How Many Iterations are Sufficient for Semiparametric Estimation?

January 5, 2011

Guang Cheng*

Purdue University

Abstract: A common practice in obtaining a semiparametric efficient estimate is through iteratively maximizing the (penalized) log-likelihood w.r.t. its Euclidean parameter and functional nuisance parameter via Newton-Raphson algorithm. A rigorous theoretical study of the above semiparametric iterative estimation approach is the main purpose of this paper. We first show that the grid search algorithm produces the desirable initial estimate with the proper convergence rate. Our major contribution is to provide a formula in calculating the minimal number of iterations k^* needed to produce an efficient estimate $\widehat{\theta}_n^{(k^*)}$. We discover that (a) k^* depends on the convergence rates of the initial estimate and the nuisance functional estimate, and (b) k^* iterations are also sufficient for recovering the estimation sparsity in high dimensional data. The last contribution is the novel construction of $\widehat{\theta}_n^{(k)}$ which does not require knowing the form of the implicitly defined efficient score function. These general conclusions hold, in particular, when the nuisance parameter is not estimable at root-n rate, and apply to semiparametric models estimated under various regularizations, e.g., kernel or penalized estimation.

Keywords and phrases: Generalized Profile Likelihood, Higher Order Asymptotic Efficiency, K-step Estimation, Newton Raphson Algorithm, Semiparametric Models.

Short title: k-step Semiparametric Estimation

1. Introduction

Semiparametric models indexed by a Euclidean parameter of interest $\theta \in \Theta \subset \mathbb{R}^d$ and an infinite-dimensional nuisance parameter $\eta \in \mathcal{H}$ are proven to be useful in a variety of contexts, e.g., [1, 4, 8, 19, 24, 28, 30, 35, 40]. The semiparametric MLE for θ can be viewed as a solution of the implicitly defined efficient score function whose nonparametric estimation is only possible in some special cases, e.g., [24]. Therefore, it is generally hard

*Guang Cheng is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907-2066, Email: chengg@purdue.edu

to solve the MLE from the efficient score function analytically or numerically. A common practice is to maximize the log-profile likelihood

$$\log pl_n(\theta) = \sup_{\eta \in \mathcal{H}} \log lik_n(\theta, \eta), \tag{1}$$

where $lik_n(\theta, \eta)$ is the likelihood given n data, via some optimization algorithm. For example, the Newton-Raphson algorithm is applied to the partial likelihood of the Cox model in the software **R** (with the command *coxph*).

A general approach of obtaining a semiparametric efficient estimate of θ is to iteratively maximize the log-likelihood w.r.t. θ and η as follows:

General Semiparametric Iterative Estimation Approach

I. Identify an initial estimate $\widehat{\theta}_n^{(0)}$;

II. Construct the corresponding nuisance estimate $\widehat{\eta}(\widehat{\theta}_n^{(0)})$ either by pure nonparametric approach, e.g., isotonic estimation, or under some regularization, e.g., kernel or sieve estimation;

III. Apply the Newton-Raphson (NR) or other optimization algorithm to

$$\widehat{S}_n(\theta) = \log lik_n(\theta, \widehat{\eta}(\theta)), \tag{2}$$

at $\theta = \widehat{\theta}_n^{(0)}$ to obtain $\widehat{\theta}_n^{(1)}$;

IV. Repeat steps II-III k^* iterations until

$$|\widehat{S}_n(\widehat{\theta}_n^{(k^*)}) - \widehat{S}_n(\widehat{\theta}_n^{(k^*-1)})| \leq \epsilon$$

for some pre-determined sufficiently small ϵ.

If $\widehat{\eta}(\theta)$ is the nonparametric MLE (NPMLE) for any fixed θ, then $\widehat{S}_n(\theta)$ is just the profile likelihood defined in (1). In general, $\widehat{S}_n(\theta)$ is called the generalized profile likelihood as in [39]. The above likelihood estimation procedure or its M-estimation analog has been extensively implemented in the literature. Here is an incomplete list: (i) Odds-Rate Regression Model under Survival Data, e.g., [24, 30]; (ii) Semiparametric Regression under Shape Constraints, e.g., [4, 11]; (iii) Logistic Regression with Missing Covariates, e.g., [35]; (iv) Generalized Partly Linear (Single Index) Model, e.g., [8, 19]; (v) Conditionally Parametric Model, e.g. [39, 40]; (vi) Semiparametric Transformation Model, e.g., [28]. In addition, the above iterative procedure can also be adapted to accommodate the penalized estimation and selection of the semiparametric models by using a different criterion function than (2), see [7, 16, 29]. We will discuss that scenario in Section 5.2.

Unfortunately, the rigorous statistical analyses are unavailable for the above semiparametric estimation approach. This is because $\widehat{S}_n(\theta)$ at hand is usually implicitly defined, and thus its continuity/smoothness is unknown. In this paper, by using the empirical processes theories, we have made the following three contributions.

It is well known that identifying the preliminary estimate $\widehat{\theta}_n^{(0)}$ in a suitable neighborhood of θ_0 is critical in guaranteeing the fast convergence of the above approach. Occasionally, we can exploit the structure of some simple semiparametric models to produce a $\sqrt{n}$-consistent $\widehat{\theta}_n^{(0)}$, e.g., difference based estimate in partly linear model [46]. However, a general strategy is to conduct a search of $\widehat{S}_n(\theta)$ at finitely many θ-value and use the maximizer as $\widehat{\theta}_n^{(0)}$, e.g., [20, 44]. Our first contribution is to provide sufficient conditions under which the above grid search will produce $\widehat{\theta}_n^{(0)}$ with proper convergence rates. When the dimension or support of θ is large, $\widehat{\theta}_n^{(0)}$ may have the slower than root-n rate. Some examples will be given to illustrate that point. This well motivates our second contribution, which is particularly useful when $\widehat{\theta}_n^{(0)}$ has the sub-optimal rate, in the below.

