Statistica Sinica Preprint No: SS-2016-0534.R2	
Title	A NONPARAMETRIC REGRESSION MODEL FOR PANEL COUNT DATA ANALYSIS
Manuscript ID	SS-2016-0534.R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0534
Complete List of Authors	Huadong Zhao Ying Zhang Xingqiu Zhao and Zhangsheng Yu
Corresponding Author	Zhangsheng Yu
E-mail	yuzhangsheng@sjtu.edu.cn
Notice: Accepted version subject to English editing.	

A NONPARAMETRIC REGRESSION MODEL FOR PANEL COUNT DATA ANALYSIS

Huadong Zhao[1], Ying Zhang[2], Xingqiu Zhao[3]and Zhangsheng Yu[4]*

[1]*East China Normal University,* [2]*Indiana University,*

[3]*The Hong Kong Polytechnic University and* [2,4]*Shanghai Jiao Tong University*

Abstract: Panel count data are commonly encountered in analysis of recurrent events where the exact event times are unobserved. To accommodate the potential non-linear covariate effect, we consider a non-parametric regression model for panel count data. The regression B-splines method is used to estimate the regression function and the baseline mean function. The B-splines-based estimation is shown to be consistent and the rate of convergence is obtained. Moreover, the asymptotic normality for a class of smooth functionals of regression splines estimators is established. Numerical studies are carried out to evaluate the finite sample properties. Finally, we apply the proposed method to analyze the nonlinear effect of one of interleukin functions with the risk of childhood wheezing.

Key words and phrases: Empirical process; Maximum pseudolikelihood estimator; Regression splines; Wheezing.

1. Introduction

In many longitudinal studies, subjects' information is observed at several random discrete time points during the follow-up period. Instead of the event times, only the number of events before each encounter (visit) is observed. The number of visits and visit times vary among individuals. This kind of data is often referred to as panel count data. Unlike recurrent event data, in panel count data, the exact times of the events are not observed, and the number of events in each observation interval can be greater than one. For example, in a childhood wheezing study conducted at the Indiana University School of Medicine, 105 infants at a high risk of developing childhood asthma were followed for 5 years. The interleukin function, airway reactivity, and demographic information were collected at enrollment and the occurrence (number) of wheezing episodes were collected on a monthly basis over the entire follow-up time, which resulted in panel count data.

Statistical methods for panel count data have been studied extensively in the past three decades. Thall and Lachin (1988) studied the data from NCGS using a marginal model. They proposed a non-parametric estimation of the rate of the counting process. Sun and Kalbfleisch (1995) estimated the mean function using the isotonic regression method. Lee and Kim (1998) analyzed correlated panel data. Wellner and Zhang (2000) studied the large sample theory for the

likelihood-based non-parametric estimates for panel count data. They showed that the non-parametric maximum pseudolikelihood estimator (NPMPLE) based on the nonhomogeneous Poisson process is exactly the isotonic regression estimator of Sun and Kalbfleisch (1995). In addition, they proved the consistency of NPMPLE and derived the convergence rate. Zhang (2002) investigated a semi-parametric regression model of panel count data with the pseudolikelihood approach. Nielsen and Dean (2008a,b) considered an estimating equation for recurrent event panel data without providing theoretical properties. Other methods of semi-parametric regression analysis for panel count data were developed by Sun and Wei (2000), Wellner and Zhang (2007) and Zhu et al. (2015). For a comprehensive review of statistical methods for panel count data analysis, see Sun and Zhao (2013).

The spline-based functional analysis has also been conducted for panel count data. Lu, Zhang, and Huang (2007, 2009) employed monotonic B-splines to model the mean function, and developed a spline-based semi-parametric proportional mean model. These papers demonstrated the benefits of spline-based estimators in analyzing panel count data. However, the effects of covariates on the mean function were assumed to be multiplicative in Lu, Zhang, and Huang (2009), which may be too restrictive in some applications.

In practice, the functional forms of covariate effects are often unknown or

might be too complicated to be explicitly specified. For example, in the aforementioned childhood wheezing study, it is of interest to ascertain the effect of the interleukin function during infancy with the risk of wheezing occurrence. However the functional form of the interleukin 5 effect is unknown and can possibly be non-linear as shown in Figure 3. Proper analysis of such data is lacking because of non-existence of non-parametric regression in the literature of panel count data.

In this article, the regression spline technique is used to model the regression function of the covariates and the baseline mean function using B-splines. We modify the proportional mean model for panel count data of Lu, Zhang, and Huang (2009) by incorporating non-linear covariate effects, and conduct a B-splines based functional analysis for the covariate effect using the pseudolikelihood approach of Zhang (2002) for its numerical advantages. Our method shows that the B-splines-based NPMPLEs of the baseline mean and the regression function are consistent and converge to the true corresponding functions at the rate of $r/(1+2r)$, where r is the degree of smoothness of the baseline mean and the regression function. The asymptotic normality for a class of functionals of the B-splines-based NPMPLE of the regression function is derived for making a statistical inference.

The remainder of this paper is organized as follows. Section 2 presents the

model and estimation procedure. Section 3 illustrates the asymptotic properties. The finite sample performance of the proposed estimators is numerically evaluated by simulation studies in Section 4. Section 5 applies the proposed method to the aforementioned wheezing study. Section 6 concludes the paper with some discussions. Technical details are outlined in the online Supplementary Materials, which can be accessed via *ftp://public.sjtu.edu.cn/* using account name *yuzhangsheng* and password *public*.

2. Model and Regression B-spline Estimators

First, we introduce some notation. Let $\mathbb{N} = \{\mathbb{N}(t) : t \geq 0\}$ be a counting process and K be the number of observation times. Denote $T = \{T_j : j = 1, 2, \ldots, K\}$ the vector of ordered observation times with $0 < T_1 < \ldots < T_K$. The counting process $\mathbb{N}(t)$ registers the number of events in a sequence of intervals made by $T_1, T_2, \cdots, T_K$, which results in panel observed event counts satisfying $0 \leq \mathbb{N}(T_1) \cdots \leq \mathbb{N}(T_K)$. The observed data of a subject consist of $X = (K, T, \mathbb{N}, \underline{Z})$, where $T = (T_1, \cdots, T_K)$, $\mathbb{N} = (\mathbb{N}(T_1), \cdots, \mathbb{N}(T_K))$ and $\underline{Z}$ is a vector of p-dimensional covariates. Suppose we have n independent and identically distributed copies of X denoted by

$$\{X_i = (K_i, T_i, \mathbb{N}^{(i)}, \underline{Z}_i), i = 1, \cdots, n\}$$

with $T_i = (T_{i,1}, \cdots, T_{i,K_i})$, $\underline{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$ and $\mathbb{N}^{(i)} = (\mathbb{N}(T_{i,1}), \cdots, \mathbb{N}(T_{i,K_i}))$.

