.1651	0.0140	95.1
α_{20}	0.1897	0.0342	95.3	0.1419	0.0182	95.2
α_{30}	0.5860	0.0508	94.6	0.4429	0.0274	95.8
α_{40}	0.8437	0.0673	95.7	0.6403	0.0365	94.3
α_{50}	0.9747	0.0755	95.4	0.7387	0.0415	94.3
β_{110}	0.0199	0.0034	94.7	0.0144	0.0021	96.5
β_{120}	0.0057	0.0017	96.2	0.0039	0.0010	95.5
β_{210}	0.0337	0.0046	95.1	0.0247	0.0028	93.6
β_{220}	0.0387	0.0056	95.1	0.0284	0.0034	93.0
β_{230}	0.0418	0.0056	95.2	0.0304	0.0034	92.4

Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95 % confidence interval (CP) of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_0^{\mathrm{T}})^{\mathrm{T}}$, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

5 Simulation study

Results from Monte Carlo experiments are now presented to evaluate the finite-sample performance of the proposed method. We conducted 1000 replications for each configuration of the experiments. The covariate vector is assumed to take the form $\mathbf{w} = (\mathbf{x}^{\mathrm{T}}, \mathbf{z}^{\mathrm{T}})^{\mathrm{T}}$. Here $\mathbf{x} = (x_0, x_1, x_2)^{\mathrm{T}}$ for $x_0 = 1$. $\mathbf{z} = (z_1, z_2, z_3)^{\mathrm{T}}$. The data of x_1 were generated from the normal distribution with mean 1 and standard deviation 1. The data of x_2 were generated from the exponential distribution with rate 0.5

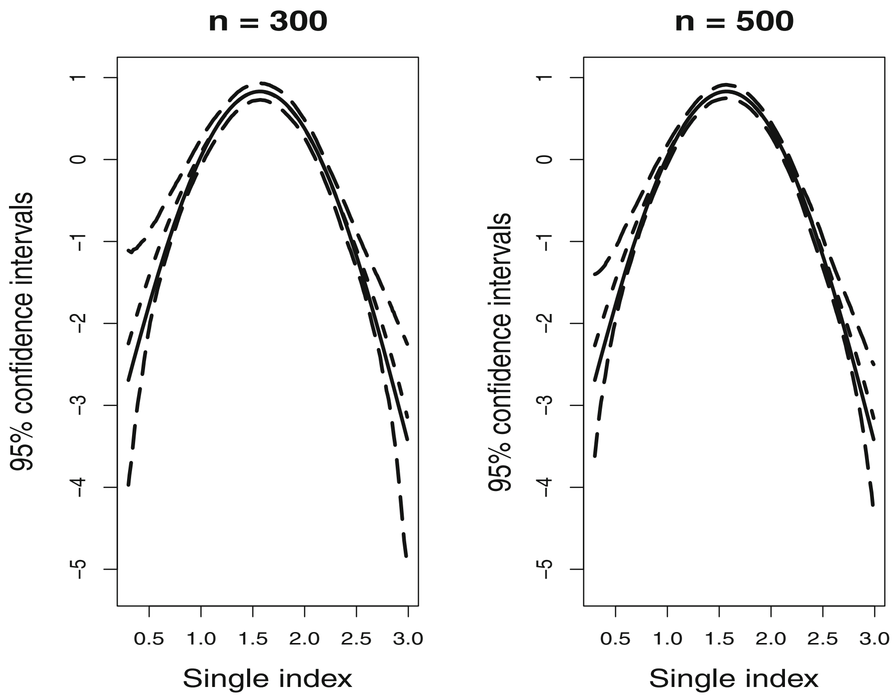

Fig. 1 ($\rho = 0.0$) Curve estimates and corresponding 95 % confidence intervals for $\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20})$. The *solid curves* are the true mean functions. The *dashed curves* are the average *B*-spline fits and the *long dashed curves* are the corresponding 2.5 and 97.5 % quantiles, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

(i.e., mean 2). We considered four scenarios of correlation between z_i and z_j, $i \neq j$, $i, j = 1, 2, 3$, where the correlation coefficient $\rho = \mathrm{corr}(z_i, z_j) = 0.0, 0.3, 0.55$, and 0.80 respectively when the data of z_1, z_2, and z_3 were generated from the uniform $[0, 2]$ distribution. The data of count outcome variable Y were generated from the ZIP regression single-index model that consists of the following two sub-regression models: the logistic regression model for the probability π that an individual is not at risk of the event

$$\mathrm{logit}\,[\pi(\mathbf{w}; \boldsymbol{\alpha}_0)] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}_0$$

and the Poisson regression single-index model with mean λ for the event count when an individual is at risk of the event

$$\log[\lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_0)] = \beta_{110}x_1 + \beta_{120}x_2 + \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}).$$

Here $\boldsymbol{\alpha}_0 = (\alpha_{00}, \alpha_{10}, \alpha_{20}, \alpha_{30}, \alpha_{40}, \alpha_{50})^{\mathrm{T}} = (0.4, 0.5, -0.8, -0.7, 0.5, -0.8)^{\mathrm{T}}$. $\boldsymbol{\beta}_{10} = (\beta_{110}, \beta_{120})^{\mathrm{T}} = (1, 0.4)^{\mathrm{T}}$. $\boldsymbol{\beta}_{20} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^{\mathrm{T}}$. $\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 5\sin(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) - 15\sqrt{3}/8\left[3\cos(2/\sqrt{3}) - 3\cos(4/\sqrt{3}) - 1 + \cos(6/\sqrt{3})\right]$.

