A Least-Squares Approach to Consistent Information Estimation in Semiparametric Models

Jian Huang
Department of Statistics
University of Iowa

Ying Zhang and Lei Hua
Department of Biostatistics
University of Iowa

May 5, 2008

Abstract

A method is proposed for consistent information estimation in a class of semiparametric models. The method is based on the geometric interpretation of the efficient score function, that it is the residual of the orthogonal projection of the score function for the finite-dimensional parameter onto the tangent space for the infinite-dimensional parameter. The empirical version of this projection is a least-squares nonparametric regression problem. Under appropriate conditions, the sum of squared residuals of this regression is shown to be a consistent estimator of the efficient Fisher information and is actually the observed information for a class of sieve maximum likelihood estimators. Simulations studies are conducted to evaluate finite sample performance of the estimator in two illustrating examples: Poisson proportional mean model for panel count data and Cox model for interval-censored data. Finally the method is applied to two real-life examples: bladder tumor study and breast cosmesis study.

Key words and phrases. counting process, Cox model, empirical processes, information, interval censoring, maximum (profile) likelihood, monotone polynomial splines, panel count data, proportional hazards model, semiparametric model

1. Introduction

In a regular parametric model, the maximum likelihood estimator (MLE) is asymptotically normal with variance equal to the inverse of the Fisher information, and the Fisher information can be estimated by the observed information. This result provides large sample justification for the use of normal approximation to the distribution of MLE. An important factor making this approximation useful in statistical inference is that the observed information can be readily computed and is consistent. In many situations, consistency of the observed information follows directly from the law of large numbers and consistency of MLE.

Asymptotic normality of MLE of the regular parameters continues to hold in many semiparametric and nonparametric models. See for example, Chen (1988, 1995), Chang (1990), Geskus and Groeneboom (1996), Gill (1989), Groeneboom (1996), Groeneboom and Wellner (1992), Gu and Zhang (1993), Murphy (1995), Murphy, Rossini and van der Vaart (1997), Qin and Lawless (1993), Severini and Wong (1991), van der Laan (1993), van der Vaart (1994, 1996), Wong and Severini (1991), Huang (1996), Huang and Rossini (1997) and Wellner and Zhang (2007). In all these examples, the MLE or a smooth functional of the MLE is asymptotically normal with variance equal to the inverse of the efficient Fisher information. The asymptotic normality and efficiency results provide insight to the theoretical properties of maximum likelihood estimators. Unfortunately, in many semiparametric models studied in the aforementioned articles, the efficient Fisher information is either very complicated or may not have an explicit expression and they often can not be estimated directly as in the case of parametric MLE. Thus effort is required to estimate the asymptotic variance in order to apply these results to statistical inference.

In this paper, we consider consistent information estimation in a class of semiparametric

models that are parametrized in terms of a finite-dimensional parameter θ and a parameter ϕ whose dimension increases with sample size. Hence ϕ is often called an infinite-dimensional parameter. Two important examples are Cox's (1972) proportional hazards model for interval-censored data studied by Huang and Wellner (1995) and the proportional mean model for panel count data proposed by Sun and Wei (2000) and Wellner and Zhang (2007). In these two examples, θ is the finite-dimensional regression coefficient, ϕ is the log of baseline hazard function or the log of baseline mean function.

At least two methods used in parametric models for estimating the variance have been suggested in semiparametric models. The first method is to use the inverse observed information based on the likelihood, correcting for the presence of the infinite-dimensional parameter ϕ. However, when ϕ cannot be estimated at the usual root-n rate, consistency of this estimator has not been proved in general.

The second method is to use the second derivative of the *profile likelihood* of θ at the maximum likelihood estimate. For a fixed value of θ, the profile likelihood is the maximum of the likelihood with respect to ϕ. Because the profile likelihood often can only be computed numerically, discretized versions of its second derivative are proposed by Nielson, Gill, Andersen and Sorensen (1992), and used by Huang and Wellner (1995) and Murphy and van der Vaart (1996). Using an ingenious sandwich inequality, Murphy and van der Vaart (2000) showed that the discretized versions of the second derivative of profile likelihood provides consistent variance estimator in a large class of semiparametric models. Murphy and van der Vaart (1999) also proved that the profile likelihood resembles the ordinary likelihood in many essential aspects. When the dimension of θ is one or two, it is relatively easy to compute and visualize the profile likelihood. For moderate to high dimensional θ, implementation of the profile likelihood approach may be computationally demanding and

sometimes difficult.

In a general semiparametric maximum likelihood estimation problem, the joint estimation of (θ, ϕ) is often a quite challenging problem numerically. Sieve semiparametric M-estimation using polynomial splines proposed by Lu, Zhang and Huang (2008) for panel count data is shown numerically efficient. However, the estimation of standard error for the M-estimator of regression parameter still remains a big task and alternatively, they used the bootstrap standard error by taking the computation advantage in spline-based sieve M-estimation.

In this article, we propose a least-squares approach to consistent estimation of the information matrix of the semiparametric maximum likelihood estimator of θ. The proposed method is based on the geometric interpretation of the efficient score function, that it is the residual of the projection of the score function for θ onto the tangent space for ϕ (van der Vaart, 1991, Bickel, Klaassen, Ritov and Wellner, 1993, Chapter 3). Thus the theoretical information calculation is a least-squares problem in a Hilbert space. When a sample from the model is available, this theoretical information can be estimated by its empirical version. It turns out that this empirical version is essentially a least-squares nonparametric regression problem, due to the fact that the score function for the infinite-dimensional parameter is a linear operator. In this nonparametric regression problem, the "response" is the score function for the finite-dimensional parameter θ, the "covariate" is the linear score operator for the infinite-dimensional parameter ϕ, and the "regression parameter" is the *least favorable direction* which is used to define the efficient score. Computationally, the proposed method can be implemented with a least-squares nonparametric regression fitting program.

In a class of sieve MLE's using a linear approximation space, the proposed estimator of information matrix is shown to be the same as the inverse of the observed information matrix for the sieve MLE. This equivalence is useful from both the theoretical and the computational

point of view. First, this equivalence enables a simple and indirect consistency proof of the observed information matrix in the semiparametric setting. Second, it provides two ways of computing the observed information matrix: one can either directly compute the observed information matrix or fit a least-squares nonparametric regression. Because of its numerical convenience and good theoretical properties, the class of sieve MLE's using polynomial splines is utilized in our numerical demonstration and is recommended for applications of general semiparametric estimation.

The paper is organized as follows. Section 2 describes the motivation and the least-squares approach. Section 3 specializes the general approach to a class of sieve MLE's. Section 4 applies the proposed method along with the spline-based sieve MLE to two models, semiparametric Poisson mean model for panel count data and Cox proportional hazards model for interval censored data, studied in Wellner and Zhang (2007) and Huang and Wellner (1995), respectively. Section 5 renders numerical results via simulations and applications in real-life examples for the models discussed in Section 4. Section 6 concludes with some discussions. Some technical details are included in appendices.

2. The Least-Squares Approach

Let $X_1, \ldots, X_n$ be independent random variables with a common probability measure $P_{\theta,\phi}$, where $(\theta, \phi) \in \Theta \times \Phi$. Here Θ is a subset of R^d and Φ is a general space. Assume that $P_{\theta,\phi}$ has a density $p(\cdot; \theta, \phi)$ with respect to a σ-finite measure. Denote $\tau = (\theta, \phi)$ and let $\tau_0 = (\theta_0, \phi_0) \in \Theta \times \Phi$ be the true parameter value under which the data are generated. The maximum likelihood estimator of τ_0 is the value $\hat{\tau}_n \equiv (\hat{\theta}_n, \hat{\phi}_n)$ that maximizes the

log-likelihood

$$l_n(\tau) = \sum_{i=1}^{n} \log p(X_i; \theta, \phi)$$

over the parameter space $\mathcal{T} \equiv \Theta \times \Phi$. Let $\|\cdot\|$ be the Euclidean distance of R^d, and $\|\cdot\|_\Phi$ be an appropriate norm defined on Φ. We assume it has been shown that

$$\|\hat{\tau}_n - \tau_0\|_{\mathcal{T}} \equiv \left\{ \|\hat{\theta}_n - \theta_0\|^2 + \|\hat{\phi}_n - \phi_0\|_\Phi^2 \right\}^{1/2} = O_p(r_n^{-1}), \tag{2.1}$$

where r_n is a sequence of numbers converging to infinity. Consistency and rate of convergence in nonparametric and semiparametric models have been addressed by many authors, see for example, Birg*é* and Massart (1993), van der Geer (1993), Shen and Wong (1995), and van der Vaart and Wellner (1996). The results and methods developed by these authors can often be used to verify (2.1).

The motivation to study consistent information estimation is the following. In many semiparametric models, in addition to (2.1), it can be shown that

$$n^{1/2}\left(\hat{\theta}_n - \theta_0\right) \to_d N\left(0, I^{-1}(\theta_0)\right), \tag{2.2}$$

where $I(\theta_0)$ is the efficient Fisher information for θ_0, adjusting for the presence of nuisance parameter ϕ. The definition of $I(\theta_0)$ is given below. This holds for models cited in the previous section and for the examples in Section 4. Thus estimation of the asymptotic variance of $\hat{\theta}_n$ is equivalent to estimation of $I(\theta_0)$ provided $I(\theta_0)$ is nonsingular. Of course, for the problem of estimating $I(\theta_0)$ to be meaningful, we need to first establish (2.2).

The calculation of $I(\theta_0)$ and its central role in asymptotic efficiency theory for semiparametric models have been systematically studied by Begun, Hall, Huang and Wellner

(1983), van der Vaart (1991), and BKRW (1993) and the references therein. In the following, We first briefly describe how the information $I(\theta_0)$ is defined in a general semiparametric MLE setting in order to motivate the proposed information estimator.