We consider the following mean model for panel count data:

$$E\{\mathbb{N}(t)|\underline{Z}\} = \Lambda_0(t)\exp\{\beta_0(\underline{Z})\}, \tag{1}$$

where $\beta_0(\underline{Z}) = \beta_{01}(Z_1) + \ldots + \beta_{0p}(Z_p)$, with $\beta_{0j}(\cdot)$ an unknown function without a pre-specified functional form of Z_j for $j = 1, 2, \cdots, p$, and $\Lambda_0(t)$ is the unknown cumulative baseline mean function. To ensure the mean model (1) identifiable, we require $\beta_{0j}(0) \equiv 0$ for $j = 1, 2, \cdots, p$. For the rest of the paper, we only present the model for regression function $\beta_0(Z)$ with one covariate unless otherwise specified for the sake of algebraic convenience. The estimation procedure and theoretical justification can be readily generalized to models with $p > 1$.

Note that for subject i, the panel cumulative counts $\mathbb{N}(T_{i,1}), \ldots, \mathbb{N}(T_{i,K_i})$ are correlated and it is generally difficult to specify their correlation structure. In this paper, we adopt the pseudolikelihood approach developed by Zhang (2002), which models the marginal distribution for each $\mathbb{N}(T_{ij})$ and assumes mutually independence among $\mathbb{N}(T_{i,1}), \ldots, \mathbb{N}(T_{i,K_i})$. It follows that the log pseudolikelihood function for (β, Λ) is given by

$$l_{ps}(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\left[\mathbb{N}(\mathrm{T}_{i,j})\left\{\log\Lambda(\mathrm{T}_{i,j}) + \beta(Z_i)\right\} - \Lambda(\mathrm{T}_{i,j})\exp\{\beta(Z_i)\}\right]. \tag{2}$$

We propose to estimate the functions $\log\Lambda$ and $\beta(z)$ using the regression B-splines method. Let $T = \{t_i\}_{i=1}^{m_{n1}+2l_1}$ be a sequence of knots that partition the time interval, $[\sigma_1, \tau_1]$, into m_{n1} subintervals $J_{1i} = [t_{l_1+i}, t_{l_1+i+1})$, $i = 0, \ldots, m_{n1}-1$ and $J_{1m_{n1}} = [t_{l_1+m_{n1}}, t_{l_1+m_{n1}+1}]$ with $\sigma_1 = t_1 = \cdots = t_{l_1} < t_{l_1+1} < \ldots < t_{m_{n1}+l_1} < t_{m_{n1}+l_1+1} = \cdots = t_{m_{n1}+2l_1} = \tau_1$, where $m_{n1} = O(n^{v_1})$ for $0 < v_1 < \frac{1}{2}$ is the number of interior knots and l_1 boundary knots $t_1 = \cdots = t_{l_1}$ and $t_{m_{n1}+l_1+1} = \cdots = t_{m_{n1}+2l_1}$ on each side are needed to complete the B-splines basis functions for the l_1th order B-splines (Schumaker, 2007).

Let $\Psi_{l_1,t}$ denote the class of B-splines of order $l_1 \geq 1$ that consists of functions s_1 satisfying (i) for each interval J_{1i}, s_1 is a polynomial of order l_1 for $i = 0, \ldots, m_{n1}$; (ii) for $l_1 \geq 2$, s_1 is $l_1^{'}$ times continuously differentiable on $[\sigma_1, \tau_1]$ for $0 \leq l_1^{'} \leq l_1 - 2$. The class $\Psi_{l_1,t}$ can be spanned by the B-splines basis functions $\{B_i^{(1)}(t), 1 \leq i \leq q_{n1}\}$ with $q_{n1} = m_{n1} + l_1$,

$$\Psi_{l_1,t} = \left\{ \sum_{i=1}^{q_{n1}} \alpha_i B_i^{(1)}(t) : \underline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{q_{n1}}) \in \mathcal{R}^{q_{n1}} \right\}.$$

Since $\log\Lambda_0(t)$ is a monotonically non-decreasing function, we restrict the non-parametric estimator in the subclass of $\Psi_{l_1,t}$.

$$\psi_{l_1,t} = \left\{ \sum_{i=1}^{q_{n1}} \alpha_i B_i^{(1)}(t) : \alpha_1 \leq \alpha_2 \ldots \leq \alpha_{q_{n1}} \right\},$$

as any member of $\psi_{l_1,t}$ is monotonically non-decreasing (Schumaker, 2007).

Similarly, let $Z = \{z_i\}_{i=1}^{m_{n2}+2l_2}$ be a sequence of knots that partition the interval, $[\sigma_2, \tau_2]$, into m_{n2} subintervals $J_{2i} = [z_{l_2+i}, z_{l_2+i+1})$, for $i = 0, \ldots, m_{n2}-1$ and $J_{2m_{n2}} = [z_{l_2+m_{n2}}, z_{l_2+m_{n2}+1}]$ with $\sigma_2 = z_1 = \cdots = z_{l_2} < z_{l_2+1} < \ldots < z_{m_{n2}+l_2} < z_{m_{n2}+l_2+1} = \cdots = z_{m_{n2}+2l_2} = \tau_2$, where $m_{n2} = O(n^{v_2})$ for $0 < v_2 < \frac{1}{2}$ is the number of interior knots, and l_2 is the order of B-splines.

Let $\Phi_{l_2,z}$ denote the class of B-splines of order $l_2 \geq 1$ that consists of functions s_2 satisfying the similar conditions as the $\Psi_{l_1,t}$. The class $\Phi_{l_2,z}$ can be spanned by the B-splines basis function $\{B_i^{(2)}(z), 1 \leq i \leq q_{n2}\}$ with $q_{n2} = m_{n2} + l_2$,

$$\Phi_{l_2,z} = \left\{ \sum_{i=1}^{q_{n2}} \eta_i B_i^{(2)}(z) : \underline{\eta} = (\eta_1, \eta_2, \ldots, \eta_{q_{n2}}) \in \mathcal{R}^{q_{n2}} \right\}.$$

In the forthcoming simulation studies and data application, we use the popular cubic regression B-splines ($l_1 = l_2 = 4$) to estimate the model parameters $(\beta, \log \Lambda)$. Specifically, the baseline mean function $\Lambda(t)$ and the regression function $\beta(Z)$ are modelled by the B-splines,

$$\log \Lambda(t) = \sum_{l=1}^{q_{n1}} \alpha_l B_l^{(1)}(t) = \underline{\alpha}^{\mathrm{T}} B^{(1)}(t) \quad \text{and} \quad \beta(z) = \sum_{l=1}^{q_{n2}} \eta_l B_l^{(2)}(z) = \underline{\eta}^{\mathrm{T}} B^{(2)}(z),$$

respectively, where $B^{(1)}(t) = \{B_1^{(1)}(t), \cdots, B_{q_{n1}}^{(1)}(t)\}^{\mathrm{T}}$ and $B^{(2)}(z) = \{B_1^{(2)}(z), \cdots, B_{q_{n2}}^{(2)}(z)\}^{\mathrm{T}}$ for $q_{n1} = m_{n1} + 4$ and $q_{n2} = m_{n2} + 4$.