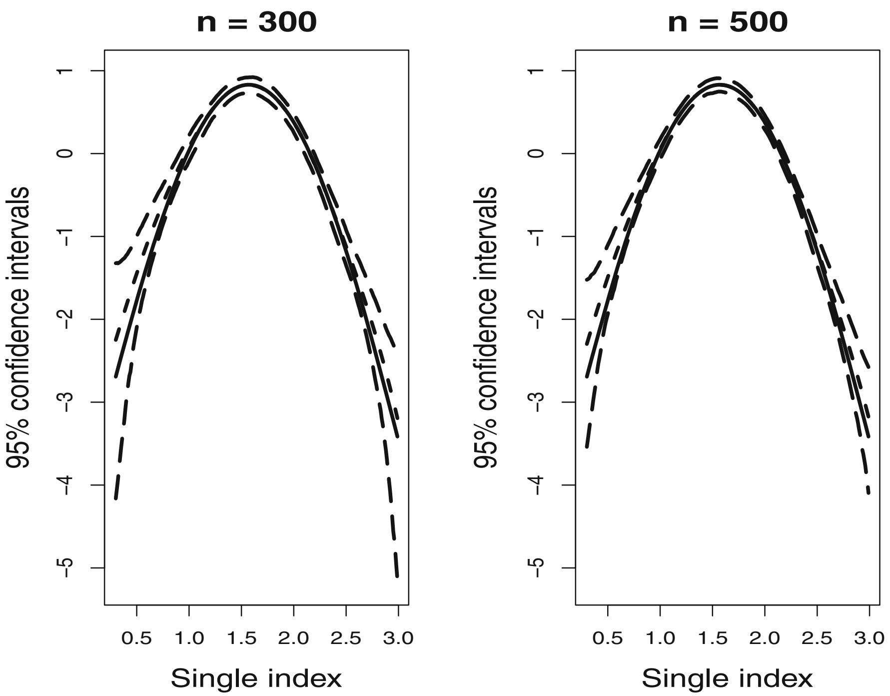

Fig. 2 ($\rho = 0.30$) Curve estimates and corresponding 95 % confidence intervals for $\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20})$. The *solid curves* are the true mean functions. The *dashed curves* are the average B-spline fits and the *long dashed curves* are the corresponding 2.5 and 97.5 % quantiles, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

The unknown function $\psi_0(\cdot)$ was approximated by cubic B-splines. We selected the number of interior knots m_n based on the AIC criteria that was discussed in Sect. 2.3. When m_n is determined, we considered equally spaced and quantile methods to choose the knot locations. Because the experiments show that the results were not sensitive to the selection of knot locations, the simulation results are only presented by the quantile method.

Tables 1, 2, 3, and 4 present the summary statistics for the estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, $\pi_{10} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.5998884$ and $\lambda_{10} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_{10}) = 12.298234$ for $\psi_{10} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.7694557$ at $\mathbf{w} = (1.0, 1.5, 0.6, 0.7, 1.25, 0.5)^{\mathrm{T}}$, and $\pi_{20} = \pi(\mathbf{w}, \boldsymbol{\alpha}_0) = 0.8053384$ and $\lambda_{20} = \lambda(\mathbf{w}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \psi_0) = 48.102577$ for $\psi_{20} = \psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20}) = 0.5533357$ at $\mathbf{w} = (1.0, 3.2, 0.3, 0.9, 1.70, 0.7)^{\mathrm{T}}$ when $\rho = 0.0, 0.3, 0.55$, and 0.8 respectively from the ZIP regression single-index model by B-spline, including sample mean (mean), bias, standard deviation (SD), and mean squared error (MSE). Because the ZIP regression model is misspecified, the estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, π_{10}, λ_{10}, π_{20}, and λ_{20} for the ZIP regression model were far from the true values. On the other hand, it was observed that the B-spline estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, π_{10}, λ_{10}, π_{20}, and λ_{20} are much more accurate than those for the ZIP regression model. The summary statistics for the estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$, π_{10}, λ_{10}, π_{20}, and λ_{20} are not shown when $\rho = 0.0, 0.3, 0.55$, and 0.8 respectively from

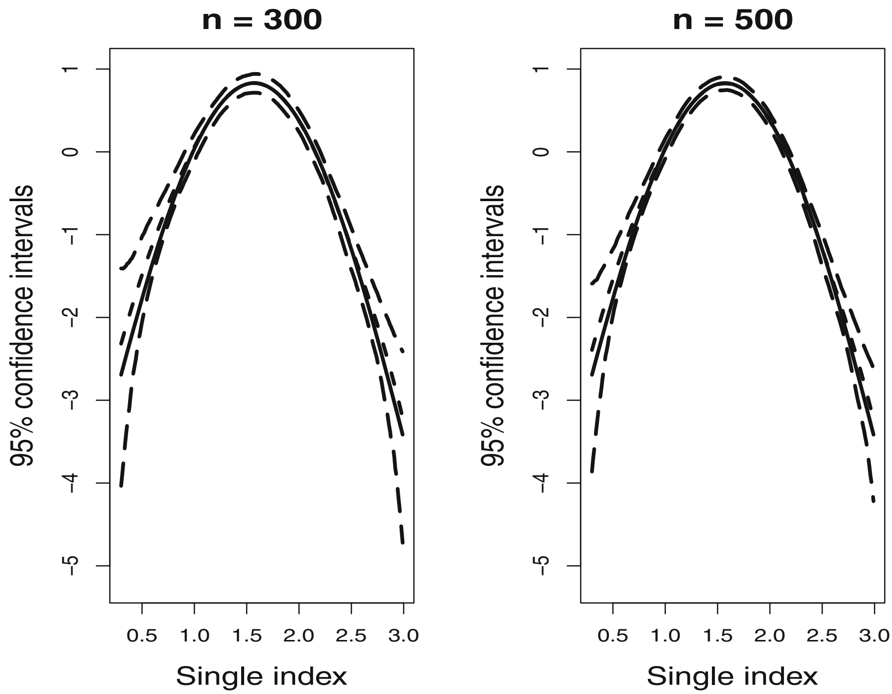

Fig. 3 ($\rho = 0.55$) Curve estimates and corresponding 95 % confidence intervals for $\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20})$. The *solid curves* are the true mean functions. The *dashed curves* are the average B-spline fits and the *long dashed curves* are the corresponding 2.5 and 97.5 % quantiles, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