Let $l(\theta, \phi; x) = \log p(x; \theta, \phi)$ be the log-likelihood for a sample of size one. Consider a parametric smooth submodel with parameter $(\theta, \phi_{(s)})$, where $\phi_{(0)} = \phi$ and

$$\left.\frac{\partial \phi_{(s)}}{\partial s}\right|_{s=0} = h.$$

Let $\mathcal{H}$ be the class of functions h defined by this equation. Usually, $\mathcal{H}$ is a Hilbert space. The score operator for ϕ is

$$\dot{l}_2(\tau; x)(h) = \left.\frac{\partial}{\partial s} l(\theta, \phi_{(s)}; x)\right|_{s=0}. \tag{2.3}$$

Observe that $\dot{l}_2$ is a linear operator mapping $\mathcal{H}$ to $L_2(P_{\theta,\phi})$. So for constants c_1, c_2 and $h_1, h_2 \in \mathcal{H}$,

$$\dot{l}_2(\tau; x)(c_1 h_1 + c_2 h_2) = c_1 \dot{l}_2(\tau; x)(h_1) + c_2 \dot{l}_2(\tau; x)(h_2). \tag{2.4}$$

The linearity of $\dot{l}_2$ is crucial to the proposed method. For a d-dimensional θ, $\dot{l}_1(\tau; x)$ is the vector of partial derivatives of $l(\tau; x)$ with respect to θ. For each component of $\dot{l}_1$, a score operator for ϕ is defined as in (2.3). So the score operator for ϕ corresponding to $\dot{l}_1$ is defined as

$$\dot{l}_2(\tau; x)(\mathbf{h}) \equiv (\dot{l}_2(\tau; x)(h_1), \ldots, \dot{l}_2(\tau; x)(h_d))', \tag{2.5}$$

where $\mathbf{h} \equiv (h_1, \ldots, h_d)'$ with $h_k \in \mathcal{H}, 1 \leq k \leq d$.

Let $\dot{\mathcal{P}}_2$ be the closed linear span of $\{\dot{l}_2(h) : h \in \mathcal{H}\}$. Then the efficient score function for the kth component of θ is $\dot{l}_{1,k} - \Pi(\dot{l}_{1,k}|\dot{\mathcal{P}}_2)$, where $\dot{l}_{1,k}$ is the kth component of $\dot{l}_1(\tau; x)$ and $\Pi(\dot{l}_{1,k}|\dot{\mathcal{P}}_2)$ is the projection of $\dot{l}_{1,k}$ onto $\dot{\mathcal{P}}_2$. Equivalently, $\Pi(\dot{l}_{1,k}|\dot{\mathcal{P}}_2)$ is the minimizer of the squared residual $P[\dot{l}_{1,k}(\tau_0; X) - \eta]^2$ over $\eta \in \dot{\mathcal{P}}_2$. See for example, van der Vaart (1991), Section 6, and BKRW (1993), Theorem 1, page 70. In general, η may not be a score function for ϕ, that is, there may not exist a $h \in \mathcal{H}$ such that $\eta = \dot{l}_2(h)$. However, in models with regularity conditions, η either is a score function or can be approximated by a score function. So we assume that the efficient score vector for θ is $\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(\xi_0)$, where ξ_0 is an element of $\mathcal{H}^d$ that minimizes

$$\rho(\mathbf{h}) \equiv E\|\dot{l}_1(\tau; X) - \dot{l}_2(\tau; X)(\mathbf{h})\|^2 \tag{2.6}$$

over $\mathcal{H}^d$. The minimizer $\xi_0 = (\xi_{01}, \xi_{02}, \ldots, \xi_{0d})'$ is called the *least favorable direction*. Denote the efficient score by $l^*(\tau; x) \equiv \dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(\xi_0)$. Then the information for θ is

$$I(\theta) = E\|l^*(\tau; X)\|^2 = E\|\dot{l}_1(\tau; X) - \dot{l}_2(\tau; X)(\xi_0)\|^2. \tag{2.7}$$

Therefore, to estimate $I(\theta)$, it is natural to consider minimizing an empirical version of (2.6). In particular, with the random sample $X_1, \ldots, X_n$ and the consistent estimator $\widehat{\tau}_n$, we can estimate $I(\theta)$ by the minimum value of

$$\rho_n(\mathbf{h}) \equiv n^{-1} \sum_{i=1}^{n} \|\dot{l}_1(\widehat{\tau}_n; X_i) - \dot{l}_2(\widehat{\tau}_n; X_i)(\mathbf{h})\|^2 \tag{2.8}$$

over $\mathcal{H}^d$. That is, if $\widehat{\xi}_n$ is a minimizer of ρ_n over $\mathcal{H}^d$, then a natural estimator of $I(\theta_0)$

is $\widehat{\mathcal{I}}_n \equiv \rho_n(\widehat{\xi}_n)$. Because $\dot{l}_2$ is a linear operator, this minimization problem is essentially a least-squares nonparametric regression problem and it can be solved for each component separately. In the next section, we show that, in a class of sieve MLE's, the minimum value $\rho_n(\widehat{\xi}_n)$ is actually the observed information based on the outer product of the first derivatives of the log-likelihood.

In the following, we state a simple proposition which provides sufficient conditions ensuring consistency of the estimated least favorable direction and $\widehat{\mathcal{I}}_n$ as an estimator of $I(\theta_0)$. This proposition appears to be useful in a large class of models and is easy to apply.

As in nonparametric regression, if the space $\mathcal{H}$ is too large, minimization over this space may not yield consistent estimators of ξ_0 and $I(\theta_0)$. We can use an approximation space $\mathcal{H}_n$ (a sieve) which is smaller than $\mathcal{H}$ and converges to $\mathcal{H}$ as n tends to infinity. Under appropriate conditions, minimization of ρ_n over $\mathcal{H}_n$ will yield a consistent estimator of ξ_0. On the other hand, if minimizing over $\mathcal{H}$ does yield consistent estimators, then $\mathcal{H}_n$ can be taken to be $\mathcal{H}$.

For simplicity and without loss of generality, it suffices to consider the case of one-dimensional θ. In the following and throughout, we will use the linear functional notations for integrals. So for any probability measure Q, $Qf = \int f dQ$ as long as the integral is well defined. Below, $P = P_{\theta_0,\phi_0}$. $\mathbb{P}_n$ is the empirical measure of $X_i, 1 \leq i \leq n$. So $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$.

PROPOSITION 2.1. *Denote* $\widehat{\xi}_n = argmin_{\mathcal{H}_n}\rho_n(h)$ *and* $\ell(\tau, h; x) = [\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(h)]^2$.

If the class of functions $\Im = \{\ell(x, \tau, h) : \tau \in \mathcal{T}, h \in \mathcal{H}\}$ *is Glivenko-Cantelli and* $\widehat{\tau}_n \to_p \tau_0$, *then*

$$\rho_n(\widehat{\xi}_n) \to_p I(\theta_0).$$

The proof is given is Appendix A.

3. Observed Information in Sieve MLE

We now apply the method described in Section 2 to a class of sieve MLE's. We show that, if the parameter space Φ and the space $\mathcal{H}$ can be approximated by a common approximation space, and if this space has a basis, then the least-squares calculation in Section 2 yields the *observed information matrix*. In other words, computation of the observed information matrix is equivalent to solving the least-squares problem of Section 2. So there is no need to actually carry out the least-squares computation when the observed information can be computed as in the ordinary setting of parametric estimation. This is computationally convenient, because the observed information matrix is based on either the first derivatives or the second derivatives of the log-likelihood function and these derivatives are often already computed in a numerical algorithm for computing the MLE of unknown regression parameters.

On the other hand, for problems in which direct computation of the observed information matrix is difficult, one can instead solve the least-squares nonparametric regression problem to obtain the observed information matrix. These nonparametric regression problems can be solved by using standard least-squares fitting programs for linear regression.

As in finite-dimensional parametric models, some regularity conditions are required for the MLE $\widehat{\theta}_n$ to be root-n consistent and asymptotically normal. These regularity conditions usually include certain smoothness assumptions on the infinite-dimensional parameter ϕ and the underlying probability model. Consequently, the least favorable direction will be a smooth function such as a bounded Lipschitz function. Then we can take $\mathcal{H}$ to be the class of such smooth functions. From the approximation theory, many spaces designed for efficient computation can be used to approximate any element in $\mathcal{H}$ with arbitrary precision under appropriately defined distance. For example, we may use the space of polynomial spline

functions (Schumaker, 1981). This class not only has good approximation power, but is also computationally convenient. We will use this approximation space in the next section.

Let Φ_n be an approximation space for both Φ and $\mathcal{H}$. Suppose it has a set of basis functions $\mathcal{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_{q_n})'$, such that every $\phi \in \Phi_n$ can be represented as $\phi = \sum_{j=1}^{q_n} \beta_j \mathbf{b}_j \equiv \mathcal{B}_n'\beta$, where $\beta = (\beta_1, \ldots, \beta_{q_n})' \in B_n \subset R^{q_n}$ is a vector of real numbers. So every $\phi \in \Phi_n$ can be identified with a vector $\beta \in B_n$. Here the dimension q_n is a positive integer depending on sample size n. To ensure consistency of $\widehat{\tau}_n$, we need $q_n \to \infty$ as $n \to \infty$. In general, for $\widehat{\theta}_n$ to be asymptotically normal, we need to control the growth rate of q_n appropriately.

The sieve MLE of $\tau_0 = (\theta_0, \phi_0)$ is defined to be the $(\widehat{\theta}_n, \widehat{\phi}_n)$ that maximizes the log-likelihood $l_n(\theta, \phi)$ over $\Theta \times \Phi_n$. Equivalently, one can find $(\widehat{\theta}_n, \widehat{\beta}_n)$ that maximizes $l_n(\theta, \mathcal{B}_n'\beta)$ over $\Theta \times B_n$. Then $\widehat{\phi}_n = \mathcal{B}_n'\widehat{\beta}_n$.