Our main contribution is to answer the title of this paper from a theoretical point of view. The offered theoretical suggestions are important since k^* or ϵ is arbitrarily chosen in practice, and they may also be used to further reduce computational cost of the bootstrap inferences for semiparametric models, see [3, 12]. Specifically, we provide a formula in calculating the minimal number of iterations k^* needed to produce a semiparametric efficient $\widehat{\theta}_n^{(k^*)}$. We discover that (a) k^* depends on the convergence rates of $\widehat{\theta}_n^{(0)}$ and $\widehat{\eta}(\theta)$; (b) more than k^* iterations, i.e., k, will not change the limiting distribution of $\widehat{\theta}_n^{(k)}$, but will improve its higher order asymptotic efficiency; (c) k^* iterations are also sufficient for recovering the estimation sparsity under high dimensional data. Surprisingly, the value of k^* could be quite large when η is estimated at a very slow rate. For example, we need 8 iterations in conditionally exponential models, see Table 3. The above conclusions hold, in particular, when η is not root-n estimable, and apply to semiparametric models estimated under various regularizations, e.g., kernel or penalized estimation.

The last contribution is the novel construction of $\widehat{\theta}_n^{(k)}$. In contrast with the literature, i.e., [5, 36, 37], our construction does not require knowing the form of the implicitly defined efficient score function or applying the sample splitting (drop-one-out) technique. We also note that the sample splitting may be avoided by a conditioning argument in [21].

Our results can be useful in practice to obtain an estimator that has the same desirable

higher order asymptotic efficiency properties as the semiparametric efficient estimate without having to compute it. Our results also can be useful to obtain a root-n consistent estimator starting from an initial estimator that is only n^{ψ} consistent for some $\psi \in (0, 1/2)$. On the other hand, one has to be careful in applying the theoretical results of the paper, because we focus more on the theoretical explorations in general semiparametric context. In particular semiparametric models, one may make necessary adaptations on the grid search of $\widehat{\theta}_n^{(0)}$ or the construction of $\widehat{\theta}_n^{(k)}$ to better capture model features so that the finite sample behaviors become better. Due to the space limitation, we only consider the NR algorithm in this paper, but notice that the extensions to the slight modifications of NR are possible by considering the discussions in Page 534 of [34].

Section 2 provides some necessary background material on the semiparametric estimation. In Section 3, we propose two grid search algorithms for identifying the initial estimate whose convergence rate will be rigorously proven. In Section 4, we consider the semiparametric maximum likelihood estimation in which $\widehat{S}_n(\theta)$ is the possibly non-smooth profile likelihood (1). In Section 5, we consider the semiparametric estimation under two types of regularization, i.e., kernel estimation and penalized estimation, in which $\widehat{S}_n(\theta)$ is smooth. In that section, we also consider the *sparse* and efficient estimation of the partial linear models as an important application of penalized estimation. Several semiparametric models ranging from survival models, mixture models to conditionally exponential models are treated to illustrate the applicability of our theories. All the proofs are postponed to the Appendix.

2. Preliminary

We assume that the data $X_1, \ldots, X_n$ are i.i.d. throughout the paper. In what follows, we first briefly review the concepts of the efficient score function and the least favorable curve (LFC), and then relate the estimation of LFC to that of θ as discussed in [39]. Unless otherwise specified, the notation E is reserved for the expectation taken under (θ_0, η_0).

The score functions for θ and η are defined as, respectively,

$$\begin{aligned} \dot{\ell}_0(X_i) &= \frac{\partial}{\partial \theta} \log lik(X_i; \theta_0, \eta_0), \\ A_{\theta_0,\eta_0} h(X_i) &= \frac{\partial}{\partial t}|_{t=0} \log lik(X_i; \theta_0, \eta(t)), \end{aligned} \tag{3}$$

where h is a "direction" along which $\eta(t) \in \mathcal{H}$ approaches η_0 as $t \to 0$. $A_{\theta_0,\eta_0} : \mathbf{H} \mapsto L_2^0(P_{\theta_0,\eta_0})$ is the score operator for η, where $\mathbf{H}$ is some closed and linear direction set.

The efficient score function $\widetilde{\ell}_0$ is defined as the residual of the projection of $\dot{\ell}_0$ onto the tangent space $\mathcal{T}$, which is defined as the closed linear span of the tangent set $\{A_{\theta_0,\eta_0}H = (A_{\theta_0,\eta_0}h_1, \ldots, A_{\theta_0,\eta_0}h_d)' : h_j \in \mathbf{H}\}$. Therefore, we can write the efficient score function at (θ_0, η_0) as

$$\widetilde{\ell}_0 = \dot{\ell}_0 - \Pi_0\dot{\ell}_0, \tag{4}$$

where $\Pi_0\dot{\ell}_0 = \arg\min_{t\in\mathcal{T}} E\|\dot{\ell}_0 - t\|^2$. The variance of $\widetilde{\ell}_0$ is defined as the efficient information matrix $\widetilde{I}_0$. The inverse of $\widetilde{I}_0$ is shown to be Cramér-Rao bound for estimating θ in the presence of an infinite dimensional η, see [6].