the ZIP regression model. The results indicate that the biases of spline estimates are very small. These standard deviations of the estimates decrease at a rate of $n^{-1/2}$ as n increases when $\rho = 0$; however, their standard deviations tend to increase when ρ is larger.

The performance of the proposed standard error method for the B-spline estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ was also evaluated based on the conditional expected information given in (7). Tables 5, 6, 7, and 8 show the standard errors of the B-spline estimates of $(\boldsymbol{\alpha}_0^{\mathrm{T}}, \boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$ when $\rho = 0.0$, 0.3, 0.55, and 0.8 respectively. Overall, the proposed standard error estimation method was found to work reasonably well. The empirical coverage probabilities were not statistically significantly different from the nominal coverage probability 95 % except for β_{210} ($n = 300$) and α_{20} and β_{210} ($n = 500$) when $\rho = 0.0$, β_{220} ($n = 300$) and β_{120}, β_{220}, and β_{230} ($n = 500$) when $\rho = 0.3$, β_{210} ($n = 300$) and α_{50} and β_{220} ($n = 500$) when $\rho = 0.55$, and β_{110}, β_{210}, β_{220}, and β_{230} ($n = 500$) when $\rho = 0.8$ and tended to be closer to this nominal coverage probability when the sample size was increased. In addition, the averages of estimated standard errors were all close to the corresponding Monte Carlo standard deviations of the estimators.

Figures 1, 2, 3 and 4 depict the pointwise mean estimate of ψ_0 and the corresponding 2.5 and 97.5 % quantiles obtained based on 1000 Monte Carlo samples when $\rho =$

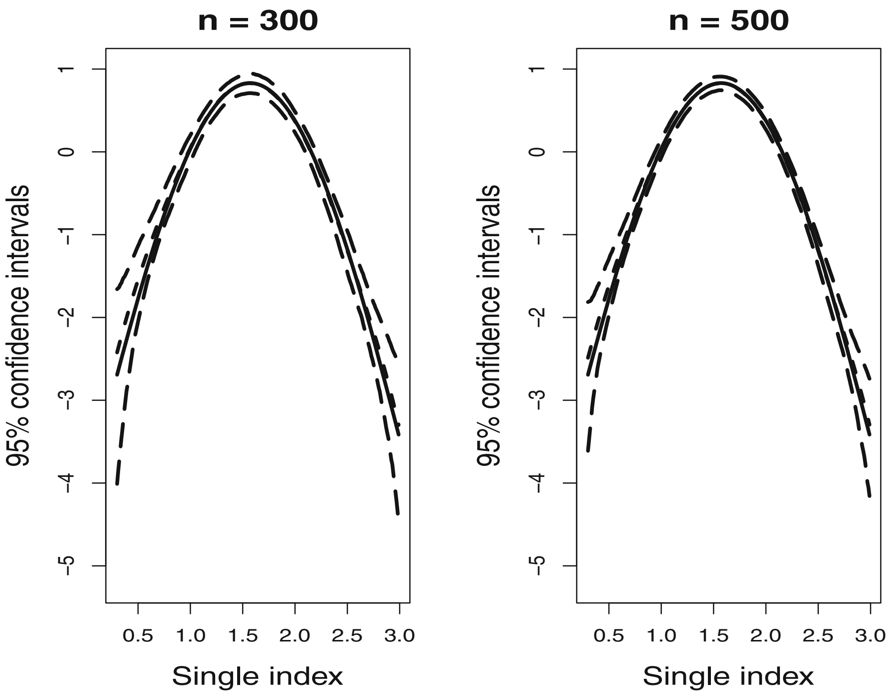

Fig. 4 ($\rho = 0.80$) Curve estimates and corresponding 95 % confidence intervals for $\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_{20})$. The *solid curves* are the true mean functions. The *dashed curves* are the average B-spline fits and the *long dashed curves* are the corresponding 2.5 and 97.5 % quantiles, based on 1000 Monte Carlo samples with sample size 300 or 500, respectively

0.0, 0.3, 0.55, and 0.8 respectively. The fitted curves followed closely the true curve ψ_0, which indicates that the bias is little. The lower and upper limits of confidence intervals were also reasonably close to the true function except on boundaries, which reveals that the variation in the estimates was small. The variability was decreased when sample size was increased. In summary, it can be seen from the Monte Carlo study that the proposed zero-inflated Poisson regression single-index model is a practical model for zero-inflated count data when the dimension of covariates that have possibly non-linear effects is large.