Now consider estimation of $I(\theta_0)$. First we introduce some convenient notations. Denote

$$\dot{l}_2(\widehat{\tau}_n; x)(\mathcal{B}_n) = (\dot{l}_2(\widehat{\tau}_n; x)(\mathbf{b}_1), \ldots, \dot{l}_2(\widehat{\tau}_n; x)(\mathbf{b}_{q_n}))^T,$$

and

$$A_{11} = \mathbb{P}_n \left(\dot{l}_1(\widehat{\tau}_n; X)\right)^{\otimes 2}, \quad A_{12} = \mathbb{P}_n \dot{l}_1(\widehat{\tau}_n; X) \dot{l}_2^T(\widehat{\tau}_n; X)(\mathcal{B}_n),$$

$$A_{21} = A_{12}^T, \quad A_{22} = \mathbb{P}_n \left(\dot{l}_2(\widehat{\tau}_n; X)(\mathcal{B}_n)\right)^{\otimes 2},$$

where $a^{\otimes 2} = aa^T$. The outer product version of the observed information for θ is given by

$$\widehat{\mathcal{O}}_n = A_{11} - A_{12} A_{22}^- A_{21}. \tag{3.1}$$

Here for any matrix A, A^- denotes its generalized inverse. Although A_{22}^- may not be unique, $A_{12} A_{22}^- A_{21}$ is unique by the results on the generalized inverse, see for example,

Rao (1973), Chapter 1, result 1b.5 (vii), page 26. The use of the generalized inverse in (3.1) is for generality of these formulas. When the nonparametric component ϕ is a smooth function, then for any fixed sample size, the sieve MLE is obtained over a finite-dimensional approximation space whose (theoretical) dimension is $O(n^\nu)$, where ν is typically less than $1/2$, A_{22} is usually invertible.

We now show that $\widehat{\mathcal{O}}_n$ equals $\widehat{\mathcal{I}}_n$, which is the minimum value of

$$\rho_n(\mathbf{h}) = \mathbb{P}_n \|\dot{l}_1(\widehat{\tau}_n; X) - \dot{l}_2(\widehat{\tau}_n; X)(\mathbf{h})\|^2$$

over $\mathbf{h} \in \Phi_n$ for $\mathbf{h} = (h_1, h_2, \cdots, h_d)'$. Write $h_j = \mathcal{B}_n' c_j$ for $j = 1, 2, \cdots, d$. Let $Y_{i,j} = \dot{l}_{1,j}(\widehat{\tau}_n; X_i)$, the jth component of $Y_i = \dot{l}_1(\widehat{\tau}_n; X_i)$ and $Z_i = \dot{l}_2(\widehat{\tau}_n; X_i)(\mathcal{B}_n)$. This minimization problem becomes a least-squares problem of finding $\widehat{c}_n = (\widehat{c}_{n,1}', \widehat{c}_{n,2}', \cdots, \widehat{c}_{n,d}')'$ that minimizes

$$\mathbb{P}_n \left[\sum_{j=1}^{d} (Y_{i,j} - Z_i' c_j)^2 \right].$$

By standard least-squares calculation,

$$\widehat{c}_{n,j} = \left(\sum_{i=1}^{n} Z_i Z_i' \right)^{-} \left(\sum_{i=1}^{n} Z_i Y_{i,j} \right) \quad for \quad j = 1, 2, \cdots, n.$$

Hence, we have

$$\begin{aligned}
\widehat{\mathcal{I}}_n &= \rho_n(\hat{\xi}_n) = \mathbb{P}_n \left[\dot{l}_1(\widehat{\tau}_n; X) - \dot{l}_2(\widehat{\tau}_n; X)(\widehat{\xi}_n) \right]^{\otimes 2} \\
&= \mathbb{P}_n \left[\dot{l}_1(\widehat{\tau}_n; X) - (\widehat{c}_{n,1}, \widehat{c}_{n,2}, \cdot, \widehat{c}_{n,d})' \, \dot{l}_2(\widehat{\tau}_n; X)(\mathcal{B}_n) \right]^{\otimes 2} \\
&= \mathbb{P}_n \left[\dot{l}_1(\widehat{\tau}_n; X) - \left(\sum_{i=1}^{n} Y_i Z_i' \right) \left(\sum_{i=1}^{n} Z_i Z_i' \right) \dot{l}_2(\widehat{\tau}_n; X)(\mathcal{B}_n) \right]^{\otimes 2}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}_n \left[\dot{l}_1(\widehat{\tau}_n; X) - A_{12} A_{22}^{-} \dot{l}_2(\widehat{\tau}_n; X)(\mathcal{B}_n) \right]^{\otimes 2} \\
&= A_{11} - A_{12} A_{22}^{-} A_{21} - A_{12} A_{22}^{-} A_{21} + A_{12} A_{22}^{-} A_{22} A_{22}^{-} A_{21} \\
&= A_{11} - A_{12} A_{22}^{-} A_{21} = \widehat{\mathcal{O}}_n
\end{aligned}
$$

We summarize the above calculation in the following proposition.

PROPOSITION 3.1. *Assume that there exist* $\beta_n^* \in B_n$ *and* $c_{n,j}^* \in R^{q_n}$ *for* $j = 1, 2, \cdots, d$ *such that*

$$
\|\mathcal{B}_n' \beta_n^* - \phi_0\|_\Phi = O(k_{1n}^{-1}), \quad \textit{and} \quad \|\mathcal{B}_n' c_{n,j}^* - \xi_{0,j}\|_\Phi = O(k_{2n}^{-1}), \quad j = 1, 2 \cdots, d, \tag{3.2}
$$

where k_{1n} *and* k_{2n} *are two sequences of numbers satisfying* $k_{1n} \to \infty$ *and* $k_{2n} \to \infty$ *as* $n \to \infty$. *Suppose that the conditions of Proposition 2.1 are satisfied. Then the observed information matrix* $\widehat{\mathcal{O}}_n$ *defined in (3.1) is a consistent estimator of* $I(\theta_0)$.

4. Examples

In this section, we illustrate the proposed method in two semiparametric regression models, including Poisson proportional mean model for panel count data studied in Wellner and Zhang (2007) and Cox model (1972) for interval-censored data studied in Huang and Wellner (1995). In these examples, the parameter space Φ will be the space of smooth functions defined below. The sieve Φ_n is the space of polynomial splines. The polynomial splines have been used in many fully nonparametric regression models, see for example, Stone (1985, 1986).

Let $a = d_0 < d_1 < \cdots < d_{K_n} < d_{K_n+1} = b$ be a partition of $[a, b]$ into K_n subintervals $I_{Kt} = [d_t, d_{t+1}), t = 0, \ldots, K-1$ and $I_{KK} = [d_K, d_{K+1}]$, where $K \equiv K_n \approx n^v$ is a positive integer such that $\max_{1 \le k \le K+1} |d_k - d_{k-1}| = O(n^{-v})$. Denote the set of partition points by $D_n = \{d_1, \ldots, d_{K_n}\}$. Let $\mathcal{S}_n(D_n, K_n, m)$ be the space of polynomial splines of order $m \ge 1$ consisting of functions s satisfying: (i) the restriction of s to I_{Kt} is a polynomial of order m for $m \le K$; (ii) for $m \ge 2$ and $0 \le m' \le m-2$, s is m' times continuously differentiable on $[a, b]$. This definition is phrased after Stone (1985), which is a descriptive version of Schumaker (1981), page 108, Definition 4.1. According to Schumaker (1981), page 117, Corollary 4.10, there exists a *local* basis $\mathcal{B}_n \equiv \{\mathbf{b}_t, 1 \le t \le q_n\}$, so called B-splines, for $\mathcal{S}_n(D_n, K_n, m)$, where $q_n \equiv K_n + m$. These basis functions are nonnegative and sum up to one at each point in $[a, b]$, and each $\mathbf{b}_t$ is zero outside the interval $[d_t, d_{t+m}]$.

4.1. Poisson Proportional Mean Model for Panel Count Data

Let $\{\mathbb{N}(t) : t \ge 0\}$ be a univariate counting process. K is the total number of observations on the counting process and $\underline{T} = (T_{K,1}, \cdots, T_{K,K})$ is a sequence of random observation times with $0 < T_{K,1} < \cdots < T_{K,K}$. The counting process is only observed at those times with the cumulative events denoted by $\underline{\mathbb{N}} = \{\mathbb{N}(T_{K,1}), \cdots, \mathbb{N}(T_{K,K})\}$. This type of data is referred to to panel count data by Sun and Kalbfleish (1995). In this manuscript, we assume that $(K, \underline{T})$ is conditionally independent of $\underline{\mathbb{N}}$ given a vector of covariates Z and we denote the observed data consisting of independent and identically distributed $X_1, \cdots, X_n$, where $X_i = (K_i, \underline{T}^{(i)}, \underline{\mathbb{N}}^{(i)}, Z_i)$ with $\underline{T}^{(i)} = (T^{(i)}_{K_i,1}, \cdots, T^{(i)}_{K_i,K_i})$ and $\underline{\mathbb{N}}^{(i)} = (\mathbb{N}^{(i)}(T^{(i)}_{K_i,1}), \cdots, \mathbb{N}^{(i)}(T^{(i)}_{K_i,K_i}))$, for $i = 1, \cdots, n$.

Panel count data are often seen in clinical trials, social demographic and industrial reliability studies. Sun and Wei (2000) and Zhang (2002) proposed the proportional mean

model

$$\Lambda(t|Z) = \Lambda_0(t)\exp(\theta_0' Z) \tag{4.1}$$

to analyze panel count data semiparametrically, where $\Lambda(t|Z) = E(\mathbb{N}(t)|Z)$ is the expected cumulative events observed at time t, conditional on Z with the true baseline mean function given by $\Lambda_0(t)$. Wellner and Zhang (2007) proposed a nonhomogeneous Poisson process with the conditional mean function given by (4.1) to study the MLE of $\tau_0 = (\theta_0, \Lambda_0(t))$. The log likelihood for $(\theta, \Lambda(t))$ under the Poisson proportional mean model is given by

$$l_n(\theta, \Lambda) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\left[\Delta\mathbb{N}_{K_i,j}^{(i)}\log\Delta\Lambda_{K_i,j} + \Delta\mathbb{N}_{K_i,j}^{(i)}\theta' Z_i - \exp(\theta' Z_i)\Delta\Lambda_{K_i,j}\right],$$

where

$$\Delta\mathbb{N}_{K_i,j}^{(i)} = \mathbb{N}^{(i)}(T_{K_i,j}^{(i)}) - \mathbb{N}^{(i)}(T_{K_i,j-1}^{(i)})$$

and

$$\Delta\Lambda_{K_i,j} = \Lambda(T_{K_i,j}^{(i)}) - \Lambda(T_{K_i,j-1}^{(i)}),$$

for $j = 1, 2, \cdots, K$.