After substituting the B-splines expression of $\Lambda(\cdot)$ and $\beta(\cdot)$ in (2), the log pseudolikelihood function can be written as

$$\begin{aligned} l_{ps}(\gamma) &= \sum_{i=1}^{n}\sum_{j=1}^{K_i}[\mathbb{N}(\mathrm{T}_{i,j})\{\underline{\alpha}^{\mathrm{T}}B^{(1)}(\mathrm{T}_{i,j}) + \underline{\eta}^{\mathrm{T}}B^{(2)}(\mathrm{Z}_i)\} - \exp\{\underline{\alpha}^{\mathrm{T}}B^{(1)}(\mathrm{T}_{i,j}) + \underline{\eta}^{\mathrm{T}}B^{(2)}(Z_i)\}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{K_i}[\mathbb{N}(\mathrm{T}_{i,j})\{B(\mathrm{T}_{i,j}, Z_i)^{\mathrm{T}}\gamma\} - \exp\{B(\mathrm{T}_{i,j}, Z_i)^{\mathrm{T}}\gamma\}], \end{aligned}$$

where $B(t,z) = (B^{(1)}(t)^{\mathrm{T}}, B^{(2)}(z)^{\mathrm{T}})^{\mathrm{T}}$ and $\gamma = (\underline{\alpha}^{\mathrm{T}}, \underline{\eta}^{\mathrm{T}})^{\mathrm{T}}$.

Hence, computation of a B-splines-based NPMPLE is converted to a convex programming problem with the linear equality-inequality constraints

$$A\gamma = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{q_{n1}} \\ \eta_1 \\ \vdots \\ \eta_{q_{n2}} \end{pmatrix} \geq 0.$$

To ensure identifiability, we also need to add the zero-intercept constraint $\sum_{l=1}^{q_{n2}} \eta_l B_l^{(2)}(0) = 0$ for the regression splines.

The algorithm for the convex programming problem subject to linear equality-inequality constraints developed by Lange (1994) is applied to this problem for

computing the B-splines based NPMPLE of (β_0, Λ_0).

3. Asymptotic Results

In this section we study the asymptotic properties of the B-splines-based NPMPLE of (β_0, Λ_0), $(\widehat{\beta}_n(\cdot), \widehat{\Lambda}_n(\cdot))$. Denote H the distribution for covariate Z on $\mathbb{R}$ and $\mathfrak{B}$ the collection of Borel sets on $\mathbb{R}$. Let $\mathcal{B}[0,\tau] = \{B \bigcap [0,\tau] : B \in \mathfrak{B}\}$ and let μ_2 denote the probability measure induced by H. Following the notation in Wellner and Zhang (2007), for $B \in \mathcal{B}[0,\tau]$ and $C \in \mathfrak{B}$, we define

$$\nu(B \times C) = \int_C \sum_{k=1}^{\infty} P(K = k|Z = z) \sum_{j=1}^{k} P(T_{k,j} \in B|K = k, Z = z) d\mu_2(z)$$

with $\mu_1(B) = \nu(B \times \mathbb{R})$.

We consider two classes of functions

$$\mathcal{F}_1 = \{\beta : \beta \text{ is a continuous function in } \mathbb{R}\} \quad \text{and}$$

$$\mathcal{F}_2 = \{\Lambda : [0,\infty) \to [0,\infty) | \Lambda \text{ is nondecreasing}, \Lambda(0) = 0\}.$$

We aim to find the B-splines NPMPLE $(\widehat{\beta}_n(\cdot), \widehat{\Lambda}_n(\cdot))$ in the parameter space $\Theta = \mathcal{F}_1 \times \mathcal{F}_2$. In order to study the asymptotic properties of $(\widehat{\beta}_n(\cdot), \widehat{\Lambda}_n(\cdot))$, we define an L_2-metric $d(\theta_1, \theta_2)$ for the parameter space Θ given by

$$d(\theta_1, \theta_2) = \left\{ \|\beta_1 - \beta_2\|^2_{L_2(\mu_2)} + \|\Lambda_1 - \Lambda_2\|^2_{L_2(\mu_1)} \right\}^{1/2} \text{ for } \theta_i = (\beta_i, \Lambda_i) \in \Theta \ i = 1, 2.$$

We also postulate the following regularity conditions.

C1 The true parameter, $\theta_0 = (\beta_0, \Lambda_0) \in \Theta = \mathcal{F}_1 \times \mathcal{F}_2$.

C2 The maximum spacing between the consecutive knots, defined as

$$\Delta_1 = \max_{l_1+1 \le i \le m_{n1}+l_1+1} |t_i - t_{i-1}| = O(n^{-v_1})$$

$$\Delta_2 = \max_{l_2+1 \le i \le m_{n2}+l_2+1} |z_i - z_{i-1}| = O(n^{-v_2})$$

satisfying $\Delta_j/\delta_j \le M_j$ uniformly in n_j, where $M_j > 0$ is a constant for $j = 1, 2$ with $\delta_1 = \min_{l_1+1 \le i \le m_{n1}+l_1+1} |t_i - t_{i-1}|$ and $\delta_2 = \min_{l_2+1 \le i \le m_{n2}+l_2+1} |z_i - z_{i-1}|$.

C3 The true baseline mean function Λ_0 is rth bounded differentiable in $O[T]$ with $r > 1$, where $O[T] = [\sigma_1, \tau_1]$. The regression function β_0 is also rth bounded differentiable in $O[Z]$ with $r > 1$, where $O[Z] = [\sigma_2, \tau_2]$. Moreover, there exists a positive constant c_1 such that $\Lambda_0'(t) \ge c_1$ for $t \in O[T]$.

C4 For $j = 1, \ldots, K$, $T_{K,j} \in [0, \tau]$ for some $\tau \in (0, \infty)$. The measure $\mu_1 \times \mu_2$ on $([0, \tau], \mathcal{B}[0, \tau])$ is absolutely continuous with respect to ν.