A Appendix

A.1 Notations and lemmas

Let $P_{\boldsymbol{\tau}}$ be the distribution of $(y, \mathbf{w}^{\mathrm{T}})^{\mathrm{T}}$ under the parameter vector $\boldsymbol{\tau}$ and $p_{\boldsymbol{\tau}}$ the corresponding density. Denote $P_0 \equiv P_{\boldsymbol{\tau}_0}$ and $p_0 \equiv p_{\boldsymbol{\tau}_0}$. For a measurable function f, define Pf as the expectation of f under P. For any class of measurable functions $\mathcal{F}$, the bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$ is defined as the minimum number of brackets $[f_i^L, f_i^R], i = 1, \ldots, m$, such that, for $f \in \mathcal{F}$, there exists $1 \leq i \leq m$ such that $f_i^L \leq f \leq f_i^R$ and $\|f_i^R - f_i^L\|_2 \leq \varepsilon$. Define the bracketing integral

$J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \left[1 + N_{[]}(\varepsilon, \mathcal{F}, L_2(P))\right]^{1/2} d\varepsilon$. Denote $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$ and $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$. In the following, C represents a positive constant that may vary from place to place.

Lemma 1 *For any* $\delta > 0$, *define* $\mathcal{L} = \{\ell(\boldsymbol{\tau}; y, \mathbf{w}) : \psi \in \mathcal{S}_{0,n}, \boldsymbol{\theta} \in \mathcal{R}^{2p+2d-1}, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \leq \delta\}$. *Then, for any* $0 < \varepsilon \leq \delta$, $\log N_{[]}(\varepsilon, \mathcal{L}, \|\cdot\|_{P,B}) \leq Cq_n \log(\delta/\varepsilon)$, *and, hence,* $J_{[]}(\delta, \mathcal{L}, \|\cdot\|_{P,B}) \leq Cq_n^{1/2}\delta$, *where* $\|\cdot\|_{P,B}$ *is the Bernstein norm defined as* $\|f\|_{P,B}^2 = 2P\left[\exp(|f|) - |f| - 1\right]$ *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–C6 *hold, then there exists* $C > 0$ *such that*

$$P[\ell(\boldsymbol{\tau}_0; y, \mathbf{w}) - \ell(\boldsymbol{\tau}; y, \mathbf{w})] \geq C\|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2^2,$$

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

Lemma 3 (Consistency) *If conditions* C1–C6 *hold, then* $\|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0\|_2 = o_p(1)$.

Remark 3 Lemma 1 and the similar entropy calculations are used to derive the consistency of $\hat{\boldsymbol{\tau}}$ and to prove Theorems 1–4. Lemma 2 is a key result to derive the consistency and rate of convergence of $\hat{\boldsymbol{\tau}}$. Lemma 3 shows $\hat{\boldsymbol{\tau}}$ is asymptotically consistent to $\boldsymbol{\tau}_0$.

A.2 Proof of Lemma 1

Proof According to the bracketing calculation in Shen and Wong (1994), for any $\delta > 0$ and $0 < \varepsilon \leq \delta$, the logarithm of bracketing number of $\mathcal{S}_{0,n}$, computed with $L_2(P)$, is bounded by $q_n \log(\delta/\varepsilon)$ up to a constant. It is known that the neighborhoods $\mathbf{A}(\delta) = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|_2 \leq \delta\}$, $\mathbf{Z}(\delta) = \{\boldsymbol{\zeta} : \|\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\|_2 \leq \delta\}$, and $\mathbf{B}(\delta) = \{\boldsymbol{\beta}_1 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\|_2 \leq \delta\}$ can be covered by $O((\delta/\varepsilon)^{p+d})$, $O((\delta/\varepsilon)^d)$, and $O((\delta/\varepsilon)^p)$ balls with radius ε, respectively. In view of Theorem 9.23 of Kosorok (2008), the bracketing numbers for $\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|_2 \leq \delta\}$, $\{\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta} : \|\boldsymbol{\zeta} - \boldsymbol{\zeta}_0\|_2 \leq \delta\}$, and $\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 : \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\|_2 \leq \delta\}$ are bounded by $O((\delta/\varepsilon)^{p+d})$, $O((\delta/\varepsilon)^d)$, and $O((\delta/\varepsilon)^p)$, respectively. It follows that, for sufficiently large n,

$$\log N_{[]}\left(\varepsilon, \{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 + \psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}) : \psi \in \mathcal{S}_{0,n}, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \leq \delta\}, L_2(P)\right) \leq Cq_n \log(\delta/\varepsilon).$$

and, hence,

$$\log N_{[]}\left(\varepsilon, \{\lambda(\mathbf{w}; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi) : \psi \in \mathcal{S}_{0,n}, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \leq \delta\}, L_2(P)\right) \leq Cq_n \log(\delta/\varepsilon)$$

because the function $x \mapsto \exp(x)$ is Lipschitz and monotonic. By inequality $2[\exp(|x|) - 1 - |x|] \leq x^2 \exp(|x|)$,

$$\log N_{[]}\left(\varepsilon, \{\lambda(\mathbf{w}; \boldsymbol{\beta}_1, \boldsymbol{\phi}, \psi) : \psi \in \mathcal{S}_{0,n}, \|\boldsymbol{\tau} - \boldsymbol{\tau}_0\|_2 \leq \delta\}, \|\cdot\|_{P,B}\right) \leq Cq_n \log(\delta/\varepsilon).$$

Similarly, we can show that

$$\log N_{[]}\left(\varepsilon, \{\pi(\mathbf{w}; \boldsymbol{\alpha}) : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|_2 \leq \delta\}, \|\cdot\|_{P,B}\right) \leq C \log(\delta/\varepsilon).$$