To study the asymptotic properties of the MLE, Wellner and Zhang (2007) defined the following L_2-norm,

$$d(\tau_1, \tau_2) = \left\{|\theta_1 - \theta_2|^2 + \int|\Lambda_1(t) - \Lambda_2(t)|^2 d\mu_1(t)\right\}^{1/2}$$

with

$$\mu_1(t) = \int_{R^d}\sum_{k=1}^{\infty}P(K = k|Z = z)\sum_{j=1}^{k}P(T_{K,j} \le t|K = k, Z = z)dF(z).$$

They showed that the semiparametric MLE, $\widehat{\tau}_n = (\hat{\theta}_n, \hat{\Lambda}_n)$ converges to the true parameters $\tau_0 = (\theta_0, \Lambda_0)$ (under some mild regularity conditions) in a rate lower than $n^{1/2}$, i.e. $d(\widehat{\tau}_n, \tau_0) = O_p(n^{-1/3})$, however, the MLE of θ_0 is still semiparametric efficient, that is

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N\left(0, I^{-1}(\theta_0)\right)$$

with the Fisher information matrix given by

$$I(\theta_0) = E\left\{\Delta\Lambda_0(T_{K,j})e^{\theta_0' Z}\left[Z - \frac{E(Ze^{\theta_0' Z}|K, T_{K,j-1}, T_{K,j})}{E(e^{\theta_0' Z}|K, T_{K,j-1}, T_{K,j})}\right]^{\otimes}\right\}.$$

The computation of the semiparametric MLE is very time consuming as it requires the joint estimation of θ_0 and the infinite dimensional parameter Λ_0. Although the Fisher information has a nice explicit form, there is no easy method available to calculate the observed information.

Lu (2007) studied the sieve MLE for the above semiparametric Poisson model using monotone polynomial splines. Let $\mathcal{B}_n = (\mathbf{b}_1, \ldots, \mathbf{b}_{q_n})$ be the basis of B-splines defined earlier. The monotone polynomial spline space is defined to be

$$\mathcal{M}_n(D_n, K_n, m) = \left\{\phi_n : \phi_n(t) = \sum_{j=1}^{q_n} \beta_j \mathbf{b}_j(t) \in \mathcal{S}_n(D_n, K_n, m),\ \beta \in B_n,\ \ t \in [a, b]\right\}.$$

where $B_n = \{\beta : \beta_1 \leq \beta_2 \leq \cdots \leq \beta_{q_n}\}$. Each element of $\mathcal{M}_n(D_n, K_n, m)$ is a nondecreasing function because of the monotonicity constraints on $\beta_1, \ldots, \beta_{q_n}$. This fact is a consequence of the *variation diminishing properties* of B-splines. See for instance, Schumaker (1981), Example 4.75 and Theorem 4.76, pages 177-178. Replaying $\Lambda(t)$ by the $\exp(\sum_{l=1}^{q_n} \beta_l \mathbf{b}_j(t))$ in the likelihood above, Lu (2007) solved the constraint optimization problem over the space

$\Theta \times B_n$. It turns out that, compared to Wellner and Zhang's method, the B-splines sieve MLE is much less computational demanding (Lu, Zhang and Huang, 2008) with a better overall convergence rate. In addition, the sieve MLE of θ is still semiparametric efficient with the same Fisher information matrix, $I(\theta_0)$. The bootstrap procedure was implemented to estimate $I(\theta_0)$ consistently by Lu, Huang and Zhang (2008) due to the computation advantage of the B-splines sieve MLE method.

A straightforward algebra leads that

$$\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(h) = \sum_{j=1}^{K} \left(\Delta \mathbb{N}_{K,j} - \exp(\beta^T Z)\Delta\Lambda_{Kj}\right)\left(Z - \frac{\Delta h_{Kj}}{\Delta\Lambda_{Kj}}\right)$$

Under the regularity conditions given in Wellner and Zhang (2007), following the empirical process arguments used in Lu (2007) for monotone B-splines estimation, it can be easily shown that the class $\{\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(h) : \tau \in \mathcal{T}, \ \ h \in \mathcal{H}\}$ is Glivenko-Canteli, by showing the bracketing number with $L_1(P)$ norm of this class is bounded. This will imply that $\{(\dot{l}_1(\tau; x) - \dot{l}_2(\tau; x)(h))^2 : \tau \in \mathcal{T}, \ \ h \in \mathcal{H}\}$ is Glivenko-Canteli as well, based on the given regularity conditions. Moreover, using monotone cubic B-splines in the sieve estimation automatically gives

$$\|\mathcal{B}_n'\beta_n^* - \phi_0\|_\Phi = O(n^{-p\nu}) \ \text{ and } \ \|\mathcal{B}_n' c_n^* - \xi_0\|_\Phi = O(n^{-p\nu})$$

due to Corollary 6.20 of Schumaker (1981). Hence the Fisher information $I(\theta_0)$ in semiparametric proportional mean model for panel count data can be consistently estimated using the proposed least-squares approach.

4.2. Cox Model for Interval-Censored Data

Interval-censored data occur very frequently in long-term follow-up study for an event time of interest. The exact event time T is not observable; it is only known with certainty that T is bracketed between two adjacent examination times, or occurs before the first or after the last follow-up examination. Let (L, R) be the pair of examination times bracketing the event time T. That is, L is the last examination time before and R is the first examination time after the event. If $0 < L < R < \infty$, then T is interval-censored. If the event occurs before the first examination, then T is left-censored. If the event has not occurred after last examination, then T is right-censored. Such data is called "case 2" interval-censored data. Nonparametric estimation of a distribution function and its smooth functionals with interval-censored data have been studied by Groeneboom and Wellner (1992), Groeneboom (1996), and Geskus and Groeneboom (1996).

In this example, we consider the B-splines sieve MLE of the Cox proportional hazards model for the interval-censored data. With the proportional hazards model, the conditional hazard of T given a covariate vector $Z \in R^d$ is proportional to the baseline hazard. In terms of cumulative hazard functions, this model is

$$\Lambda(t|z) = \Lambda_0(t)e^{\theta_0' z}, \tag{4.2}$$

where θ_0 is a d-dimensional regression parameter and Λ_0 is the unspecified baseline cumulative hazard function.

Denote the two censoring variables by U and V, where $P(U \leq V) = 1$. Let G be the joint distribution function of (U, V). Let $\delta_1 = 1_{[T \leq U]}, \delta_2 = 1_{[U < T \leq V]}$ and $\delta_3 = 1 - \delta_1 - \delta_2$. Assume that conditional on Z, T is independent of (U, V), and that the joint distribution of

(U, V) and Z does not depend on the parameters of interest. Then the density function of $X \equiv (\delta_1, \delta_2, U, V, Z)$ with respect to the product of the counting measure on $\{0,1\} \times \{0,1\}$ and the probability measure induced by the distribution of (Z, U, V), is

$$p(x;\theta,\Lambda) = (1-\exp(-\Lambda(u)e^{\theta' z}))^{\delta_1}(\exp(-\Lambda(u)e^{\theta' z}) - \exp(-\Lambda(v)e^{\theta' z}))^{\delta_2}(\exp(-\Lambda(v)e^{\theta' z})^{\delta_3}.$$

Let $\phi = \log \Lambda$. we reparametrize this log-likelihood in terms of (θ, ϕ). The resulting log-likelihood for a sample of size one is, up to an additive term not dependent on (θ, ϕ) is

$$\begin{aligned} l(\theta,\phi;x) &= \log p(x,\theta,\phi) \\ &= \delta_1 \log(1-\exp(-e^{z'\theta+\phi(u)})) + \delta_2 \log(\exp(-e^{z'\theta+\phi(u)}) - \exp(-e^{z'\theta+\phi(v)})) \\ &\quad -\delta_3 e^{z'\theta+\phi(v)}. \end{aligned}$$

Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ with $X_i = (\delta_{1i}, \delta_{2i}, U_i, V_i, Z_i)$, for $1 \le i \le n$ being a random sample with the same distribution as $X = (\delta_1, \delta_2, U, V, Z)$. The log-likelihood for this random sample is

$$l_n(\theta,\phi) = \sum_{i=1}^{n} l(\theta,\phi;X_i).$$

Because ϕ is a nondecreasing function, it is desirable to restrict its estimator to be also nondecreasing. Therefore, we seek an estimate of ϕ in the space of $\mathcal{M}_n$. (The abbreviation of $\mathcal{M}_n(D_n, K_n, m)$) The B-splines sieve MLE of $\tau_0 = (\theta_0, \phi_0)$ is the $\widehat{\tau}_n = (\widehat{\theta}_n, \widehat{\phi}_n)$ that maximizes $l_n(\theta, \phi)$ over $\Theta \times \mathcal{M}_n$. This is equivalent to maximizing $l_n(\theta, \mathcal{B}_n'\beta)$ over $\Theta \times B_n$. No restriction will be placed on Θ. Thus Θ can be taken to be R^d.

As the same as in the model for panel count data, the B-splines sieve MLE is much easy to compute than the semiparametric MLE studied in Huang and Wellner (1995). In

addition, we can show that under some mild regularity conditions, the sieve MLE of θ, $\widehat{\theta}_n$, achieves the semiparametric efficiency as well, with the information given by

$$I(\theta_0) = P(l^*(\theta_0, \phi_0; X)^{\otimes 2} = P\left(\dot{l}_1(\theta_0, \phi_0; X) - \dot{l}_2(\theta_0, \phi_0; X)(\xi_0)\right)^{\otimes 2},$$

where $\xi_0(t)$ is the solution of a Fredholm integral equation of the second kind,

$$\xi_0(t) - \int K(t, x)\xi_0(x)dx = d(t)$$

with two complicate functions $K(t, x)$ and $d(t)$ described in Huang and Wellner (1995). It is obvious that a direct estimation of the information matrix is impossible for this model. However, with the B-splines sieve MLE approach, variance of $\widehat{\theta}_n$ can be readily estimated using the observed information matrix defined in (3.1) due to Propositions 2.1 and 3.1.

For the asymptotic normality of $\widehat{\theta}_n$ and consistency of the inverse observed information matrices, the following conditions are assumed.

(C1) (a) $E(ZZ')$ is nonsingular; (b) Z is bounded, that is, there exists $z_0 > 0$ such that $P(\|Z\| \leq z_0) = 1$.

(C2) (a) There exists a positive number η such that $P(V - U \geq \eta) = 1$; (b) the union of the supports of U and V is contained in an interval $[a, b]$, where $0 < a < b < \infty$, and $0 < \Lambda_0(a) < \Lambda_0(b) < \infty$.