A main idea of estimating θ is to reduce a high dimensional semiparametric model to a low dimensional random submodel of the same dimension as θ called the least favorable submodel (LFS). The LFS can be constructed as $t \mapsto \log lik(t, \eta_*(t))$ and satisfies

$$\eta_*(\theta_0) = \eta_0. \tag{5}$$

and

$$\frac{\partial}{\partial t}|_{t=\theta_0} \log lik(t, \eta_*(t)) = \widetilde{\ell}_0 \tag{6}$$

Note that the LFS may not exist unless $\Pi_0\dot{\ell}_0$ can be expressed as a nuisance score (the tangent set is closed). In all our examples, the LFS exists or can be approximated sufficiently closely. The $\eta_*(t)$ in the LFS is called as the least favorable curve. Under regularity conditions, it is shown that

$$\eta_*(t) = \arg\sup_{\eta\in\mathcal{H}} E\log lik(t, \eta) \quad \text{for any fixed } t \in \Theta. \tag{7}$$

By (7) and standard arguments, we can establish that the maximizer of

$$S_n(\theta) \equiv \sum_{i=1}^{n} \log lik(\theta, \eta_*(\theta))(X_i)$$

is semiparametric efficient. In addition, based on (6), we can derive that

$$\widetilde{I}_0 = E\left(\frac{\partial \log lik(t, \eta_*(t))}{\partial t}|_{t=\theta_0}\right)^{\otimes 2} = -E\left(\frac{\partial^2 \log lik(t, \eta_*(t))}{\partial t^2}|_{t=\theta_0}\right). \tag{8}$$

Recall that $\widehat{S}_n(\theta) = \sum_{i=1}^n \log lik(\theta, \widehat{\eta}(\theta))(X_i)$. Define

$$\widehat{\theta}_n = \arg\sup_{\theta\in\Theta} \widehat{S}_n(\theta). \tag{9}$$

In view of the above discussions, we can show that $\widehat{\theta}_n$ is semiparametric efficient if $\widehat{\eta}(\theta)$ is a consistent estimate of $\eta_*(\theta)$. The technical derivations in the above can be referred to Section 4 of [39]. However, the form of $\widehat{\theta}_n$ depends on how we estimate the abstract $\eta_*(\theta)$ defined in (7). For example, $\widehat{\theta}_n$ is just the semiparametric MLE if $\widehat{\eta}(\theta)$ is the well defined NPMLE. When the infinite dimensional $\mathcal{H}$ is too large, we may consider estimating $\eta_*(\theta)$ under some form of regularization, e.g., penalization. It is well known that the convergence rate of $\widehat{\eta}(\theta)$ is determined by the size of $\mathcal{H}$ in terms of its entropy number and the smoothing parameters associated with regularization methods (if used), e.g., smoothing parameter in penalized estimation.

In this paper, we will consider two types of $\widehat{\theta}_n$ defined in (9) according to the way we estimate $\eta_*(\theta)$: (i) pure nonparametric estimation in Section 4; (ii) nonparametric estimation under regularization in Section 5. Define $R_n \asymp r_n$ if $r_n/M \leq R_n \leq r_n M$ for some $M \geq 1$. We use $\mathcal{N}(\theta_0)$ to denote a neighborhood of θ_0. Let v_i denote the i-th unit vector in $\mathbb{R}^d$. Define the i-th ((i,j)-th) element of a vector V (Matrix M) as V_i (M_{ij}). For a tensor $T^{(3)}(\theta)$, we define $V^T \otimes T^{(3)}(\theta) \otimes V$ as a d-dimensional vector with i-th element $V^T(\partial^2/\partial\theta^2)(\dot{T}(\theta))_i V$, where $\dot{T}(\theta)$ is the first derivative of $T(\theta)$. Denote $int[x]$ and $\widetilde{int}[x]$ as the smallest nonnegative integer $\geq x$ and $> x$, respectively. The symbols $\mathbb{P}_n$ and $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ are used for the empirical distribution and the empirical process of the observations, respectively.

3. Initial Estimate

Throughout the paper, we assume that $\widehat{\theta}_n^{(0)}$ is n^{ψ}-consistent for $0 < \psi \leq 1/2$, i.e., $\|\widehat{\theta}_n^{(0)} - \theta_0\| = O_P(n^{-\psi})$, just as the numerical result assumes the iterations commence in some neighborhood of θ_0. In this section, we prove that the grid search of $\widehat{S}_n(\theta)$ will produce such initial estimate. The proof is nontrivial since $\widehat{S}_n(\theta)$ usually has no explicit form and is possibly nonsmooth. In fact, our theoretical results on searching $\widehat{\theta}_n^{(0)}$, i.e., Theorem 1, can be applied to any objective functions satisfying the below Conditions I1-I2, and are thus of independent interest.

We first state two primary conditions I1-I2 on $\widehat{S}_n(\theta)$.

I1. [Asymptotic Uniqueness] For any random sequence $\{\widetilde{\theta}_n\} \in \Theta$,

$$[\widehat{S}_n(\widetilde{\theta}_n) - \widehat{S}_n(\widehat{\theta}_n)]/n = o_P(1) \text{ implies that } \widetilde{\theta}_n - \theta_0 = o_P(1). \tag{10}$$

I2. [Asymptotic Expansion] For any consistent $\widetilde{\theta}_n$, $\widehat{S}_n$ satisfies

$$\widehat{S}_n(\widetilde{\theta}_n) = \widehat{S}_n(\theta_0) + n(\widetilde{\theta}_n - \theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - \frac{n}{2}(\widetilde{\theta}_n - \theta_0)'\widetilde{I}_0(\widetilde{\theta}_n - \theta_0) + \Delta_n(\widetilde{\theta}_n), \tag{11}$$

where $\Delta_n(\theta) = n\|\theta - \theta_0\|^3 \vee n^{1-2v}\|\theta - \theta_0\|$, for some $1/4 < v \leq 1/2$.