The transformation $\left(\pi(\mathbf{w};\boldsymbol{\alpha}),\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi)\right)\mapsto\ell(\pi(\mathbf{w};\boldsymbol{\alpha}),\lambda(\mathbf{w};\boldsymbol{\beta}_1,\boldsymbol{\phi},\psi);\boldsymbol{\tau})$ is essentially Lipschitz, so it follows that $\log N_{[]}\left(\varepsilon,\mathcal{L},\|\cdot\|_{P,B}\right)\le Cq_n\log(\delta/\varepsilon)$, and, hence, the bracketing integral is bounded by $q_n^{1/2}\delta$, up to a constant. □

A.3 Proof of Lemma 2

Proof Let $\mathbb{M}_n(\boldsymbol{\tau})=\mathbb{P}_n\ell(\boldsymbol{\tau};y,\mathbf{w})$ and $\mathbb{M}(\boldsymbol{\tau})=P\ell(\boldsymbol{\tau};y,\mathbf{w})$. For any $\boldsymbol{\tau}$ in a neighborhood of $\boldsymbol{\tau}_0$, a Taylor's expansion yields

$$\mathbb{M}(\boldsymbol{\tau}_0)-\mathbb{M}(\boldsymbol{\tau})\ge P[(\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}_0)^2]+P\{[\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{10}+\psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta})-\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_0)]^2\},$$

up to a constant. Let $g_1(\mathbf{x})=\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10})$ and $g_2(\mathbf{z})=\psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta})-\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_0)$. According to the law of total expectation and Cauchy-Schwarz inequality, $\{E[g_1(\mathbf{x})g_2(\mathbf{z})]\}^2\le E_{\mathbf{z}}[g_2^2(\mathbf{z})]E_{\mathbf{z}}[\{E_{\mathbf{x}|\mathbf{z}}[g_1(\mathbf{x})|\mathbf{z}]\}^2]$. By the orthogonality of a conditional expectation, there exists $0<\xi<1$ such that $E_{\mathbf{z}}[\{E_{\mathbf{x}|\mathbf{z}}[g_1(\mathbf{x})|\mathbf{z}]\}^2]=\xi E_{\mathbf{x}}[g_1^2(\mathbf{x})]$. Hence, $E[g_1^2(\mathbf{x})g_2^2(\mathbf{z})]\le\xi E[g_1^2(\mathbf{x})]E[g_2^2(\mathbf{z})]$. In view of Lemma 25.86 of van der Vaart (2000),

$$\mathbb{M}(\boldsymbol{\tau}_0)-\mathbb{M}(\boldsymbol{\tau})\ge\|\psi(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta})-\psi_0(\mathbf{z}^{\mathrm{T}}\boldsymbol{\zeta}_0)\|_2^2+\|\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}-\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}_0\|_2^2+\|\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_{10}\|_2^2,$$

up to a constant. By Lemma 1 of Stone (1985) and conditions C3 and C4, it follows that $\mathbb{M}(\boldsymbol{\tau}_0)-\mathbb{M}(\boldsymbol{\tau})\ge C\|\boldsymbol{\tau}-\boldsymbol{\tau}_0\|_2^2$. □

A.4 Proof of Lemma 3

Proof We verify the conditions of Theorem 5.7 in van der Vaart (2000) to prove the consistency of $\hat{\boldsymbol{\tau}}$. According to Lemma 1, $\mathcal{L}$ is a Donsker class, and is therefore a Glivenko-Cantelli class. Thus, $\sup_{\boldsymbol{\tau}}|(\mathbb{P}_n-P)\ell(\boldsymbol{\tau};y,\mathbf{w})|=o_p(1)$ for $\boldsymbol{\tau}$ in a neighborhood of $\boldsymbol{\tau}_0$. The first condition of the theorem holds. It follows from Lemma 2 that $\sup_{\|\boldsymbol{\tau}-\boldsymbol{\tau}_0\|_2\ge\varepsilon}\mathbb{M}(\boldsymbol{\tau})\le\mathbb{M}(\boldsymbol{\tau}_0)-C\varepsilon^2<\mathbb{M}(\boldsymbol{\tau}_0)$. Hence, the second condition of the theorem also holds.

According to Jackson's theorem for polynomials (de Boor 2001), there exists a spline of order $l\ge 2$ $\psi_{0,n}\in\mathcal{S}_{0,n}$ such that $\|\psi_{0,n}-\psi_0\|_\infty=O(n^{-r\nu})$ for $1/(2r+2)<\nu<1/(2r)$. Let $\boldsymbol{\tau}_{0,n}=(\boldsymbol{\theta}_0,\psi_{0,n})$. By definition of $\hat{\boldsymbol{\tau}}$,

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

where $I_{n1}=(\mathbb{P}_n-P)[\ell(\boldsymbol{\tau}_{0,n};y,\mathbf{w})-\ell(\boldsymbol{\tau}_0;y,\mathbf{w})]$ and $I_{n2}=P[\ell(\boldsymbol{\tau}_{0,n};y,\mathbf{w})-\ell(\boldsymbol{\tau}_0;y,\mathbf{w})]$. As shown in the proof of Lemma 1, $\mathcal{S}_{0,n}$ is a Donsker class. Because $\ell(\boldsymbol{\theta}_0,\psi;y,\mathbf{w})$ is essentially Lipschitz with respect to ψ, the preservation theorem of Donsker class yields the class of functions $\ell(\boldsymbol{\theta}_0,\psi;y,\mathbf{w})-\ell(\boldsymbol{\theta}_0,\psi_0;y,\mathbf{w})$, for $\psi\in\mathcal{S}_{0,n}$ and $\|\psi-\psi_0\|_2\le\delta$, is a Donsker class. Moreover, by the mean value theorem, $P[\ell(\boldsymbol{\tau}_{0,n};y,\mathbf{w})-\ell(\boldsymbol{\tau}_0;y,\mathbf{w})]^2\le C\|\psi_{0,n}-\psi_0\|_\infty^2\to 0$ as $n\to\infty$. In view of Lemma 19.24 of van der Vaart (2000), $I_{n1}=o_p(n^{-1/2})$. Observe that