(C3) Λ_0 belongs to Φ, a class of functions with bounded pth derivative in $[a, b]$ for $p \geq 1$ and the first derivative of Λ_0 is strictly positive and continuous on $[a, b]$.

(C4) The conditional density $g(u, v|z)$ of (U, V) given Z has bounded partial derivatives with respect to (u, v). The bounds of these partial derivatives do not depend on (u, v, z).

(C5) For some $\kappa \in (0,1)$, $a^T var(Z|U)a \geq \kappa a^T E(ZZ^T|U)a$ and $a^T var(Z|V)a \geq \kappa a^T E(ZZ^T|V)a$ a.s. for all $a \in R^d$.

It should be noted that in applications, implementation of the proposed estimation method does not require these conditions to be satisfied. These conditions are sufficient but may not be necessary to prove the following asymptotic theorem. Some conditions may be weakened but will make the proof considerably more difficulty. However, from a view point of practice, these conditions are usually satisfied.

To study the asymptotic properties, we facilitate a metric defined as follows: for any $\phi_1, \phi_2 \in \Phi$, define

$$\|\phi_1 - \phi_2\|_\Phi^2 = E[\phi_1(U) - \phi_2(U)]^2 + E[\phi_1(V) - \phi_2(V)]^2.$$

and for any $\tau_1 = (\theta_1, \phi_1)$ and $\tau_2 = (\theta_2, \phi_2)$ in the space of $\mathcal{T} = \Theta \times \Phi$, define

$$d(\tau_1, \tau_2) = \|\tau_1 - \tau_2\|_\mathcal{T} = \left\{\|\theta_1 - \theta_2\|^2 + \|\phi_1 - \phi_2\|_\Phi^2\right\}^{1/2}.$$

THEOREM 4.1. *Let $K_n = O(n^\nu)$, where ν satisfies the restriction $\frac{1}{2(1+p)} < \nu < \frac{1}{2p}$. Suppose that T and (U,V) are conditionally independent given Z and that the distribution of (U,V,Z) does not involve (θ, Λ). Furthermore, suppose that conditions (C1)–(C5) hold. Then*

(i) $d(\widehat{\tau}_n, \tau_0) \to_p 0$.

(ii) $d(\widehat{\tau}_n, \tau_0) = O_p\left(n^{-\min(p\nu,(1-\nu)/2)}\right)$. *Thus if* $\nu = 1/(1+2p)$, $d(\widehat{\tau}_n, \tau_0) = O_p(n^{-p/(1+2p)})$. *This is the optimal rate of convergence in nonparametric regression with comparable smoothness assumptions.*

(iii) $n^{1/2}(\widehat{\theta}_n - \theta_0) = n^{-1/2} I^{-1}(\theta_0) \sum_{i=1}^{n} l^*(\theta_0, \phi_0; X_i) + o_p(1) \to_d N(0, I^{-1}(\theta_0))$. *Thus* $\widehat{\theta}_n$ *is asymptotically normal and efficient.*

(iv) The inverse observed information matrix is a consistent estimator of $I^{-1}(\theta_0)$*, the asymptotic variance of* $n^{1/2}(\widehat{\theta}_n - \theta_0)$.

The proof of this theorem is considerably long, we will provide a sketch of the proof in Appendix A.

5. Numerical Results

5.1. Simulations Studies

In this section, we conduct simulation studies for the examples discussed in the preceding section to evaluate the finite sample performance of the proposed estimator. In each example, we estimate the unknown parameters using the cubic B-splines sieve maximum likelihood estimation and estimate the standard error of the regression parameter estimates using the proposed least-squares method based on the cubic B-splines as well. For the B-splines sieve, the number of knots is chosen as $K_n = [N^{1/3}]$, the largest integer below $N^{1/3}$, where N is the number of distinct observation time points in the data, and the knots are evenly placed between $(0, 1)$ in the first example and $(0, 5)$ in the second example.

Simulation 1: Panel Count Data. We generate the data with the setting given in Wellner and Zhang (2007). For each subject, we independently generate $X_i = (Z_i, K_i, \underline{T}^{(i)}_{K_i}, \underline{\mathbb{N}}^{(i)}_{K_i})$, for $i = 1, 2, \cdots, n$, where $Z_i = (Z_{i,1}, Z_{i,2}, Z_{i,3})$ with $Z_{i,1} \sim$ Unif$(0, 1)$, $Z_{i,2} \sim N(0, 1)$, and $Z_{i,3} \sim$ Bernoulli(0.5); K_i is sampled randomly from the discrete set, $\{1, 2, 3, 4, 5, 6\}$; Given K_i, $\underline{T}^{(i)}_{K_i} = (T^{(i)}_{K_i,1}, T^{(i)}_{K_i,2}, \cdots, T^{(i)}_{K_i,K_i})$ are the order statistics of

Table 1: **Monte-Carlo simulations results of the B-splines sieve MLE of θ_0 with 1000 repetitions for semiparametric analysis of panel count data**

	n=50			n=100			n=200		
	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$
Bias	0.0001	-0.0003	0.0014	0.0003	0.0005	0.0001	-0.0012	0.0003	-0.0002
M-C sd	0.1029	0.0286	0.0712	0.0685	0.0188	0.0488	0.0474	0.0141	0.0337
ASE	0.1365	0.0418	0.0865	0.0805	0.0239	0.0542	0.0519	0.0152	0.0359
95%-CP	98.4%	98.5%	97.5%	97.6%	97.9%	96.5%	96.2%	95.1%	96.6%

K_i random draws from Unif$(0, 1)$; The panel counts $\underline{\mathbb{N}}_{K_i}^{(i)} = (\mathbb{N}_{K_i,1}^{(i)}, \mathbb{N}_{K_i,2}^{(i)}, \cdots, \mathbb{N}_{K_i,K_i}^{(i)})$ are generated from the Poisson process with the conditional mean function given by $\Lambda(t|Z_i) = 2t\exp(\theta_0^T Z_i)$ with $\theta_0 = (-1.0, 0.5, 1.5)^T$.

We conduct the simulation study for n =50, 100 and 200, respectively. In each case, we perform Monte-Carlo study with 1000 repetitions. Table 1 displays the estimation bias(Bias), Monte-Carlo standard deviation(M-C s.d.), the average of standard errors using the proposed method(ASE), and the coverage probability of 95% Wald-confidence interval using the proposed estimator of standard error(95%-PC).

The results show that the B-splines sieve MLE performs quite well. It has very little bias with seemingly decrease of estimation variability as sample size increases. The proposed method tends to overestimate the standard error slightly but the overestimation lessens as sample size increases. As the result of overestimation, the coverage probability of 95% confidence interval exceeds the nominal value when sample size is 50 or 100 but gets closer to 95% with sample size increasing to 200.

Simulation 2: Interval-Censored Data. We generate the data in a way similar to

Table 2: **Monte-Carlo simulations results for B-splines sieve MLE of θ_0 with 1000 repetitions for semiparametric analysis of interval-censored data**

	n=50	n=100	n=200
Bias	0.1194	0.0572	0.0196
M-C sd	0.7850	0.4966	0.3422
ASE	0.8649	0.5212	0.3506
95%-CP	97.4%	96.4%	95.6%

what was used in Huang and Rossini (1997). For each subject, we independently generate $X_i = (U_i, V_i, \delta_{i,1}\delta_{i,2}, Z_i)$, for $i = 1, 2, \cdots, n$, where $Z_i \sim$ Bernoulli(0.5); we simulate a series of examination times by the partial sum of interarrival times that are independently and identically distributed according to exp(1), then U_i is the last examination time within 5 at which the event has not occurred yet and V_i is the first observation time within 5 at which the event has occurred; the event time is generated according to Cox proportional hazards model $\Lambda(t) = t\exp(Z)$ for which the true parameters are: $\theta_0 = 1$ and $\Lambda_0(t) = t$. Similarly, Monte-Carlo study with 1000 repetition is performed and the corresponding results are displayed in Table 2.

The results show that both bias and Monte-Carlo standard deviation decrease as sample size increases. As observed earlier, the proposed least square estimate of the standard error may overestimate the true value but the overestimation lessens as sample size increasing to 200. With sample size 200, the Wald 95% confidence-interval achieves the desired coverage probability.

5.2. Applications

This section illustrates the method in two real life examples: the bladder tumor randomized clinical trial conducted by Byar, Blackard and VACURG (1977) and the breast cosmesis trial described by Finkelstein and Wolfe (1985). We adopt the cubic B-splines sieve semiparametric MLE method and we estimate the asymptotic standard error of the estimates of regression parameter using the least-squares approach with cubic B-splines as well. The inference is made based on asymptotic theorem developed in this paper.

Example 1: Bladder Tumor Study. The data set of the bladder tumor randomized clinical trial conducted by the Veterans Administration Cooperative Urological Research Group (Byar, Blackard and VACURG, 1977) is extracted from Andrews and Herzberg (1985, p. 253-260). In this study, a randomized clinical trial of three treatments : placebo, pyridoxine pill and thiotepa instillation into bladder, was conducted for patients with superficial bladder tumor (a total of 116 subjects: 40 were randomized to placebo, 31 to pyridoxine pill and 38 to thiotepa instillation) when entering the trial. At each follow-up visit, tumors were counted, measured and then removed after being found. The treatments as originally assigned will continue after each visit. The number of follow-up visits and follow-up times varied greatly from patient to patient and hence the observation of bladder tumor counts in this study falls in the framework of panel count data as described in Section 4.