Condition I1 is usually implied by the model identifiability conditions. Condition I2 is very weak since we only assume the existence of the asymptotic expansion (11) but not require the continuity of $\widehat{S}_n(\cdot)$. When $\widehat{S}_n(\cdot)$ is the possibly nonsmooth $\log pl_n(\cdot)$, I2 is implied by model Assumptions M1-M4 in Section 4, see (A.5), with v being the convergence rate of $\widehat{\eta}(\theta)$ defined in (15). As for the smooth regularized $\widehat{S}_n$, we can verify I2 using a three term Taylor expansion of $\widehat{S}_n(\cdot)$ with v being g in Condition G of Section 5. In fact, Conditions I1-I2 are very mild and can be satisfied in a wide range of semiparametric models, e.g., proportional odds model and penalized semiparametric logistic regression. See the following two examples for more details.

Now we consider two types of grid search: deterministic type and stochastic type. In the former, we form a grid of cubes with sides of length $sn^{-\psi}$ over $\mathbb{R}^d$ for some $s > 0$ and $0 < \psi \leq 1/2$, and thus obtain a set of points $\mathcal{D}_n = \{\theta_{iD}\}$ regularly spaced throughout Θ with cardinality $card(\mathcal{D}_n) \geq Cn^{d\psi}$ for some $C > 0$. The grid point maximizing $\widehat{S}_n(\theta)$ is thought of as $\widehat{\theta}_n^{(0)}$. However, this deterministic search could be very slow if the dimension d of θ is high. This motivates us to propose the stochastic search in which the search points are the realizations of some independent random variable $\bar{\theta}$ with strictly positive density around θ_0, e.g., $\bar{\theta} \sim Unif[\Theta]$. And we require that the magnitude of the stochastic search points remains n^{ψ} no matter how large the dimension d is. In theory, the stochastic grid search has significant computational savings over the deterministic alternative. In the below Theorem 1 we rigorously prove that the convergence rates of the above numerical outcomes are n^{ψ}-consistent for $0 < \psi \leq 1/2$.

THEOREM 1. *Let $\mathcal{D}_n$ be a set of points regularly spaced throughout Θ with $card(\mathcal{D}_n) \geq Cn^{d\psi}$ for some $C > 0$ and $0 < \psi \leq 1/2$. Assume that $\bar{\theta}$ is independent of the data and admits a density having support Θ and bounded away from zero in some neighborhood of θ_0. Let $\mathcal{S}_n$ be a set of realizations of $\bar{\theta}$ with $card(\mathcal{S}_n) \geq \widetilde{C}n^{\psi}$ for some $\widetilde{C} > 0$ and $0 < \psi \leq 1/2$. Suppose that Conditions I1-I2 hold, and that the parameter space Θ is compact. Then, if $\widehat{\theta}_n$ defined in (9) is consistent and $\widetilde{I}_0$ is nonsingular, we have*

$$\theta_n^D - \theta_0 = O_P(n^{-\psi}), \tag{12}$$

$$\theta_n^S - \theta_0 = O_P(n^{-\psi}), \tag{13}$$

where $\theta_n^D = \arg\max_{\theta \in \mathcal{D}_n} \widehat{S}_n(\theta)$ and $\theta_n^S = \arg\max_{\theta \in \mathcal{S}_n} \widehat{S}_n(\theta)$.

Similar theorem is also proven in Robinson (1988) but for parametric models.

In the below Cox model and semiparametric mixture model, there is no theoretically justified initial estimate available in the literature. Hence, we naturally take the above grid search to obtain $\widehat{\theta}_n^{(0)}$. In reality, due to the limited computational resources, the obtained $\widehat{\theta}_n^{(0)}$ may have *slower than root-n* rate when θ has high dimension or large support as shown in Theorem 1.

Example 1: Cox Model under Current Status Data

In the Cox proportional hazards model, the hazard function of the survival time T of a subject with covariate Z is expressed as:

$$\lambda(t|z) \equiv \lim_{\Delta \to 0} \frac{1}{\Delta} Pr(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t) \exp(\theta' z),$$

where λ is an unspecified baseline hazard function. We consider the current status data where each subject is observed at a single examination time Y to determine if an event has occurred, but the event time T cannot be known exactly. Specifically, the observed data are n realizations of $X = (Y, \delta, Z) \in R^+ \times \{0, 1\} \times R$, where $\delta = I\{T \leq Y\}$. The cumulative hazard function $\eta(y) = \int_0^y \lambda(t)dt$ is considered as the nuisance parameter. The parameter space $\mathcal{H}$ for η is restricted to a set of nondecreasing and cadlag functions on some compact interval. In this model, it is well known that both $\widehat{\eta}(\theta)$ and $\widehat{S}_n(\theta) = \log pl_n(\theta)$ have no explicit forms, and can only be calculated numerically via the iterative convex minorant algorithm, see [24]. As for the convergence rate of η, Murphy and van der Vaart (1999) showed $\|\widehat{\eta}(\widetilde{\theta}_n) - \eta_0\|_2 = O_P(\|\widetilde{\theta}_n - \theta_0\| \vee n^{-1/3})$, where $\|\cdot\|_2$ is the L_2 norm. According to (A.5) in Appendix and the above discussions, we know that I2 is satisfied with $v = 1/3$. Condition I1 is verified in Lemma 2 of [27] for this model.