$I_{n2} \geq -C\|\psi_{0,n} - \psi_0\|_\infty^2 = -O(n^{-2r\nu})$. It follows that $\mathbb{M}_n(\hat{\boldsymbol{\tau}}) - \mathbb{M}_n(\boldsymbol{\tau}_0) > -o_p(1)$. Therefore, Theorem 5.7 of van der Vaart (2000) applies and yields the of consistency of $\hat{\boldsymbol{\tau}}$. □

A.5 Proof of Theorem 1

Proof We apply Theorem 3.4.1 of van der Vaart and Wellner (1996) to prove the rate of convergence. Let $\boldsymbol{\tau} \in \Theta_n = \{(\boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \psi) : \boldsymbol{\phi} \in \mathcal{R}^{d-1}, \boldsymbol{\alpha} \in \mathcal{R}^{p+d}, \boldsymbol{\beta}_1 \in \mathcal{R}^p, \psi \in \mathcal{S}_{0,n}\}$. Choose $d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_{0,n})$ and $M_n(\boldsymbol{\tau})$ defined in the theorem to be $\|\boldsymbol{\tau} - \boldsymbol{\tau}_{0,n}\|_2$ and $\mathbb{M}(\boldsymbol{\tau})$, respectively. By definition of $\hat{\boldsymbol{\tau}}$, $\mathbb{M}_n(\hat{\boldsymbol{\tau}}) \geq \mathbb{M}_n(\boldsymbol{\tau}_{0,n})$. In the proof of Lemma 3 for consistency, we have already shown that $\mathbb{M}(\boldsymbol{\tau}) - \mathbb{M}(\boldsymbol{\tau}_0) \leq -Cd_n^2(\boldsymbol{\tau}, \boldsymbol{\tau}_0)$. Because $\mathbb{M}(\boldsymbol{\tau}_0) - \mathbb{M}(\boldsymbol{\tau}) \leq Cd_n(\boldsymbol{\tau}_{0,n}, \boldsymbol{\tau}_0) \leq C\|\psi_{0,n} - \psi_0\|_\infty = O(n^{-r\nu})$, for any $\boldsymbol{\tau} \in \Theta_n$ such that $\delta/2 \leq d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_{0,n}) \leq \delta$, we have $d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_0) \geq d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_{0,n}) - d_n(\boldsymbol{\tau}_{0,n}, \boldsymbol{\tau}_0) > C\delta$ for sufficiently large n. It follows that

$$\begin{aligned}\mathbb{M}(\boldsymbol{\tau}) - \mathbb{M}(\boldsymbol{\tau}_{0,n}) &= \mathbb{M}(\boldsymbol{\tau}) - \mathbb{M}(\boldsymbol{\tau}_0) + \mathbb{M}(\boldsymbol{\tau}_0) - \mathbb{M}(\boldsymbol{\tau}_{0,n}) \\ &\leq -C\delta^2 + O(n^{-2r\nu}) = -C\delta^2\end{aligned}$$

for sufficiently large n.

For any $\delta > 0$, in view of Lemma 1,

$$J_{[]}\{\delta, \{\ell(\boldsymbol{\tau}; y, \mathbf{w}) - \ell(\boldsymbol{\tau}_{0,n}; y, \mathbf{w}) : \boldsymbol{\tau} \in \Theta_n, \delta/2 \leq d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_{0,n}) \leq \delta\}, \|\cdot\|_{P,B}\} \leq Cq_n^{1/2}\delta.$$

Moreover, for $\boldsymbol{\tau} \in \Theta_n$ and $\delta/2 \leq d_n(\boldsymbol{\tau}, \boldsymbol{\tau}_{0,n}) \leq \delta$, by inequality $2[\exp(|x|) - |x| - 1] \leq x^2\exp(|x|)$ and conditions C3–C5, $\|\ell(\boldsymbol{\tau}; y, \mathbf{w}) - \ell(\boldsymbol{\tau}_{0,n}; y, \mathbf{w})\|_{P,B}^2 \leq C\delta^2$. Lemma 3.4.3 of van der Vaart and Wellner (1996) yields

$$E\left[\sup_{\delta/2 \leq \|\boldsymbol{\tau} - \boldsymbol{\tau}_{0,n}\|_2 \leq \delta, \boldsymbol{\tau} \in \Theta_n} n^{1/2}|(\mathbb{M}_n - \mathbb{M})(\boldsymbol{\tau}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{\tau}_{0,n})|\right] \leq C\phi_n(\delta)$$

with $\phi_n(\delta) = q_n^{1/2}\delta + n^{-1/2}q_n$. Obviously, $\phi_n(\delta)/\delta$ is decreasing in δ. It can be readily shown that $r_n^2\phi_n(1/r_n) \leq n^{1/2}$ with $r_n = n^{\min(r\nu, (1-\nu)/2)}$. Theorem 3.4.1 of van der Vaart and Wellner (1996) is applied to yield $r_n d_n(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_{0,n}) = O_p(1)$. Because $d_n(\boldsymbol{\tau}_{0,n}, \boldsymbol{\tau}_0) = O(n^{-r\nu})$, it follows that $r_n d_n(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_0) \leq r_n d_n(\hat{\boldsymbol{\tau}}, \boldsymbol{\tau}_{0,n}) + r_n d_n(\boldsymbol{\tau}_{0,n}, \boldsymbol{\tau}_0) = O_p(1) + r_n O(n^{-r\nu}) = O_p(1)$. This completes the proof of the rate of convergence. □