C5 The support of the covariate distribution H is a bounded set in $\mathbb{R}$, denoted as $\mathcal{Z}$.

C6 For some functions h_1 and h_2, if $h_1(Z) + h_2(T) = 0$ with probability 1 for all Z and T, then $h_1 \equiv 0$ and $h_2 \equiv 0$.

C7 For some $\eta \in (0, 1)$ and bounded function $g \in \mathbb{R}^d$,

$$a^{\mathrm{T}} var(g(Z)|U)a \geq \eta a^{\mathrm{T}} E(g(Z)g(Z)^{\mathrm{T}}|U)a \quad \text{a.s. for all} \quad a \in \mathbb{R}^d,$$

where (U, Z) has distribution $\nu/\nu(\mathbb{R}^+ \times \mathcal{Z})$.

C8 The function $M_0^{ps}(X) = \sum_{j=1}^{K} \mathbb{N}(T_j) \log \mathbb{N}(T_j)$ satisfies $PM_0^{ps}(X) < \infty$.

C9 The number of observations is finite; that is, there exists a positive integer k_0 such that $P(K \leq k_0) = 1$.

C10 For some $C \geq 0$, $E(e^{C\mathbb{N}(t)}|Z)$ is uniformly bounded for $Z \in \mathcal{Z}$.

Remark 1. C1 indicates that the true model parameters are in the estimation space. C2 can be viewed as knots selection criteria that are easily satisfied using quantile spaced knots. C3 is required for the proof of rate of convergence and is a reasonable assumption in view of application. C4 and C6 are required for the identifiability of the non-parametric model. C5 and C7 are the technical conditions for proving the forthcoming theorems that were similarly provided in Wellner and Zhang (2007). C8-C10 are exactly the conditions C4, C9 and C10, respectively, given in Wellner and Zhang (2007).

Theorem 1. *Suppose that* $C1 - C10$ *hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean model (1), then*

$$d(\hat{\theta}_n, \theta_0) \to_p 0 \quad as \quad n \to \infty.$$

Theorem 2. *Suppose that* $C1 - C10$ *hold, if* $v_1 = v_2 = 1/(1+2r)$*, then*

$$n^{r/(1+2r)} d(\hat{\theta}_n, \theta_0) = O_p(1) \ as \ n \to \infty.$$

Theorem 3. *Let* H_Λ *consist of all the functions in* $[\sigma_1, \tau_1]$ *whose total variation is bounded by 1. Let* H_β *consist of all the functions in* $[\sigma_2, \tau_2]$ *whose* r*th derivatives are bounded by 1(*$r > 1$*). We take* $\{\hat{\Lambda}_n(t) - \Lambda_0(t), \hat{\beta}_n(z) - \beta_0(z)\}$ *as a stochastic class in* $l^\infty(\mathrm{H}_\Lambda \times \mathrm{H}_\beta)$ *whose value for* $(h_1, h_2) \in \mathrm{H}_\Lambda \times \mathrm{H}_\beta$ *is defined as*

$$\int (\hat{\Lambda}_n(t) - \Lambda_0(t)) dh_1(t) + \int (\hat{\beta}_n(z) - \beta_0(z)) dh_2(z).$$

Then under C1-C10, $n^{1/2}\{\hat{\Lambda}_n(t) - \Lambda_0(t), \hat{\beta}_n(z) - \beta_0(z)\}$ converges in distribution to a mean-zero Gaussian process with variance $V(h_1, h_2)$ in the metric

space $l^\infty(\mathrm{H}_\Lambda \times \mathrm{H}_\beta)$, where

$$V(h_1^*, h_2^*) = E\{\varphi^2(\theta_0; X, Z)[h_1^*, h_2^*]\}$$

with

$$\varphi(\theta_0; X, Z)[h_1^*, h_2^*] = \sum_{j=1}^{K} \Big\{ \left(\frac{\mathbb{N}(T_j)}{\Lambda_0(T_j)} - \exp\{\beta_0(Z)\} \right) h_1^*(T_j) + \big[\mathbb{N}(T_j) - \Lambda_0(T_j) \exp\{\beta_0(Z)\}\big] h_2^*(Z) \Big\},$$

in which h_1^* and h_2^* satisfy $Q_2(h_1^*, h_2^*)(t) = h_1(t)$ and $Q_1(h_1^*, h_2^*)(z) = h_2(z)$ with Q_1 and Q_2 given in the Supplementary Materials.

Theorem 3 not only describes the asymptotic distribution of a class of estimated smooth functionals of the model parameters, but is also useful to construct a non-parametric test for the covariate effect on the underlying counting process: $H_0 : \beta_0(z) = 0$ for all z. To do so, we need to identify a specific $h^* = (h_1^*, h_2^*)$ such that

$$Q_2(h_1^*, h_2^*)(t) = 0 \quad \text{and} \quad Q_1(h_1^*, h_2^*)(z) = H(z)$$

The selection of such h^* results in

$$\sqrt{n} \int \left(\hat{\beta}_n(z) - \beta_0(z) \right) dH(z) \to_d N(0, \sigma_{\beta_0}^2)$$

with $\sigma^2_{\beta_0}$ given by

$$\begin{aligned}\sigma^2_{\beta_0} &= E\left[\sum_{j=1}^{K}\left\{\left(\mathbb{N}(T_j)-\Lambda_0(T_j)\exp(\beta_0(Z))\right)h_2^*(Z)+\left(\frac{\mathbb{N}(T_j)}{\Lambda_0(T_j)}-\exp(\beta_0(Z))\right)h_1^*(T_j)\right\}\right]^2\\ &= E\left[\sum_{j=1}^{K}\left\{\left(\mathbb{N}(T_j)-\Lambda_0(T_j)\exp(\beta_0(Z))\right)\left(h_2^*(Z)-\frac{E\{h_2^*(Z)\exp(\beta_0(Z))|K,T_j\}}{E\{\exp(\beta_0(Z))|K,T_j\}}\right)\right\}\right]^2\end{aligned}$$

based on Theorem 3 (the proof is provided in the Supplementary Materials). It then leads to constructing the test statistic

$$T_n=\int\hat{\beta}_n(z)d\mathbb{H}_n(z)=\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_n(Z_i),$$

where $\mathbb{H}_n$ is the empirical distribution of Z. It can be easily shown (the proof is provided in the Supplementary Materials) that under the null hypothesis H_0, T_n is asymptotically normally distributed with mean 0.