Example 2: Semiparametric Mixture Model in Case-Control Studies

Roeder, Carroll and Lindsay (1996) consider the logistic regression model with a missing covariate for case-control studies. In this model, they observe two independent random samples: one complete component $Y_C = (D_C, W_C)$ and Z_C of the size n_C, and one reduced component $Y_R = (D_R, W_R)$ of the size n_R. Following the assumptions given in [35], the likelihood for $x = (y_C, y_R, z_C)$ is defined as

$$lik(\theta', \eta)(x) = p_{\theta'}(y_C|z_C)\eta\{z_C\} \int p_{\theta'}(y_R|z)d\eta(z),$$

where $d\eta$ denotes the density of η w.r.t. some dominating measure, and

$$p_{\theta'}(y|z) = \left(\frac{\exp(\gamma + \theta e^z)}{1 + \exp(\gamma + \theta e^z)}\right)^d \left(\frac{1}{1 + \exp(\gamma + \theta e^z)}\right)^{1-d} \phi_\sigma(w - \alpha_0 - \alpha_1 z),$$

where $\phi_\sigma(\cdot)$ denotes the density for $N(0,\sigma)$. The unknown parameters are $\theta' = (\theta, \alpha_0, \alpha_1, \gamma, \sigma)$ ranging over the compact $\Theta' \subset \mathbb{R}^4 \times (0,\infty)$ and the distribution η of the regression variable restricted to the set of nondegenerate probability distributions with a known compact support. In this semiparametric mixture model, we will concentrate on the regression coefficient θ, considering $\theta_2 = (\alpha_0, \alpha_1, \gamma, \sigma)$ and η as nuisance parameters. The NPMLE $\widehat{\eta}(\theta)(z)$ is a weighted average of two empirical distributions, and the log-profile likelihood defined as

$$\widehat{S}_n(\theta) = \log pl_n(\theta) = \sup_{\theta_2,\eta} \log lik_n(\theta', \eta)$$

has no explicit form. Let $(\widehat{\theta}_{2,\theta}, \widehat{\eta}(\theta))$ be the profile likelihood estimator for (θ_2, η) so that $\widehat{\theta}'_\theta = (\theta, \widehat{\theta}_{2,\theta})$. Both $\widehat{\eta}(\theta)$ and $\widehat{S}_n(\theta)$ can be computed efficiently via the iterative algorithm in Section 4 of [35], a special case of our general algorithm. Murphy and van der Vaart (1999) showed that, for any $\widetilde{\theta}_n \xrightarrow{P} \theta_0$,

$$\|\widehat{\eta}(\widetilde{\theta}_n) - \eta_0\|_{BL_1} + \|\widehat{\theta}'_{\widetilde{\theta}_n} - \theta'_0\| = O_P(|\widetilde{\theta}_n - \theta_0| \vee n^{-\frac{1}{2}}), \tag{14}$$

where $\|\cdot\|_{BL_1}$ is the weak topology. According to (A.5) in Appendix and the above discussions, we know that I2 is satisfied with $v = 1/2$. Condition I1 is verified in Lemma 3 of [27] for this model.

4. Semiparametric Maximum Likelihood Estimation

In this section, we consider the maximum likelihood estimation of θ corresponding to the case that (i) $\widehat{\eta}(\theta)$ is the NPMLE for $\eta_*(\theta)$ given any fixed θ and (ii) $\widehat{S}_n(\theta) = \log pl_n(\theta)$. The NPMLE $\widehat{\eta}(\theta)$ is well defined when η is under shape restrictions, e.g. the monotone cumulative hazard function. In general, the profile likelihood does not have a closed form, and thus can only be calculated numerically, e.g., see Examples 1-2. We first discuss the construction of $\widehat{\theta}_n^{(k)}$, and then show that the minimal number of iterations k^* is jointly determined by the convergence rates of $\widehat{\theta}_n^{(0)}$ and $\widehat{\eta}(\theta)$. In the end, our theories are applied to the previous two models.

In this section, we assume the following convergence rate Condition (15) and the LFS Conditions M1-M4 specified in Appendix. For any random sequence $\widetilde{\theta}_n \xrightarrow{P} \theta_0$, we assume

$$\|\widehat{\eta}(\widetilde{\theta}_n) - \eta_0\| = O_P(\|\widetilde{\theta}_n - \theta_0\| \vee n^{-r}), \tag{15}$$

where $\|\cdot\|$ is some norm in $\mathcal{H}$ and $1/4 < r \leq 1/2$. Of course we take the largest such r in the following and call it the convergence rate for estimating η. The above range of r holds

in regular semiparametric models, which we can define without loss of generality to be models where the entropy integral converges. Theorems 3.1-3.2 in [31] can be applied to calculate the convergence rate (15). Under the above regularity conditions, Cheng and Kosorok (2008b) showed the following second order asymptotic linear expansion result.

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{I}_0^{-1} \widetilde{\ell}_0(X_i) + O_P(n^{-2r+1/2}). \tag{16}$$

We need to estimate $\mathbb{P}_n \widetilde{\ell}_0$ and $\widetilde{I}_0$ to construct $\widehat{\theta}_n^{(k)}$ which is generated from the NR algorithm. In view of (6) and (8), we can estimate them based on the derivatives of the log-profile likelihood (the sample analog of $S_n(\theta)$) as follows

$$\left[\widehat{\ell}_n(\theta, s_n)\right]_i = \frac{\log pl_n(\theta + s_n v_i) - \log pl_n(\theta)}{n s_n}, \tag{17}$$

$$\begin{aligned} \left[\widehat{I}_n(\theta, t_n)\right]_{i,j} = & -\frac{\log pl_n(\theta + t_n(v_i + v_j)) + \log pl_n(\theta)}{n t_n^2} \\ & + \frac{\log pl_n(\theta + t_n v_i) + \log pl_n(\theta + t_n v_j)}{n t_n^2}. \end{aligned} \tag{18}$$