For this trial, the treatment effects, particularly the thiotepa instillation method, on reducing the bladder tumor recurrent have been the focal point of interest in many studies, see for example, Wei, Lin and Weissfeld (1989), Sun and Wei (2000), Zhang (2002) and Wellner and Zhang (2007). In this paper, we study the proportional mean model as proposed by Wellner and Zhang (2007),

$$E\{\mathbb{N}(t)|Z\} = \Lambda_0(t)\exp(\theta_{0,1}Z_1 + \theta_{0,2}Z_2 + \theta_{0,3}Z_3 + \theta_{0,4}Z_4), \tag{5.1}$$

where Z_1 and Z_2 are the baseline tumor count and size, measured when subjects entered the study, and Z_3 and Z_4 define the indicators of the pyridoxine pill and instillation treatments, respectively. Lu, Zhang and Huang (2008) has used the cubic B-splines sieve semiparametric MLE method for this model and estimated the asymptotic standard error of the estimate of θ_0 based on the bootstrap approach. In this paper, we reanalyze the data using the same method but estimate the asymptotic standard error using the least-squares approach. The semiparametric sieve MLE of θ_0 is $\widehat{\theta}_n = (0.2076, -0.0356, 0.0647, -0.7949)$ with the asymptotic standard errors given by $(0.0066, 0.0101, 0.0338, 0.0534)$ and the corresponding p-values=(0.0000,0.0004,0.0553,0.0000) based on the asymptotic theorem developed in Wellner and Zhang (2007). We note that the result of this analysis is very different from that of Lu, Zhang and Huang (2008). The conflict of the results indirectly indicates that the working assumption of Poisson process model to form the likelihood may not be valid as Lu-Zhang-Huang's inference procedure is shown to be robust against the underlying counting process.

Example 2: Breast Cosmesis Study. The breast cosmesis study is the clinical trial for comparing radiotherapy alone with primary radiotherapy with adjuvant chemotherapy in terms of subsequent cosmetic deterioration of the breast following tumorectomy. Subjects (46 assigned to radiotherapy alone and 48 to radiotherapy plus chemotherapy) were followed for up to 60 months, with pre-scheduled follow-up visits for every 4-6 months. In this paper, we propose Cox proportional hazards model to analyze the difference in time until appearance

of breast retraction,

$$\Lambda(t|Z) = \Lambda_0(t)\exp(\theta_0 Z), \tag{5.2}$$

where Λ_0 is the baseline hazard (the hazard of the time to appearance of breast retraction) for radiotherapy alone) and Z is the indicator for the treatment of radiotherapy plus chemotherapy. Using the method proposed in this paper, the cubic B-splines sieve semiparametric MLE of θ_0 is $\widehat{\theta}_n = 0.8948$ with asymptotic standard error given by 0.2926. The Wald test statistic is $Z = 3.0582$ with the associated p-value=0.0011. This indicates that the treatment of radiotherapy with adjuvant chemotherapy significantly increases the risk of the breast retraction and the result is comparable with what has been concluded in Finkelstein and Wolfe (1985).

6. Conclusion and Discussion

When the infinite dimensional parameter as nuisance parameter can not be eliminated in estimating the finite dimensional parameter, a general semiparametric maximum likelihood estimation is often a challenging task. Sieve MLE method, proposed originally by Geman and Hwang (1982), renders a practical approach for alleviating the difficulty in semiparametric estimation problem. Particularly, the spline-based sieve semiparametric method, as exemplified by Lu, Zhang and Huang (2008) has many attractions in practice. Not only it reduces the numerical difficulty in computing the semiparametric MLE, it also achieves the semiparametric asymptotic estimation efficiency for the finite dimensional parameter. However, the estimation of the information matrix remains a difficult task in general. In this paper, we propose an easy-to-implement least-squares approach to estimate

the semiparametric information matrix. We show that the estimator is asymptotic consistent, as often a byproduct of the establishment of asymptotic normality for sieve semiparametric MLE. Interestingly, this estimator is exactly the observed information matrix if we treat the semiparametric sieve MLE as a parametric MLE problem. Although the estimator of asymptotic error overestimates the true value slightly in finite sample situation as shown in our simulation studies, the overestimation issue is generally alleviated as sample size increases, say to 200.

In addition to the expression of information matrix given in (2.7), we note that it can be expressed through the second derivatives. Denote $\ddot{l}_{11}(\tau; x) = \frac{\partial}{\partial\theta}\dot{l}_1(\tau; x)$, $\ddot{l}_{12}(\tau; x) = \frac{\partial}{\partial s}\dot{l}_1(\theta, \phi_{(s)}; x)|_{s=0}$, $\ddot{l}_{21}(\tau; x)(h) = \frac{\partial}{\partial\theta}\dot{l}_2(\tau; x)(h)$, and letting $(\partial/\partial s)\,\phi_{1(s)}|_{s=0} = h_1$, $\ddot{l}_{22}(\tau; x)(h, h_1) = \frac{\partial}{\partial s}\dot{l}_2(\theta, \phi_{1(s)}; x)(h)$. Then the information matrix can be written as

$$I(\theta_0) = -E\left[\ddot{l}_{11}(\tau_0; X) - 2\ddot{l}_{12}(\tau_0; X)(\xi_0) + \ddot{l}_{22}(\tau_0; X)(\xi_0, \xi_0)\right].$$

This expression leads to an alternative estimator of $I(\theta_0)$, given by

$$\tilde{\rho}_n(\tilde{\xi}_n) = -n^{-1}\sum_{i=1}^{n}\left[\ddot{l}_{11}(\widehat{\tau}_n; X_i) - 2\ddot{l}_{12}(\widehat{\tau}_n; X_i)(\tilde{\xi}_n) + \ddot{l}_{22}(\widehat{\tau}_n; X_i)(\tilde{\xi}_n, \tilde{\xi}_n)\right], \tag{6.1}$$

where $\tilde{\xi}_n$ is the minimizer of $\tilde{\rho}_n(h)$ over $\mathcal{H}_n$. The consistency of $\tilde{\rho}_n(\tilde{\xi}_n)$ can be similarly shown and it would be interesting to investigate how this estimator behaves compared to the one proposed in this paper.

As implied in Example 1, this semiparametric inference procedure is not robust against the underlying probability model. If the true model is not the working model for forming the likelihood, the inference may be invalid. Wellner and Zhang (2007) developed a general theorem for making robust semiparametric inference and Lu, Zhang and Huang (2007) extend

the robust semiparametric inference to the spline-based sieve semiparametric estimation. It remains an open research problem on how to generalize the proposed least-squares method to estimate the asymptotic error of the regression parameter estimates in order to make robust semiparametric inference.

7. Appendix A

This section contains the sketch of the proofs for Proposition 2.1 and Theorem 4.1. In the following, C represents a constant that may varies from place to place.

The proof for Proposition 2.1:

For the least favorable direction ξ_0, we define $\mathcal{H}_n(\epsilon) = \{h : h \in \mathcal{H}_n \text{ and } \|h - \xi_0\|_{\mathcal{H}} \geq \epsilon\}$ for $\epsilon \downarrow 0$. Note that for any $h \in \mathcal{H}$,

$$\begin{aligned}\rho_n(h) - \rho_n(\xi_0) &= (\mathbb{P}_n - P)\ell(\widehat{\tau}_n, h; X) - (\mathbb{P}_n - P)\ell(\widehat{\tau}_n, \xi_0; X) \\ &\quad + P\{\ell(\widehat{\tau}_n, h; X) - \ell(\tau_0, h; X)\} - P\{\ell(\widehat{\tau}_n, \xi_0; X) - \ell(\tau_0, \xi_0; X)\} \\ &\quad + P\{\ell(\tau_0, h; X) - \ell(\tau_0, \xi_0; X)\}\end{aligned}$$

Because the class $\Im = \{\ell(\tau, h; X) : \tau \in \mathcal{T} \text{ and } h \in \mathcal{H}\}$ is Glivenko-Cantelli and $\widehat{\tau}_n \in \mathcal{H}_n \subset \mathcal{H}$, we have that

$$(\mathbb{P}_n - P)\ell(\widehat{\tau}_n, h; X) = o_P(1) \text{ and } (\mathbb{P}_n - P)\ell(\widehat{\tau}_n, \xi_0; X) = o_p(1).$$

In addition, by the weak consistency of $\widehat{\tau}_n \to_p \tau_0$, Continuous Mapping Theorem and

Dominate Convergence Theorem, we can conclude that

$$P\left\{\ell(\widehat{\tau}_n, h; X) - \ell(\tau_0, h; X)\right\} \to_p 0 \quad \text{and} \quad P\left\{\ell(\widehat{\tau}_n, \xi_0; X) - \ell(\tau_0, \xi_0; X)\right\} \to_p 0.$$

Hence for any $h \in \mathcal{H}$,

$$\rho_n(h) - \rho_n(\xi_0) = \rho(h) - \rho(\xi_0) + o_p(1) > 0 \quad \text{in probability} \quad \text{as} \quad n \to \infty$$

by the characteristics of the least favorable direction. This implies that

$$P\left(\inf_{h \in \mathcal{H}_n(\epsilon)} \rho_n(h) > \rho_n(\xi_0)\right) \to_p 1$$

and hence let $\epsilon \to 0$ we can conclude that $\|\hat{\xi}_n - \xi_0\|_{\mathcal{H}} \to_p 0$.

Subsequently, we can conclude

$$\begin{aligned} \rho_n(\hat{\xi}_n) &= \mathbb{P}_n \ell(\widehat{\tau}_n, \hat{\xi}_n; X) \\ &= (\mathbb{P}_n - P)\ell(\widehat{\tau}_n, \hat{\xi}_n; X) + P\ell(\widehat{\tau}_n, \hat{\xi}_n; X) \to_p P\ell(\tau_0, \xi_0; X) = I(\theta_0), \end{aligned}$$

by the consistency of $\widehat{\tau}_n$ and $\hat{\xi}_n$ and $\Im$ being a Glivenko-Cantelli.

The proof for Theorem 4.1:

To prove Part (i) of Theorem 4.1, we verify the conditions of Theorem 5.7 in van der Vaart (1998). Let $\mathbb{M}(\tau) = Pl(\tau; X) = Pl(\theta, \phi; X)$ and $\mathbb{M}_n(\tau) = \mathbb{P}_n l(\tau; X) = \mathbb{P}_n l(\theta, \phi; X)$. Hence for any $\tau = \mathcal{T}_n = \Theta \times \mathcal{M}_n$, $\mathbb{M}_n(\tau) - \mathbb{M}(\tau) = (\mathbb{P}_n - P)l(\tau; X)$.

Let $\mathcal{L}_1 = \{l(\tau; X) : \quad \tau \in \mathcal{T}_n\}$. By the calculation of Shen and Wong (1994), page 597, $\forall \epsilon > 0$, the bracketing number of $\mathcal{M}_n$ computed with $L_1(P)$-norm is bounded by $(1/\epsilon)^{Cq_n}$.