A.6 Proof of Theorem 2

Proof Theorem 2 follows from the same arguments as those in the proof of Theorem 1(b) of Lu and Loomis (2013) and entropy calculations similar to those in Lemma 1. □

A.7 Proof of Theorem 3

Proof We apply Theorem 3.1 of Murphy and van der Vaart (1997) to prove the asymptotic distribution of the spline likelihood ratio test statistic. Let $\mathbf{t} = \left(\mathbf{t}_1^{\mathrm{T}}, \mathbf{t}_2^{\mathrm{T}}, \mathbf{t}_3^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\mathbf{t}_1 \in \mathbb{R}^{d-1}$, $\mathbf{t}_2 \in \mathbb{R}^{p+d}$, and $\mathbf{t}_3 \in \mathbb{R}^p$. Define an approximately least favorable submodel

$$\Psi_{\mathbf{t}}(\boldsymbol{\theta}, \psi) = (\mathbf{t}, \psi_{\mathbf{t}}(\boldsymbol{\theta}, \psi)),$$

where $\psi_{\mathbf{t}}(\boldsymbol{\theta}, \psi) = \psi + (\boldsymbol{\theta} - \mathbf{t})^{\mathrm{T}}\mathbf{h}^* \circ \psi_0^{-1} \circ \psi$. Let $p(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ and $\ell(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ be density and log density functions under parameters $(\mathbf{t}, \psi_{\mathbf{t}}(\boldsymbol{\theta}, \psi))$, respectively. Also denote by $\dot{\ell}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ the first derivatives of $\ell(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ with respect to $\mathbf{t}$. Some derivative calculations then yield

$$\dot{\ell}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w}) = \begin{pmatrix} \eta_{\mathbf{t}}[\psi_{\mathbf{t}}'\mathbf{D}_{\mathbf{t}}\mathbf{z} - \mathbf{h}_1^* \circ \psi_0 \circ \psi_{\mathbf{t}} + (\boldsymbol{\phi} - \mathbf{t}_1)^{\mathrm{T}}\nabla_{\mathbf{t}_1}(\mathbf{h}_1^* \circ \psi_0 \circ \psi_{\mathbf{t}})] \\ -\eta_{\mathbf{t}}\mathbf{h}_2^* \circ \psi_0 \circ \psi_{\mathbf{t}} + \xi_{\mathbf{t}}\mathbf{w} \\ \eta_{\mathbf{t}}(\mathbf{x} - \mathbf{h}_3^* \circ \psi_0 \circ \psi_{\mathbf{t}}) \end{pmatrix}.$$

Here $\xi_{\mathbf{t}}, \eta_{\mathbf{t}}, \psi_{\mathbf{t}}$, and $\psi_{\mathbf{t}}'$ represent ξ, η, ψ, and ψ' evaluated at $(\mathbf{t}, \psi_{\mathbf{t}}(\mathbf{t}, \psi))$, respectively. $\nabla_{\mathbf{t}_1}(\mathbf{h}_1^* \circ \psi_0 \circ \psi_{\mathbf{t}})$ is the gradient of $\mathbf{h}_1^* \circ \psi_0 \circ \psi_{\mathbf{t}}$ with respect to $\mathbf{t}_1$. Observe that $\dot{\ell}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ converges to $\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau}_0; y, \mathbf{w})$ as $(\mathbf{t}, \boldsymbol{\theta}, \psi) \to (\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \psi_0)$. Moreover, using the similar arguments to those in the proof of Lemma 1, we can show that, for any $\delta > 0$, the class of functions $\dot{\ell}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})$ with $\psi \in \mathcal{S}_{0,n}$, $\|\psi - \psi_0\|_2 \le \delta$, $\|\mathbf{t} - \boldsymbol{\theta}_0\|_2 \le \delta$, and $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \le \delta$ is P-Donsker. Thus, Lemma 3.2 of Murphy and van der Vaart (1997) is applicable.

Using the same arguments as above, we can show that, for $(\mathbf{t}, \boldsymbol{\theta}, \psi)$ in a neighborhood of $(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \psi_0)$, the class of $p^{-1}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})\partial^2 p(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})/\partial\mathbf{t}\partial\mathbf{t}^{\mathrm{T}}$ is P-Donsker and is therefore P-Glivenko-Cantelli. Furthermore, as $(\mathbf{t}, \boldsymbol{\theta}, \psi) \to (\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \psi_0)$,

$$E[p^{-1}(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})\partial^2 p(\mathbf{t}, \boldsymbol{\theta}, \psi; y, \mathbf{w})/\partial\mathbf{t}\partial\mathbf{t}^{\mathrm{T}}] \to -E[\ell_{\boldsymbol{\theta}}^*(\boldsymbol{\tau}_0; y, \mathbf{w})]^{\otimes 2} + \mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{0}.$$

It follows that condition 3.14 in Murphy and van der Vaart (1997) holds. Thus the conditions in Theorem 3.1 of Murphy and van der Vaart (1997) reduce to the unbiasedness condition

$$\sqrt{n}P_0\dot{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \hat{\psi}_0; y, \mathbf{w}) = o_p(1),$$

where $\hat{\psi}_0$ is the estimator of ψ_0 under $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Using the same arguments as those in the proof of the convergence rate of $\hat{\boldsymbol{\tau}}$, we can deduce that $\|\hat{\psi}_0 - \psi_0\|_2 = O_p(n^{-r/(1+2r)})$.