In the above we use the numerical derivatives since the smoothness and differentiability of $\log pl_n(\theta)$ are usually unknown. In Lemma A.1 of Appendix, we show that (17) and (18) (also called as the observed information in [31]) are indeed consistent. Thus, we can write $\widehat{\theta}_n^{(k)}$ in step (III) as

$$\widehat{\theta}_n^{(k)} = \widehat{\theta}_n^{(k-1)} + \left[\widehat{I}_n\left(\widehat{\theta}_n^{(k-1)}, t_n^{(k-1)}\right)\right]^{-1} \widehat{\ell}_n\left(\widehat{\theta}_n^{(k-1)}, s_n^{(k-1)}\right), \tag{19}$$

where step sizes $s_n^{(k-1)} \vee t_n^{(k-1)} = o(1)$. A close inspection of (19) reveals that we have constructed $\widehat{\theta}_n^{(k)}$ even without knowing the forms of $\widetilde{\ell}_0$ and $\widetilde{I}_0$.

The convergence of $\widehat{\theta}_n^{(k)}$ to $\widehat{\theta}_n$, which is exactly the maximizer of $\log pl_n(\theta)$, as $k \to \infty$ is guaranteed by the asymptotic parabolic form of $\log pl_n(\theta)$ proven in [32]. However, to figure out the minimal k^* such that $\|\widehat{\theta}_n^{(k^*)} - \widehat{\theta}_n\| = o_P(n^{-1/2})$, we need to make use of the second order asymptotic quadratic expansion of $\log pl_n(\theta)$ derived in [14] under the above regularity conditions. As seen from (19), the orders of step sizes $(s_n^{(k-1)}, t_n^{(k-1)})$ are critical in determining the convergence rate of $\widehat{\theta}_n^{(k)}$ to $\widehat{\theta}_n$, and thus need to be properly chosen at each iteration. In the below Lemma, we present the optimal step sizes, under which the fastest convergence rate is achieved, at each iteration. Denote the convergence rate of $\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|$ as $O_P(n^{-r_{k-1}})$.

LEMMA 1. *Suppose Conditions M1-M4 and (15) hold. Also suppose that the MLE $\widehat{\theta}_n$ is consistent and $\widetilde{I}_0$ is nonsingular. The convergence rate of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$ is improved through the following three stages:*

(i) $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^{3/2})$ *when* $r_{k-1} < r$ *and we choose* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-3r_{k-1}/2}, n^{-r_{k-1}/2})$;

(ii) $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^{1/2} n^{-r})$ *when* $r \le r_{k-1} < 1/2$ *and we choose* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-r-r_{k-1}/2}, n^{-r_{k-1}/2})$;

(iii) $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-r-1/4})$ *when* $r_{k-1} \ge 1/2$ *and we choose* $(s_n^{(k-1)}, t_n^{(k-1)}) \asymp (n^{-r-1/4}, n^{-r_{k-1}/2})$.

Now we present our second main theorem, i.e., Theorem 2. Let $\widehat{\theta}_n^{(0)}$ be n^{ψ}-consistent. We first show that $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-S(\psi,r,k)})$ based on which we figure out the value of k^* in (21). According to the above Lemma 1, it is easily seen that $S(1/2, r, k) = r + 1/4$ for any $1/4 < r \le 1/2$ and $k \ge 1$ (thus $k^* = 1$); and $S(1/3, 1/2, 1) = 1/2$ and $S(1/3, 1/2, k) = 3/4$ for any $k \ge 2$ (thus $k^* = 2$). This justifies the well known one-step iteration result given $\sqrt{n}$-consistent initial estimate. As discussed in Section 1, $\widehat{\theta}_n^{(0)}$ may also have the sub-optimal rate. This motivates us to give the general form of $S(\psi, r, k)$ for $\psi < 1/2$ in the below. Define, if $\widetilde{S}_1(\psi, r) \ge 1/2$,

$$S(\psi, r, k) = \begin{cases} S_1(\psi, k) & k \le K_1(\psi, r) \\ r + 1/4 & k \ge K_1(\psi, r) + 1 \end{cases},$$

where $S_1(\psi, k) = \psi(3/2)^k$, $K_1(\psi, r) = int\,[\log(r/\psi)/\log(3/2)]$ and $\widetilde{S}_1(\psi, r) = S_1(\psi, K_1(\psi, r))$, and if $r \le \widetilde{S}_1(\psi, r) < 1/2$,

$$S(\psi, r, k) = \begin{cases} S_1(\psi, k) & k \le K_1(\psi, r) \\ S_2(\widetilde{S}_1(\psi, r), r, k - K_1(\psi, r)) & K_1(\psi, r) < k \le K_1(\psi, r) + \widetilde{K}_2(\psi, r) \\ r + 1/4 & k \ge K_1(\psi, r) + \widetilde{K}_2(\psi, r) + 1 \end{cases},$$

where $S_2(\psi, r, k) = 2r + 2^{-k}(\psi - 2r)$, $K_2(\psi, r) = int[\log\{(2r - \psi)/(2r - 1/2)\}/\log 2]$ and $\widetilde{K}_2(\psi, r) = K_2(\widetilde{S}_1(\psi, r), r)$.