Remark 2. Theorem 2 shows that the proposed B-splines-based NPMPLE $(\hat{\beta}_n(\cdot),\hat{\Lambda}_n(\cdot))$ achieves the $r/(1+2r)$ convergence rate. This convergence rate was shown to be optimal in spline-smoothing non-parametric estimation literatures (Speckman, 1985; Zhang et al., 2010). Theorem 3 demonstrates the asymptotic normality of the functionals of the proposed non-parametric estimators and facilitates possible procedures for making an inference on $\beta_0(\cdot)$ and $\Lambda_0(\cdot)$. However,

to do so, we need to estimate the standard error of a functional of the B-splines-based NPMPLE, which is a daunting job in view of the complicated expression of the asymptotic variance. The bootstrap estimation for standard error is a viable alternative for inference due to the numerical advantage in NPMPLE.

4. Simulation Studies

In this section, we use simulation studies to evaluate the finite sample performance of the proposed estimator. We generated n independent and identically distributed observations $X_i = (K_i, T_i, \mathbb{N}^{(i)}, Z_i)$ for $i = 1, \ldots, n$. For subject i, the number of visits (or encounters), K_i, was generated from a discrete uniform distribution on $\{1, 2, \ldots, N_1\}$ for a finite N_1. Given the number of visits K_i, the visit times vector $T_i = (T_{i,1}, \ldots, T_{i,K_i})$ were K_i ordered random draws from a uniform distribution Unif$(0, T^{\infty})$, where T^{∞} was the maximum length of follow up time. Between two consecutive visit times $T_{i,j-1}$ and $T_{i,j}$, the number of recurrent events was generated from a Poisson process model with interval event counts following the Poisson distribution given by

$$\mathbb{N}_{i,j} - \mathbb{N}_{i,j-1} \sim Po\left\{2(T_{i,j} - T_{i,j-1}) \exp(\beta(Z_i))\right\} \ (j = 1, \ldots, K_i),$$

which results in $\mathbb{N}_{i,j}$, the number of cumulative events at $T_{i,j}$ following the Poisson distribution with mean $2T_{i,j} \exp\{\beta(Z_i)\}$, conditional on Z_i.

We used cubic B-splines for the non-parametric estimation of $\log \Lambda(t)$. Seven interior knots were used with the locations determined by quantiles of the total observation times $\{T_{i,j} : i = 1, 2, \ldots, n; j = 1, 2, \ldots, K_i;\}$ so that there were approximately equal numbers of observations in each interval. Similarly, for the regression function $\beta(Z)$, the cubic B-splines non-parametric estimate was also calculated with the seven interior knots chosen to be the quantiles of observed covariate values $\{Z_i; i = 1, 2, \ldots, n\}$.

We performed the simulation studies under four different settings with sample sizes 100 and 400. In all these settings, the cumulative baseline mean function was $\Lambda(t) = 2t + 1$. The maximum number of visits per subject was set as $N_1 = 6$, and the maximum follow-up time was $T^{\infty} = 10$. The covariate was generated from a uniform $[0, 1]$ distribution, and the simulation was repeated for 1000 times in each scenario.

S1. Null regression function $\beta(Z) = 0$.

S2. Linear regression function $\beta(Z) = 2Z$.

S3. Non-linear regression function $\beta(Z) = Beta(Z, 2, 2)$, where $Beta(\cdot)$ is the $Beta$ density function.

S4. Non-linear regression function that oscillates at 0: $\beta(Z) = 1.5 \sin(2\pi Z) I(Z \leq 0.5) + 0.5 \sin(2\pi Z) I(Z > 0.5)$ where $I(\cdot)$ is the indicator function.

Estimation results are presented in Figures 1 and 2 for the four considered settings. In all these figures, the solid curve is the true regression function $\beta_0(z)$, and the dotted, dashed and dash-dotted curves are the pointwise 2.5-quantile, mean and 97.5-quantile of $\hat{\beta}_n(z)'s$ respectively, with the estimated $\beta_0(z)$ based on 1000 replicates. It can be seen from Figures 1 and 2 that the mean curves of the estimated regression function are almost overlapped with the corresponding true curves in all the settings. The bandwidth between the pointwise 97.5 and 2.5 quantile curves decreases as sample size increases from 100 to 400. The simulation studies numerically justify the estimation consistency stated in Theorem 1.

As described in the preceding section, the result of Theorem 3 allows us to construct a test statistic to make an inference about whether covariate Z affects the underlying counting process $\mathbb{N}(t)$ by testing the null hypothesis $H_0 : \ \beta_0(z) = 0$ for all z. To evaluate the test statistic T_n, we estimate the standard error of T_n by the bootstrap method based on 100 resamplings with a replacement: that is, the estimate of $se(T_n)$ is given by the standard deviations of the 100 estimates of T_n from the bootstrap samples.

Table 1 presents the simulation results for T_n and the probability of rejecting H_0 under various scenarios. It can be seen from the table that the estimation bias of $\int \hat{\beta}_n(z) d\mathbb{H}_n(z)$ is virtually negligible, and the empirical standard deviation of the estimator based on 1000 repetitions decreases as sample size increases.