Since $\Theta \subset R^d$ is compact, Θ can be covered by $C(1/\epsilon)^d$ balls with radius ϵ. Then by Conditions (C1)-(C3), we can easily construct ϵ-brackets for $\mathcal{L}_1$ with the bracketing number with $L_1(P)$-norm bounded by $C(1/\epsilon)^{Cq_n+d}$. Hence $\mathcal{L}_1$ is Glivenko-Cantelli by Theorem 2.4.1 of van der Vaart and Wellner (1996). Therefore, $\sup_{\tau\in\mathcal{T}_n} |\mathbb{M}_n(\tau) - \mathbb{M}(\tau)| \to_p 0$.

Let $g(z,t) = \exp(\theta' z + \phi(t))$ and $g_0(z,t) = \exp(\theta_0' z + \phi_0(t))$. A straightforward algebra yields that

$$\begin{aligned}
\mathbb{M}(\tau_0) - \mathbb{M}(\tau) &= E\left\{[1-\exp(-g_0(Z,U))]\log\frac{1-\exp(-g_0(Z,U))}{1-\exp(-g(Z,U))}\right. \\
&\quad +[\exp(-g_0(Z,U))-\exp(-g_0(Z,V))]\log\frac{\exp(-g_0(Z,U))-\exp(-g_0(Z,V))}{\exp(-g(Z,U))-\exp(-g(Z,V))} \\
&\quad \left.+\exp(-g_0(Z,V))\log\frac{\exp(-g_0(Z,V))}{\exp(-g(Z,V))}\right\} \\
&= E\left\{[1-\exp(-g(Z,U))]m\left(\frac{1-\exp(-g_0(Z,U))}{1-\exp(-g(Z,U))}\right)\right. \\
&\quad +[\exp(-g(Z,U))-\exp(-g(Z,V))]m\left(\frac{\exp(-g_0(Z,U))-\exp(-g_0(Z,V))}{\exp(-g(Z,U))-\exp(-g(Z,V))}\right) \\
&\quad \left.+\exp(-g(Z,V))m\left(\frac{\exp(-g_0(Z,V))}{\exp(-g(Z,V))}\right)\right\},
\end{aligned}$$

where $m(x) = x\log x - x + 1 \geq (x-1)^2/4$ for $0 \leq x \leq 5$. Further analysis using Taylor expansion and Conditions (C1)-(C3) leads to

$$\begin{aligned}
\mathbb{M}(\tau_0) - \mathbb{M}(\tau) &\geq CE\left\{\frac{1}{1-\exp(-g(Z,U))}\left[\exp(-g_0(Z,U))-\exp(-g(Z,U))\right]^2\right. \\
&\quad \left.+\frac{1}{\exp(-g(Z,V))}\left[\exp(-g_0(Z,U))-\exp(-g(Z,U))\right]^2\right\} \\
&\geq CE\left\{\left[(\theta_0-\theta)'Z+(\phi_0-\phi)(U)\right]^2+\left[(\theta_0-\theta)'Z+(\phi_0-\phi)(V)\right]^2\right\}.
\end{aligned}$$

With Conditions (C1)-(C5), using the same arguments as those in Wellner and Zhang (2007),

page 2126-2127 leads to

$$\mathbb{M}(\tau_0) - \mathbb{M}(\tau) \geq C\left(\|\theta - \theta_0\|^2 + \|\phi - \phi_0\|_\Phi^2\right) = Cd^2(\tau_0, \tau).$$

Then it implies that $\sup_{\tau: d(\tau,\tau_0)\geq\epsilon} \mathbb{M}(\tau) \leq \mathbb{M}(\tau_0) - C\epsilon^2 < \mathbb{M}(\tau_0)$.

For $\phi_0 \in \Phi$, Lu (2007) has shown that there exists a $\phi_{0,n} \in \mathcal{M}_n$ of order $m \geq p+2$ such that

$$\|\phi_{0,n} - \phi_0\|_\infty \leq Cq_n^{-p} = O\left(n^{-p\nu}\right).$$

This also implies that $\|\phi_{0,n} - \phi_0\|_\Phi \leq Cq_n^{-p} = O\left(n^{-p\nu}\right)$. Now let $\tau_{0,n} = (\theta_0, \phi_{0,n})$, we have

$$\begin{aligned}
\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) &= \mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_{0,n}) + \mathbb{M}_n(\tau_{0,n}) - \mathbb{M}_n(\tau_0) \\
&\geq \mathbb{P}_n l(\tau_{0,n}; X) - \mathbb{P}_n l(\tau_0; X) \\
&= (\mathbb{P}_n - P)\left\{l(\tau_{0,n}; X) - l(\tau_0; X)\right\} + \mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0).
\end{aligned}$$

We can easily show that the class $\mathcal{L}_2 = \{l(\theta_0, \phi; x) - l(\theta_0, \phi_0; x) : \ \phi \in \mathcal{M}_n \ \text{ and } \ \|\phi - \phi_0\|_\Phi \leq Cn^{-p\nu}\}$ is a P-Donsker by calculating the bracketing number with $L_2(P)$-norm. It is obvious that in this class $P(l(\theta_0, \phi; X) - l(\theta_0, \phi_0; X))^2 \to 0$ as $n \to \infty$. Hence

$$(\mathbb{P}_n - P)\left\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\right\} = o_p(n^{-1/2})$$

by the relationship between P-Donsker and asymptotic equicontinuity given by Corollary 2.3.12 of van der Vaart and Wellner (1996). By the Dominated Convergence Theorem, it is easy to see that $\mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0) > -o(1)$. Therefore,

$$\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq o_p(n^{-1/2}) - o(1) = -o_p(1).$$

This completes the proof of $d(\widehat{\tau}_n, \tau_0) \to_p 0$.

To prove Part (ii), we verify the conditions of Theorem 3.2.5 of van der Vaart and Wellner (1996). First, we have already shown in the proof of consistency that $\mathbb{M}(\tau_0) - \mathbb{M}(\tau) \geq Cd^2(\tau_0, \tau)$.

Next, we further explore $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0)$. In the proof of Part (i), we know that $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq I_{1,n} + I_{2,n}$, where $I_{1,n} = (\mathbb{P}_n - P)\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\}$ and $I_{2,n} = P\{l(\theta_0, \phi_{0,n}; X) - l(\theta_0, \phi_0; X)\}$. By Taylor expansion, we have

$$I_{1,n} = (\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)(\phi_{0,n} - \phi_0)\right\} = n^{-p\nu+\epsilon}(\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)\frac{\phi_{0,n} - \phi_0}{n^{-p\nu+\epsilon}}\right\}$$

for any $0 < \epsilon < 1/2 - p\nu$. Because $\|\phi_{0,n} - \phi_0\|_\infty = O(n^{-p\nu})$, using Corollary 2.3.12 of van der Vaart and Wellner (1996), we can show that $(\mathbb{P}_n - P)\left\{\dot{l}_2(\theta_0, \tilde{\phi}; X)\frac{\phi_{0,n} - \phi_0}{n^{-p\nu+\epsilon}}\right\} = o_p(n^{-1/2})$. Hence $I_{1,n} = o_p(n^{-p\nu+\epsilon}n^{-1/2}) = o_p(n^{-2p\nu})$. Using the fact that the function $m(x) = x\log x - x + 1 \leq (x-1)^2$ in the neighborhood of $x = 1$, it can be easily argued that $\mathbb{M}(\tau_0) - \mathbb{M}(\tau_{0,n}) \leq C\|\phi_{0,n} - \phi_0\|_\Phi^2 = O(n^{-2p\nu})$, which implies that $I_{2,n} = \mathbb{M}(\tau_{0,n}) - \mathbb{M}(\tau_0) \geq -O(n^{-2p\nu})$. Thus we conclude that $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq -O_p(n^{-2p\nu}) = -O_p\left(n^{-2\min(p\nu,(1-\nu)/2)}\right)$.

Let $\mathcal{L}_3(\eta) = \{l(\tau; x) - l(\tau_0; x) : \quad \phi \in \mathcal{M}_n \quad \text{and} \quad d(\tau, \tau_0) \leq \eta\}$. Using the same argument as that in the proof of consistency, we obtain that the logarithm of the ϵ-bracketing number of $\mathcal{L}_3(\eta)$, $\log N_{[\,]}(\epsilon, \mathcal{L}_3(\eta), L_2(P))$ is bounded by $Cq_n\log(\eta/\epsilon)$. This leads to $J_{[\,]}(\eta, \mathcal{L}_3(\eta), L_2(P)) = \int_0^\eta \sqrt{1 + \log N_{[\,]}(\epsilon, \mathcal{L}_3(\eta), L_2(P))}d\epsilon \leq Cq_n^{1/2}\eta$. Because Conditions (C1) and (C2) guarantee the uniform boundedness of $l(\tau; x)$, using Theorem 3.4.1 of van der Vaart and Wellner (1996), the key function $\phi_n(\eta)$ in Theorem 3.2.5 of van der Vaart and

Wellner (1996) is given by $\phi_n(\eta) = q_n^{1/2}\eta + q_n/n^{1/2}$. Note that

$$n^{2p\nu}\phi_n(1/n^{p\nu}) = n^{p\nu}n^{\nu/2} + n^{2p\nu}n^{\nu} + n^{2p\nu}n^{\nu}/n^{1/2} = n^{1/2}\left\{n^{p\nu-(1-\nu)/2} + n^{2p\nu-(1-\nu)}\right\}.$$

Therefore, if $p\nu \leq (1-\nu)/2$, $n^{2p\nu}\phi_n(1/n^{p\nu}) \leq n^{1/2}$. This implies that if we choose $r_n = \min(p\nu, (1-\nu)/2)$, it follows that $r_n^2\phi_n(1/r_n) \leq n^{1/2}$ and $\mathbb{M}_n(\widehat{\tau}_n) - \mathbb{M}_n(\tau_0) \geq -O_p(r_n^{-2})$. Hence $r_n d(\widehat{\tau}_n, \tau_0) = O_p(1)$.