THEOREM 2. *Suppose that Conditions in Lemma 1 hold and proper step sizes are chosen according to Lemma 1. Let $\widehat{\theta}_n^{(k)}$ be the k-step estimator defined in (19) and $\widehat{\theta}_n^{(0)}$ be n^{ψ}-consistent for $0 < \psi \le 1/2$. Recall that $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-r_k})$. We show that r_k increases from ψ to $(r + 1/4)$ as $k \to \infty$. Specifically, we have*

$$\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-S(\psi,r,k)}). \tag{20}$$

This implies that

$$\|\widehat{\theta}_n^{(k^*)} - \widehat{\theta}_n\| = o_P(n^{-1/2}), \tag{21}$$

where $k^* = K_1(\psi, r) + \widetilde{int}[\log((2r - \widetilde{S}_1(\psi, r))/(2r - 1/2))/\log 2]$.

Interestingly, we notice that the optimal bound of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$, i.e. $O_P(n^{-r-1/4})$, is intrinsically determined by how accurately we estimate the nuisance parameter, i.e., the value of r. This bound can not be further improved unless we are willing to make stronger assumptions than M1-M4, which seem unrealistic. From the form of $S(\psi, r, k)$, we find that more accurate initial estimate leads to higher order asymptotic efficiency of $\widehat{\theta}_n^{(k)}$.

We apply Theorem 2 to the previous two examples, and the required Conditions are verified in [13, 14] for them.

Example 1: Cox Model under Current Status Data (Cont')

According to Theorem 2, we establish the following table to depict the convergence of $\widehat{\theta}_n^{(k)}$ to $\widehat{\theta}_n$ given different initial estimates until it reaches the lower bound $O_P(n^{-7/12})$.

Table 1. *Cox Model under Current Status Data* ($r = 1/3$)

	$\psi = 1/2$	$\psi = 1/3$	$\psi = 1/4$
Cox	$r_1 = 7/12$	$r_1 = 1/2, r_2 = 7/12$	$r_1 = 3/8, r_2 = 25/48, r_3 = 7/12$
Models	$k^* = 1$	$k^* = 2$	$k^* = 2$

Remark: Define $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-r_k})$.

Example 2: Semiparametric Mixture Model in Case-Control Studies (Cont')

The following Table 2 is similar as Table 1. From Table 2, we can tell that the fast convergence rate of the nuisance estimate, i.e., $r = 1/2$, leads to fast convergence of $\widehat{\theta}_n^{(k)}$ to $\widehat{\theta}_n$.

Table 2. *Semiparametric Mixture Model in Case-Control Studies* ($r = 1/2$)

	$\psi = 1/2$	$\psi = 1/3$	$\psi = 1/4$
Mixture	$r_1 = 3/4$	$r_1 = 1/2, r_2 = 3/4$	$r_1 = 3/8, r_2 = 9/16, r_3 = 3/4$
Models	$k^* = 1$	$k^* = 2$	$k^* = 2$

Remark: Define $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-r_k})$.

5. Semiparametric Estimation under Regularization

In this section, we consider the semiparametric estimation under two types of regularizations, i.e., kernel estimation and penalized estimation. In contrast with the profile likelihood estimation, the regularized $\widehat{S}_n(\theta)$ is usually differentiable although its form may vary under different regularizations. We first present a unified framework for studying $\widehat{\theta}_n^{(k)}$ when $\widehat{S}_n(\theta)$ is third order differentiable, and then present several examples corresponding to different regularizations which fit into this framework. We also discuss the variable selection in partly linear models as an extension of the penalized estimation.

In this section, we construct $\widehat{\theta}_n^{(k)}$ in step (III) as follows:

$$\widehat{\theta}_n^{(k)} = \widehat{\theta}_n^{(k-1)} + \left[\widehat{I}_n(\widehat{\theta}_n^{(k-1)})\right]^{-1} \widehat{\ell}_n(\widehat{\theta}_n^{(k-1)}), \tag{22}$$

where $\widehat{\ell}_n(\cdot) = \widehat{S}_n^{(1)}(\cdot)/n$ and

$$\widehat{I}_n(\cdot) = -\widehat{S}_n^{(2)}(\cdot)/n, \tag{23}$$

where $\widehat{S}_n^{(j)}(\cdot)$ is the j-th derivative of $\widehat{S}_n(\cdot)$. When $\widehat{S}_n^{(2)}(\theta)$ has no explicit form or is hard to compute, we may prefer constructing $[\widehat{I}_n(\theta)]_{ij}$ as

$$-n^{-1/2}\frac{[\widehat{S}_n^{(1)}(\theta + n^{-1/2}t_2 v_j)]_i - [\widehat{S}_n^{(1)}(\theta + n^{-1/2}t_1 v_j)]_i}{t_2 - t_1}, \tag{24}$$

where t_1 and t_2 $(t_1 < t_2)$ are arbitrarily fixed real numbers.

Recall that

$$S_n(\theta) = n\mathbb{P}_n \log lik(\theta, \eta_*(\theta))$$

and define $S_n^{(j)}(\cdot)$ as the j-th derivative of $S_n(\cdot)$. In view of the discussions in Section 2, i.e. (6) & (8), we expect that $\widehat{\theta}_n^{(k)}$ converges to $\widehat{\theta}_n$ if $\widehat{S}_n^{(j)}(\cdot)$ approximates $S_n^{(j)}(\cdot)$ well enough round θ_0 for $j = 1, 2, 3$. Therefore, we assume the following general condition G.

G. Assume that

$$\frac{1}{n}\widehat{S}_n^{(1)}(\theta_0) - \frac{1}{n}S_n^{(1)}(\theta_0) = O_P(n^{-2g}), \tag{25}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\widehat{S}_n^{(2)}(\theta) - \frac{1}{n}S_n^{(2)}(\theta)\right| = O_P(n^{-g}), \tag{26}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\widehat{S}_n^{(3)}(\theta)\right| = O_P(1), \tag{27}$$

where $1/4 < g \leq 1/2$.