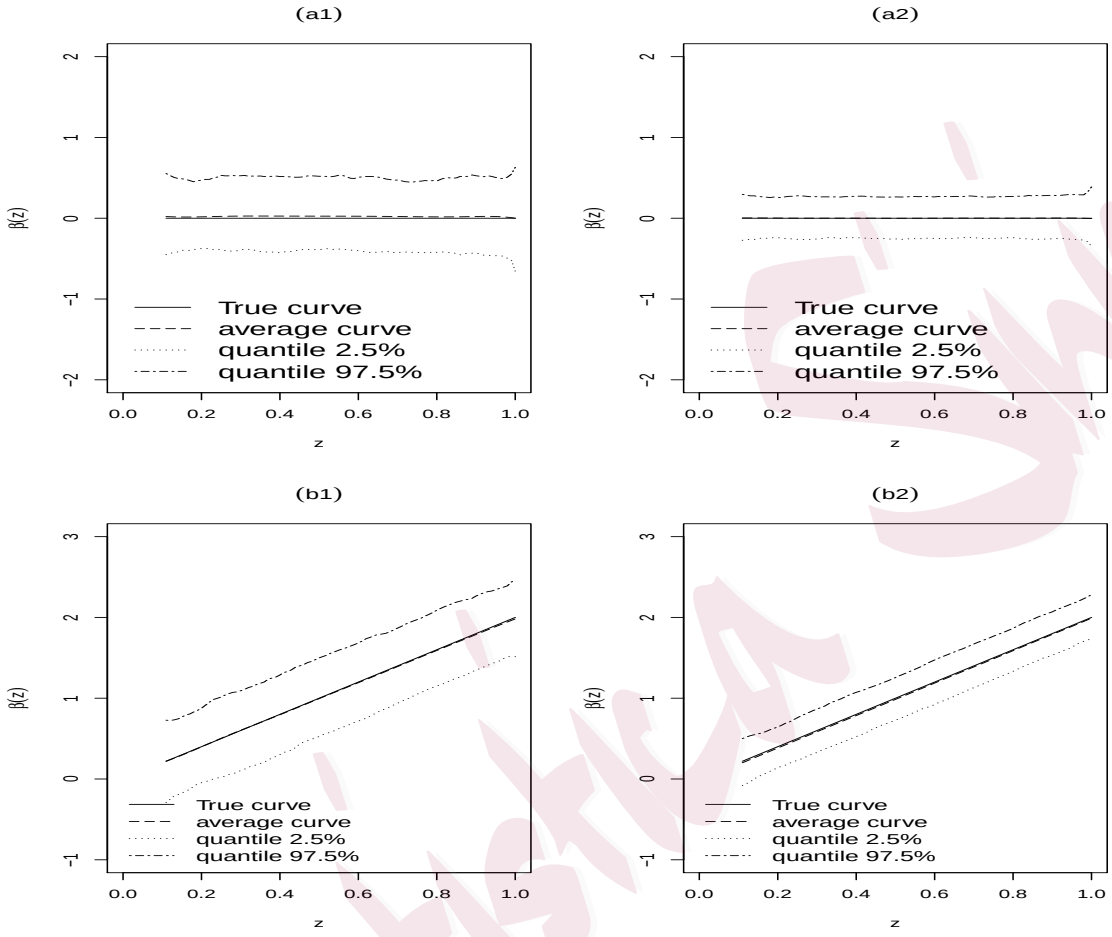

Figure 1: Estimation results for the regression function: The solid curve is the true regression function $\beta_0(z)$, the dotted, dashed and dash-dotted curves are the pointwise 2.5-quantile, mean and 97.5-quantile of $\hat{\beta}_n(z)$'s, respectively; (a1)-(a2) are the results of $\beta_0(Z) = 0$ under sample sizes 100 and 400; (b1)-(b2) are the results of $\beta_0(Z) = 2 * Z$ under sample sizes 100 and 400.

However, the average of the bootstrap standard error estimates is slightly smaller than the empirical standard deviation in all the settings, which results in a slightly inflated type I error and the testing power. We believe the reason for

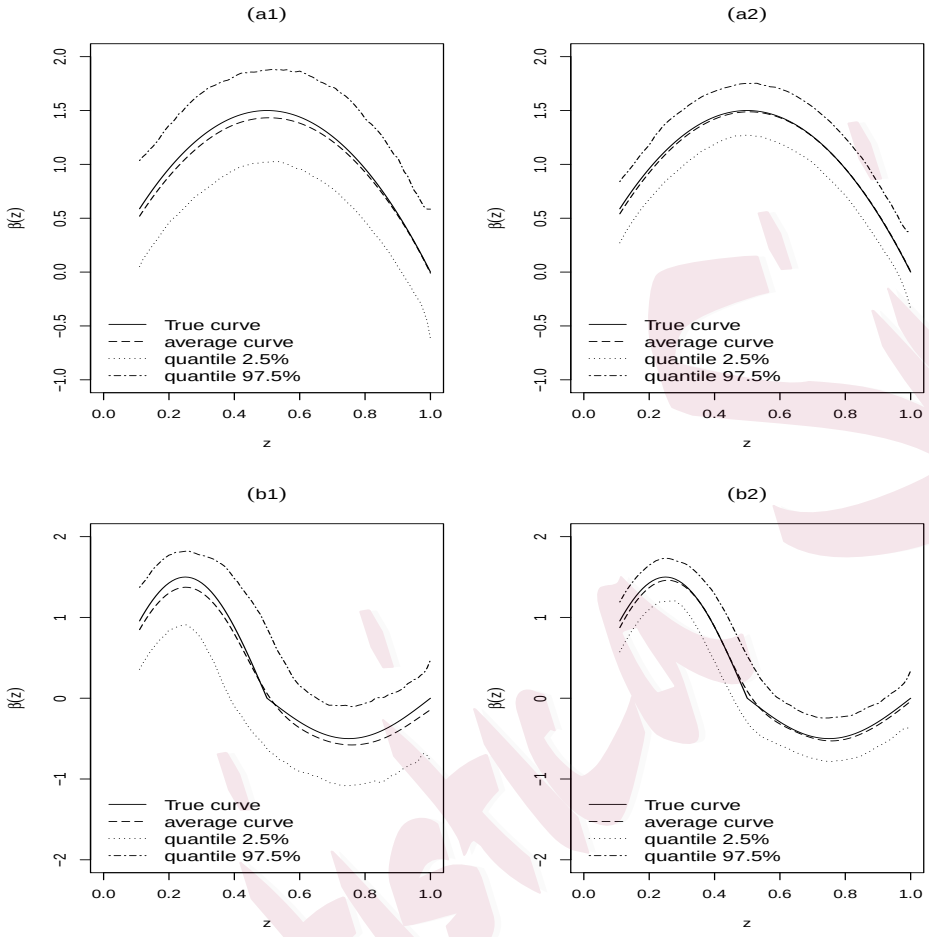

Figure 2: Estimation results for the regression function: The solid curve is the true regression function $\beta_0(z)$, the dotted, dashed and dash-dotted curves are the pointwise 2.5-quantile, mean and 97.5-quantile of $\hat{\beta}_n(z)$'s, respectively; (a1)-(a2) are the results of $\beta_0(Z) = Beta(Z, 2, 2)$ under sample sizes 100 and 400; (b1)-(b2) are the results of $\beta_0(Z) = 1.5\sin(2\pi Z)I(Z \leq 0.5) + 0.5\sin(2\pi Z)I(Z > 0.5)$ under sample sizes 100 and 400.

the underestimation of the standard error is due to the use of pseudolikelihood approach, which causes the observed count data to be "overdispersed" in terms of the proposed model. This fact of underestimation in standard error was also

Table 1: Simulation results. SP, sample size; True value, the exact value of $\int \beta_0(z)dH(z)$; M-C SD, Monte Carlo standard deviation; ASE, average of bootstrap standard errors; Prob., probability.