Abbreviate $\dot{\ell}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \psi; y, \mathbf{w})$ to $\dot{\ell}(\psi; y, \mathbf{w})$. In view of the fact that $P_{\boldsymbol{\theta},\psi}\dot{\ell}(\boldsymbol{\theta}, \boldsymbol{\theta}, \psi; y, \mathbf{w}) = 0$ for all $(\boldsymbol{\theta}, \psi)$, we can decompose $P_0\dot{\ell}(\hat{\psi}_0; y, \mathbf{w})$ as $I_{n5} + I_{n6}$, where $I_{n5} = (P_0 - P_{\boldsymbol{\theta}_0,\hat{\psi}_0})\dot{\ell}(\psi_0; y, \mathbf{w})$ and $I_{n6} = (P_0 - P_{\boldsymbol{\theta}_0,\hat{\psi}_0})[\dot{\ell}(\hat{\psi}_0; y, \mathbf{w}) - \dot{\ell}(\psi_0; y, \mathbf{w})]$. Observe that $I_{n5} = P_0\{\dot{\ell}(\psi_0; y, \mathbf{w})[(p_0 - p_{\boldsymbol{\theta}_0,\hat{\psi}_0})/p_0 - \dot{\ell}_\psi(\boldsymbol{\tau}_0; y, \mathbf{w})[\psi_0 - \hat{\psi}_0]]\}$. By a Taylor's expansion, I_{n5} can be expressed as

$$I_{n5} = -(1/2)P_0[p_0^{-1}\dot{\ell}(\psi_0; y, \mathbf{w})d^2 p(\boldsymbol{\theta}_0, \psi_0 + t(\hat{\psi}_0 - \psi_0); y, \mathbf{w})/dt^2]|_{t=t^*},$$

where $0 < t^* < 1$. According to conditions C3–C5 and the rate of convergence of $\hat{\psi}_0$, $I_{n5} = o_p(n^{-1/2})$. Similarly, a Taylor's expansion and the rate of convergence of $\hat{\psi}_0$ yield $I_{n6} = o_p(n^{-1/2})$. This completes the proof of Theorem 3. □

A.8 Proof of Theorem 4

Proof In view of the consistency of $\hat{\boldsymbol{\tau}}$ and Proposition 2.1 of Huang et al. (2008), we can show that $\mathbb{P}_n[\dot{\ell}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\tau}}; y, \mathbf{w}) - \dot{\ell}_{\psi}(\hat{\boldsymbol{\tau}}; y, \mathbf{w})[\hat{\mathbf{h}}^*]]^{\otimes 2} \to \mathbf{I}(\boldsymbol{\theta}_0)$ in probability. According to some entropy calculations and the law of large numbers, it follows that $\hat{E}_{\boldsymbol{\theta\theta}} = \hat{A}_{\boldsymbol{\theta\theta}} + o_p(1)$, $\hat{E}_{\boldsymbol{\theta}\psi} = \hat{A}_{\boldsymbol{\theta}\psi} + o_p(1)$, and $\hat{E}_{\psi\psi} = \hat{A}_{\psi\psi} + o_p(1)$. We conclude that $\mathcal{E}_n \to \mathbf{I}(\boldsymbol{\theta}_0)$ in probability. This completes the proof of Theorem 4. □

Acknowledgments The authors express their thanks to an associate editor and two referees whose constructive comments improved the presentation. The Project described was supported by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), through Grant #UL1 TR000002 (C.S. Li).

References

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society Series A Statistics in Society*, *162*, 195–209.

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

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine*, *21*, 1461–1469.

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

Dietz, K., & Böhning, D. (1997). The use of two-component mixture models with one completely or partly known component. *Computational Statistics*, *12*, 219–234.

Delecroix, M., Härdle, W., & Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, *86*, 213–216.

Härdle, W., & Stoker, E. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, *84*, 986–995.

Härdle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statisitcs*, *21*, 157–178.

Hastie, T., Tibshirani, R. (1990). *Generalized additive models*. New York: Chapman & Hall/CRC.

Horowitz, J. L., & Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariate. *Journal of the American Statistical Association*, *91*, 1632–1640.

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., Zhang, Y., Hua, L. (2008). A least-squares approach to consistent information estimation in semiparametric models. Technical Report 2008–3, University of Iowa, Department of Biostatistics

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*, 71–120.

Johnson, N. L., Kotz, S., Kemp, A. W. (2005). *Univariate discrete distributions* (3rd ed.). New York: Wiley.

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

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*, 1–14.

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, S. E., Lin, Y., Shih, W. C. J. (2004). Analyzing excessive no changes in clinical trials with clustered data. *Biometrics*, *60*, 257–267.

Murphy, S. A., & van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *Annals of Statistics*, *25*, 1471–1509.

Murphy, S. A., van der Vaart, A. W. (1999). Observed information in semi-parametric models. *Bernoulli*, *5*, 381–412.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sörensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, *19*, 25–43.

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

Schumaker, L. (1981). *Spline functions: basic theory*. New York: Wiley.

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

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, *1*, 140–144.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, *13*, 689–705.

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

Sun, J., Kopciukb, K. A., & Lu, X. (2008). Polynomial spline estimation of partially linear single-index proportional hazards regression models. *Computational Statistics and Data Analysis*, *53*, 176–188.

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.

Yau, K. K. W., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, *20*, 2907–2920.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*, 1042–1054.