To derive the asymptotic normality for $\hat{\theta}_n$, we just need to verify the conditions of the general theorem given in Appendix B. For Condition (B1), we only need to verify that $\mathbb{P}_n\hat{l}_2(\widehat{\theta}_n, \widehat{\phi}_n; X)(\xi_0) = o_p(n^{-1/2})$ since $\mathbb{P}_n\dot{l}_1(\widehat{\theta}_n, \widehat{\phi}_n; X) \equiv 0$. Because ξ_0 has a bounded derivative, it is also a function with bounded variation. Then it can be easily shown using the argument in Bilingsley (1986, page 435-436) that there exist a $\xi_{0,n} \in S_n(D_n, K_n, m)$ such that $\|\xi_{0,n} - \xi_0\|_\Phi = O(q_n^{-1}) = O(n^{-\nu})$ and $\mathbb{P}_n\dot{l}_2(\widehat{\tau}_n; X)(\xi_{0,n}) = 0$. Therefore we can write $\mathbb{P}_n\dot{l}_2(\widehat{\tau}_n; X)(\xi_0) = I_{3,n} + I_{4,n}$, where $I_{3,n} = (\mathbb{P}_n - P)\dot{l}_2(\widehat{\tau}_n; X)(\xi_0 - \xi_{0,n})$ and $I_{4,n} = P\left\{\dot{l}_2(\widehat{\tau}_n; X)(\xi_0 - \xi_{0,n}) - \dot{l}_2(\tau_0; X)(\xi_0 - \xi_{0,n})\right\}$.

Let $\mathcal{L}_4 = \{\dot{l}_2(\tau; x)(\xi_0 - \xi) : \quad \tau \in \mathcal{T}_n, \quad \xi \in S_n(D_n, K_n, m) \quad \text{and} \quad \|\xi_0 - \xi\|_\Phi \leq n^{-\nu}\}$. It can be similarly shown that $\mathcal{L}_4$ is a P-Donsker and for any $r(\tau, \xi; x) \in \mathcal{L}_4$, $Pr^2 \to_p 0$ as $n \to \infty$. Hence $I_{3,n} = o_p(n^{-1/2})$ by Corollary 2.3.12 of van der Vaart and Wellner (1996). By Cauchy-Schwatz inequality, it can be easily shown that

$$\begin{aligned} I_{4,n} &\leq Cd(\widehat{\tau}_n, \tau_0)\|\xi_0 - \xi_{0,n}\|_\Phi = O_p\left(n^{-\min(p\nu,(1-\nu)/2)}n^{-\nu}\right) = O_p\left(n^{-\min(\nu(p+1),(1+\nu)/2)}\right) \\ &= o_p(n^{-1/2}). \end{aligned}$$

So (B1) holds. (B2) holds by showing that the class $\mathcal{L}_5(\eta) = \{l^*(\tau; x) - l^*(\tau_0; x) : \quad \tau \in$

$\mathcal{T}_n$ and $d(\tau, \tau_0) \leq \eta\}$ is P-Donsker and for any $r(\tau; x) \in \mathcal{L}_5(\eta)$, $Pr^2 \to_p 0$ as $\eta \to 0$. (B3) can be easily established using Taylor expansion and the rate of convergence derived in Part (ii). Hence the proof is complete.

For Part (iv), based on what we have shown in Parts (i) and (ii), it can be easily argued that the bracket number with $L_1(P)$-norm of the class $\Im$ defined in Proposition 2.1 is bounded. So $\Im$ is Glivenko-Cantelli by Theorem 2.4.1. of van der Vaart and Wellner (1996). Hence the result follows given the consistency of $(\widehat{\tau}_n, \hat{\xi}_n)$ and the approximation properties of B-splines.

8. Appendix B: A General Theorem for Asymptotic Normality

This section presents a general theorem for asymptotic normality of the MLE of final-dimensional parameter in the setting of semiparametric maximum likelihood estimation when the infinite-dimensional parameter is treated as a nuisance parameter. This theorem is the simplified version of the general theorem given in Huang (1996). The following conditions will be assumed.

(B1): $\mathbb{P}_n \dot{l}_1(\widehat{\theta}_n, \widehat{\phi}_n; X) = o_p(n^{-1/2})$ and $\mathbb{P}_n \dot{l}_2(\widehat{\theta}_n, \widehat{\phi}_n; X)(\xi_0) = o_p(n^{-1/2})$

(B2): $(\mathbb{P}_n - P)\left\{ l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X) \right\} = o_p(n^{-1/2})$

(B3): $P\left\{ l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X) \right\} = -I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2})$

THEOREM 8.1. *Suppose (B1)-(B3) are satisfied, and suppose that $I(\theta_0)$ is nonsingular. Then*

$$n^{1/2}(\widehat{\theta}_n - \theta_0) = n^{1/2} I^{-1}(\theta_0) \sum_{i=1}^{n} l^*(\theta_0, \phi_0; X_i) + o_p(1) \to_d N\left(0, I^{-1}(\theta_0)\right).$$

Proof: Combining (B2) and (B3), we have

$$\mathbb{P}_n \left\{ l^*(\widehat{\theta}_n, \widehat{\phi}_n; X) - l^*(\theta_0, \phi_0; X) \right\} = -I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2}).$$

By (B1), it follows that

$$\mathbb{P}_n l^*(\theta_0, \phi_0; X) = I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(\|\widehat{\theta}_n - \theta_0\|) + o_p(n^{-1/2})$$

Because $I(\theta_0)$ is nonsingular, and $\mathbb{P}_n l^*(\theta_0, \phi_0; X) = O_p(n^{-1/2})$, this implies that $\|\widehat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$. Thus $o_p(\|\widehat{\theta}_n - \theta_0\|) = o_p(n^{-1/2})$ and therefore

$$\mathbb{P}_n l^*(\theta_0, \phi_0; X) = I(\theta_0)(\widehat{\theta}_n - \theta_0) + o_p(n^{-1/2}).$$

The result follows.

REFERENCES

Andrews, D. F, and Herzberg, A. M. (1985). *Data; A Collection of Problems from Many Fields for the Students and Research Works.* New York: Springer-Verlag.

Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, **11**, 432-452.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Billingsley, P. (1986). *Probability and Measure.* John Wiley, New York.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields* **97**, 113-150.

Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Annals of Statistics*, **18**, 391-404.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Annals of Statistics*, **16**, 136-146.

Chen, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Annals of Statistics*, **23**, 1102-1129.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society.* Series B **34**, 187-220.

Finkelstein, D.M. and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-945.

Geman, A. and Hwang, C. R.(1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics*, **10**, 401-414.

Geskus, R. and Groeneboom, P. (1996). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *Technical Report*, Delft University.

Gill, R. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part I). *Scandinavian Journal of Statistics*, **20**, 271-288.

Groeneboom, P. (1996). Inverse problems in statistics. Proceedings of the St. Flour Summer School in Probability, 1994. *Lecture Notes in Math.*, Springer Verlag, Berlin.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar Band 19, Birkhäuser, Basel.

Gu, M. G. and Zhang, C.-H. (1993). Asymptotic Properties of self-consistent estimation based on doubly censored data. *Annals of Statistics*, **21**, 611-624.

Huang, J. (1996). Efficient Estimation for the Cox Model with Interval Censoring. *Annals of Statistics*, **24**, 540-568.

Huang, J. (1997). Limit distribution of a LS-estimator for the parametric component in the partly linear additive model. *Preprint*, Dept. of Statistics and Actuarial Sci., Univ. of Iowa.

Huang, J. and Rossini, A. J. (1997). Sieve Estimation for the Proportional Odds Failure-time Regression Model with Interval Censoring. *Journal of the American Statistical Association*, **92**, 960-967.

Huang, J. and Wellner (1995). Efficient Estimation For The Proportional Hazards Model With "Case 2" Interval Censoring. *Technical Report*, 289, Department of Statistics, University of Washington. (http://www.stat.washington.edu/tech.reports/).

Lu, M.(2007). *Analysis of Panel Count Data Using Monotone Polynomial Splines.* (Doctoral Dissertation), University of Iowa.

Lu, M., Zhang, Y. and Huang, J.(2008). Semiparametric Estimation Methods for Panel Count Data Using Monotone Polynomial Splines. *Technical Report*, 2008-1, Department of Biostatistics, University of Iowa. (http://www.public-health.uiowa.edu/biostat/research/documents).

Murphy, S. (1995). Asymptotic theory for the frailty model. *Annals of Statistics*, **23**, 182-198.

Murphy, S., Rossini, A. J., and van der Vaart, A. W. (1997). Maximum likelihood estimation in proportional odds model. *J. Amer. Statist. Assoc.*, **92**, 968-976.

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

Murphy, S. and van der Vaart, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.*, **95**, 449-465.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sorensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavia Journal of Statistics*, **19**, 25-44.

Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. *Applied Statistics* **33**, 169-175.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, **22**, 300-325.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York.

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

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditional parametric models. *Annals of Statistics*, **20**, 1768-1802.

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

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. and Kalbfleisch, J. D. (1995). Estimation of the Mean Function of Point Processes Based on Panel Count Data. *Statistical Sinica*, **5**, 279-190.

Sun, J., and Wei, L. J. (2000). Regression Analysis of Panel Count Data with Covariate-Dependent Observation and Censoring Times. *Journal of the Royal Statistical Society, Ser. B*, **62**, 293-302.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics*, **21**, 14-44.

van der Laan, M. J. (1993). *Efficient and Inefficient Estimation in Semiparametric Models* (Doctoral Dissertation). University of Utrecht, The Netherlands.

van der Vaart, A. W. (1991). On differentiable functionals. *Annals of Statistics*, **19**, 178-204.

van der Vaart, A. W. (1994). Maximum likelihood estimation with partially censored observations. *Annals of Statistics*, **22**, 1896-1916.

van der Vaart, A. W. (1996). Efficient estimation in semiparametric mixture models. *Annals of Statistics*, **24**, 862-878.

van der Vaart, A. W.(1998). *Asymptotic Statistics.* Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag, New York.

Wellner, A.J. and Zhang, Y.(2007). Likelihood-Based Semiparametric Estimation Methods for Panel Count Data with Covariates. *The Annals of Statistics*, **35**, 2106-2142.

Wong, W. H. and Severini, T. A. (1991) On maximum likelihood estimation in infinite dimensional parameter space. *Annals of Statistics*, **16**, 603-632.

Zhang, Y. (2002). A Semiparametric Pseudo-Likelihood Method for Panel Count Data. *Biometrika*, **89**, 39-48.