Setting	$\beta(Z)$	SP	True value	Bias	M-C SD	ASE	Prob. of rejecting H_0
I	0	100	0	0·021	0·209	0·176	0·116
		400		0·001	0·124	0·115	0·086
II	$2 * Z$	100	1·000	-0·005	0·224	0·186	0·980
		400		-0·014	0·127	0·116	1·000
III	$Beta(Z, 2, 2)$	100	1·000	-0·053	0·21	0·173	0·989
		400		-0·017	0·12	0·111	1·000
IV	$1.5\sin(2\pi Z)I(Z \leq 0.5)$	100	0·318	-0·089	0·221	0·193	0·256
	$+0.5\sin(2\pi Z)I(Z > 0.5)$	400		-0·030	0·121	0·113	0·727

presented in a study of B-splines-based semi-parametric models for panel count data by Hua, Zhang, and Tu (2014). It implies that one may need to have a p-value that is significantly smaller than 0.05 to reject the null hypothesis with the significance test at the usual 0.05 level. Comparison of Settings 3 and 4 (Table 1 and Figure 2) also indicates that the test power is mainly affected by the value of $\int \beta(z)dH(z)$ and is not sensitive to the shape of $\beta(z)$.

For real applications, it will be more common to consider a spline-based semi-parametric model. Following a referee's suggestion, we also considered the model $E\{\mathbb{N}(t)|Z\} = \Lambda(t)\exp\{\beta_1(Z_1)+\beta_2 Z_2+\beta_3 Z_3\}$, where $\beta_1(Z_1) = 0.5*Beta(Z_1, 2, 2)$, Z_1 and Z_2 are continuous variables generated from $Unif[0, 1]$ distribution , and Z_3 is a binary variable generated from $Bernoulli(0.5)$ distribution. For the simulated data, we fitted $\Lambda(t)$ and $\beta_1(Z_1)$ by B-splines and treated β_2 and β_3 as two unknown regression parameters to be estimated. We conducted the

Table 2: Simulation results. $\boldsymbol{\beta}^*$, $\beta_1(Z_1)$, β_2 or β_3; SP, sample size; True value, the exact value of $\int \beta(z)dH(z)$, β_2 and β_3; M-C SD, Monte Carlo standard deviation; ASE, average of bootstrap standard errors; **, probability of rejecting $H_0 : \beta_1(z_1) \equiv 0$, and coverage probability for β_2 and β_3.

Setting	$\boldsymbol{\beta}^*$	SP	True value	Bias	M-C SD	ASE	**
V	$\beta_1(Z_1) = 0.5 * Beta(Z_1, 2, 2)$	100	0·5	-0·026	0·136	0·121	0·908
		400		-0·010	0·084	0·077	1·000
	β_2	100	1·0	-0·001	0·057	0·058	0·940
		400		0·001	0·028	0·027	0·920
	β_3	100	0·5	0·001	0·033	0·033	0·943
		400		0·000	0·015	0·016	0·931

same simulation study as we did for Data Settings 1-4. Simulation results for this scenario are summarized in Table 2. The last column in Table 2 is the probability of rejecting H_0: $\beta_1(z_1) = 0$ for all z_1 at significance level 0.05, and the coverage probability of 95% confidence interval(CI) for β_2 and β_3, respectively. From Table 2 and Figure 4 (provided in the Supplementary Materials as Figure 1), it is clearly seen that the proposed methodology works well for a general splines-based semiparametric model. To facilitate the use of our method, we provide the computing code for the simulation studies online at *ftp://public.sjtu.edu.cn/*. Readers can access the code with account name *yuzhangsheng* and password *public*

5. Application

We applied the proposed method to the childhood wheezing study described in Yao et al. (2010). This is a study designed to evaluate interleukin function during infancy with the risk of asthma and wheezing symptoms for the children

with atopic dermatitis. One hundred and five infants were followed for an average of 5 years. Patients' baseline demographic information and one of interleukin functions Interleukin 5 denoted IL_5 were included as covariates for data analysis. The number of wheezing episodes was collected over time telephonically. Although the phone calls were scheduled to be made every month, the actual time for the phone call varied from month to month, and the information was not available every month as the research coordinator was not able to reach the patients. The number of the episodes since the last call was recorded, which was potentially greater than one. This data type ideally fits the framework of panel count data. The previous analysis by Yu et al. (2013) used a recurrent event model to study the recurrence of wheezing symptoms without considering the actual number of wheezing episodes between two consecutive calls. Therefore, it did not take the full advantage of the observed data collected in the study. To make better use of the collected data in studying the recurrence of wheezing, we conducted an analysis using a panel count data model and modelled a flexible non-linear effect of IL_5 on the wheezing recurrence using B-splines.

We analyzed the effect of IL_5 on the wheezing symptoms adjusted for the infant's age, sex, and mother's smoking status during pregnancy in the panel count data model. In this study, the mean age (month) at enrollment is 10.9, 53.7% are boys, and 8.04% of mothers smoked during pregnancy. The non-linear

effect of IL_5 was estimated using the proposed B-splines NPMPLE method and is depicted in Figure 3. It appears the effect of IL_5 on wheezing is more dramatic at the lower end of the IL_5 values and gradually stabilizes as IL_5 increases. The hypothesis testing procedure described in Section 4 yields a p-value of 0.005, indicating that IL_5 is indeed an influential factor for wheezing symptoms.

The multiple bumps shown in Figure 3 may present challenges in interpretation of the effect. However, this non-parametric approach is definitely informative and it provides evidence to suggest a non-linear covariate effect. To ascertain a potential non-linear effect, we also fitted the data using a piecewise linear regression model with the changing point at log(IL-5) =-5.88. The cutoff point for IL-5 was chosen based on the non-parametric estimate presented in Figure 3. The p-value for the test of slope difference between the two linear lines is < 0.001, strongly implying a non-linear effect of IL-5. The piecewise linear model has a meaningful interpretation of a piecewise proportional effect and reveals the similar information about the IL-5 effect compared to the proposed method. Virtually, the IL-5 effect on the underlying childhood wheezing process is much more pronounced when IL-5 is small.

6. Discussion and Conclusions

We propose a regression B-splines-based NPMPLE method for panel count data analysis. The proposed estimators for the baseline mean function and the

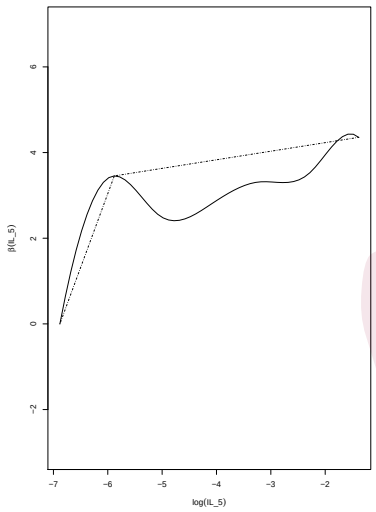

Figure 3: Estimation of the regression function. Solid line, B-splines model; dotted line, piecewise linear regression model.

regression function are consistent and converge to the corresponding true functions at the rate of $r/(1+2r)$. This convergence rate was shown to be optimal in spline-smoothing non-parametric estimation Speckman (1985); Zhang et al. (2010). Simulation studies show that the estimators have good finite sample properties. The proposed splines-based non-parametric functional analysis can be easily extended to the splines-based semi-parametric analysis with both continuous and discrete covariates included for analysis. The proposed non-parametric curve fitting is, to our knowledge, the first such method applied to panel count data analysis, and hence fills the gap of non-parametric regression for panel count data analysis.

The proposed splines-based regression method for non-parametric functional analysis requires predetermination of the number of B-spines basis functions, which increases as sample size increases. The validity of the asymptotic properties relies on the special placement of the interior knots to construct the basis functions, which can be somewhat subjective. Another approach is to use the penalized spline, in which the degree of smoothness of the estimated curve relies on the tuning parameter that can be selected using an objective approach, such as the cross-validation method. However, the asymptotic properties of penalized spline estimators have yet to be studied.

The proposed NPMPLE method has the advantage in numerical computation due to its likelihood simplicity, but the trade-off of this approach is the underestimation of the standard error for the estimated functions and their smooth functionals as shown in our simulation studies. This shortcoming is due to the fact that the proposed model does not fully account for the association among the count data. Although some standard error correction methods have yet to be developed, we believe that the complete Poisson model with the gamma frailty for B-splines estimation developed in Hua, Zhang, and Tu (2014) should be a reasonable approach with which to address the underestimation issue for the pseudolikelihood method.

The asymptotic normality theorem (Theorem 3) for a class of smooth func-

tionals not only facilitates a hypothesis testing method to test whether the covariate affects the underlying counting process, but is also potentially useful in model diagnosis for ascertaining if the covariate effect is linear. This task remains for further investigation.

Supplementary Materials

The Supplementary Materials include proofs of theorems and part simulation results.

Acknowledgements

The research is supported in part by National Natural Science Foundation of China 11671256(YU), China National Key Research and Development Program for Precision Medicine 2016YFC0902403(YU), China Scholarship Council(Zhao), and also National Natural Science Foundation of China 11271134 (Zhao).

References

De Boor, C. (2001). *A Practical Guide to Splines*. Springer-Verlag, New York.

Du, P., Jiang, Y. and Wang, Y. (2011). Smoothing spline ANOVA frailty model for recurrent event data.*Biometrics* **67**, 1330-1339.

Hu, XJ., Lagakos, SW., and Lockhart. RA. (1998). Marginal Analysis of Panel Counts through Estimating Functions. *Biometrika* **96**, 445-456.

Hua, L., Zhang, Y. and Tu, W. (2014) A spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *Canadian Journal of Statistics* **42**, 217-245.

Huang, C. Y., Qin, J., and Wang, M. C. (2010). Semiparametric analysis for recurrent event data with time-dependent covariates and informative censoring. *Biometrics* **66**, 39-49.

Lange, K.(1994) An adaptive barrier method for convex programming. *Methods and Applications of Analysis***1**, 392-402.

Lee, EW. and Kim, MY. (1998). The analysis of correlated panel data using a continuous-time Markov model. *Biometrics*. **54**, 1638-1644.

Lu, M., Zhang, Y. and Huang, J. (2007). Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* **94**, 705-718.

Lu, M., Zhang, Y. and Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *J. Amer. Statist. Assoc.* **104**, 1060-1070.

Nielsen, J. D. and Dean, C. B. (2008). Clustered mixed nonhomogeneous Poisson process spline models for the analysis of recurrent event panel data. *Biometrics* **64**, 751-761.

Nielsen, J. D. and Dean, C. B. (2008). Adaptive functional mixed NHPP models for the analysis of recurrent event panel data.*Computational Statistics and Data Analysis* **52**, 3670-3685.

Schumaker L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Shen, X. and Wong, WH. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

Speckman, P.(1985). Spline Smoothing and Optimal Rates of Convergence in Nonparametric

Regression Models. *Ann. Statist.* **13**, 970-983.

Sun, J. and Kalbfleisch, JD. (1995). Estimation of the mean function of point processes based on panel count data. *Statist. Sinica.* **5**, 279-290.

Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data.* Springer-Verlag, New York.

Sun, J. and Wei, LJ. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. Roy.Statist. Soc. B.* **62**, 293-302.

Sun, L., Zhu, L., and Sun, J. (2009). Regression analysis of multivariate recurrent event data with time-varying covariate effects. *J Multivariate Anal.* **100**, 2214-2223.

Thall, PF. and Lachin, JM. (1988) Analysis of recurrent events: Nonparametric methods for random-interval count data. *J. Amer. Statist. Assoc.* **83**, 339-347.

van der Vaart, AW. (2000). *Asymptotic Statistics.* Cambridge university press.

van der Vaart, AW. and Wellner, JA. (2000). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer-Verlag, New York.

Wellner, JA. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779-814.

Wellner, JA. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Ann. Statist.* **35**, 2106-2142.

Yao, W., Barb-Tuana, F.M., Llapur, C.J., Jones, M.H., Tiller, C., Kimmel, R., Kisling, J.,

Nguyen, E.T., Nguyen, J., Yu, Z. and Kaplan, M.H. (2010). Evaluation of airway reactivity and immune characteristics as risk factors for wheezing early in life. *Journal of Allergy and Clinical Immunology* **126**, 483-488.

Yu, Z., Liu, L., Bravata, DM., Williams, LS., Tepper, RS. (2013). A semiparametric recurrent events model with time-varying coefficients. *Statistics in Medicine*. **32**, 1016-1026.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* **89**, 39-48.

Zhang, Y., Hua, L., and Huang, J. (2010). A Spline-Based Semiparametric Maximum Likelihood Estimation Method for the Cox Model with Interval-Censored Data. *Scandinavian Journal of Statistics* **37**, 338-354.

Zhu, L., Zhao, H., Sun, J., Leisenring, W., and Robison, LL. (2015). Regression analysis of mixed recurrent-event and panel-count data with additive rate models. *Biometrics* **71**, 71-79.

School of Statistics, East China Normal University, Shanghai, China.

E-mail: huadong5359@126.com

Department of Biostatistics, Indiana University Fairbanks School of Public Health and School of Medicine, Indianapolis, Indiana 46202, USA

Department of Statistics, School of Mathematics Science, Shanghai Jiao Tong University, Shanghai 200240, China

E-mail: yz73@iu.edu

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

E-mail: xingqiu.zhao@polyu.edu.hk

Department of Statistics, School of Mathematics Science and SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail: yuzhangsheng@sjtu.edu.cn