In the kernel estimation (penalized estimation), the value of g is determined by the bandwidth order of the used kernel function (the order of the smoothing parameter). In this sense, we can think that g is a measure of the convergence rate of η as in (15). We may verify (27) by showing

$$\sup_{\theta\in\mathcal{N}(\theta_0)}\left|\frac{1}{n}\widehat{S}_n^{(3)}(\theta)-\frac{1}{n}S_n^{(3)}(\theta)\right| \quad = \quad o_P(1), \tag{28}$$

and that the class of functions $\{(\partial^3/\partial\theta^3)\log lik(x;\theta,\eta_*(\theta)) : \theta\in\mathcal{N}(\theta_0)\}$ is P-Glivenko-Cantelli and that

$$\sup_{\theta\in\mathcal{N}(\theta_0)} E\left|(\partial^3/\partial\theta^3)\log lik(X;\theta,\eta_*(\theta))\right| < \infty.$$

Now we present our third main theorem, i.e., Theorem 3. Define

$$R(\psi,g,k)=\begin{cases} R_1(\psi,g,k) & k\le L_1(\psi,g) \\ R_2(R_1(\psi,g,L_1(\psi,g)),g,k-L_1(\psi,g)) & k>L_1(\psi,g)\end{cases} \tag{29}$$

where $R_1(\psi,g,k)=(1/2-g)+2^k(\psi+g-1/2)$, $L_1(\psi,g)=int[\log(g/(g+\psi-1/2))/\log 2]$, $\widetilde{L}_1(\psi,g)=\widetilde{int}[\log(g/(g+\psi-1/2))/\log 2]$ and $R_2(\psi,g,k)=kg+\psi$.

THEOREM 3. *Suppose that Condition G holds, $\widehat{\theta}_n$ defined in (9) is consistent and $\widetilde{I}_0$ is nonsingular. We have*

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\widetilde{I}_0^{-1}\widetilde{\ell}_0(X_i)+O_P(n^{1/2-2g}). \tag{30}$$

Let $\widehat{\theta}_n^{(k)}$ be the k-step estimator defined in (22) and $\widehat{\theta}_n^{(0)}$ be n^ψ-consistent for $(1/2-g)<\psi\le 1/2$. Define $\|\widehat{\theta}_n^{(k)}-\widehat{\theta}_n\|=O_P(n^{-r_k})$. We show that r_k increases from ψ to ∞ as $k\to\infty$. Specifically, we show

$$\|\widehat{\theta}_n^{(k)}-\widehat{\theta}_n\| \quad = \quad O_P(n^{-2^k\psi}) \qquad \text{if } \widehat{I}_n(\cdot) \text{ is defined in (23)}, \tag{31}$$

$$\|\widehat{\theta}_n^{(k)}-\widehat{\theta}_n\| \quad = \quad O_P(n^{-R(\psi,g,k)}) \quad \text{if } \widehat{I}_n(\cdot) \text{ is defined in (24)}. \tag{32}$$

This implies that

$$\|\widehat{\theta}_n^{(k^*)}-\widehat{\theta}_n\|=o_P(n^{-1/2}),$$

where $k^*=\widetilde{int}[\log(1/2\psi)/\log 2]$ for (31) and $k^*=\widetilde{L}_1(\psi,g)$ for (32).

Note that (31) is a statistical counterpart to the well known quadratic convergence of the Newton-Raphson algorithm; see Page 312 of [33]. Theorems 2 and 3 imply that (i) more than k^* iterations, i.e., k, will not change the limiting distribution of $\widehat{\theta}_n^{(k)}$, but will improve its higher order asymptotic efficiency; (ii) the higher order asymptotic efficiency of $\widehat{\theta}_n^{(k)}$ is determined by how accurately η is estimated, i.e., the values of r and g; (iii) $\widehat{\theta}_n^{(k)}$ converges to $\widehat{\theta}_n$ faster when $\widehat{I}_n$ is constructed as an analytical derivative no matter whether the regularization is used or not.

REMARK 1. *Given that the initial estimate is $\sqrt{n}$ consistent, we have*

$$\begin{aligned} \|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| &= O_P(n^{-2^{k-1}}) \quad \text{if } \widehat{I}_n(\cdot) \text{ is constructed as in (23),} \\ \|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| &= O_P(n^{-(1/2+kg)}) \quad \text{if } \widehat{I}_n(\cdot) \text{ is constructed as in (24)} \end{aligned}$$

based on Theorem 3. This implies $k^* = 1$.

REMARK 2. *Theorems 2-3 together with the previous Theorem 1 offer rigorous statistical analysis for the presented semiparametric estimation approach. Those theorems indicate a tradeoff between the computational cost of searching for an initial estimate, i.e. card$(\mathcal{D}_n)$ or card$(\mathcal{S}_n)$, and that of generating an efficient estimate, i.e., k^*.*

A by-product of Theorem 3 is the application to the parametric models, i.e., η is known. In this case, $\widehat{S}_n(\theta)$ becomes $\ell_\theta(X) = \log lik(\theta; X)$, and we simplify the general Condition G to the following Conditions P1-P2. Denote the first, second and third derivative of $\ell_\theta(\cdot)$ w.r.t. θ as $\dot{\ell}_\theta(\cdot)$, $\ddot{\ell}_\theta(\cdot)$ and $\ell_\theta^{(3)}(\cdot)$, respectively. The information matrix at θ_0 is defined as I_0.

P1. $\dot{\ell}_\theta(\cdot)$ and $\ddot{\ell}_\theta(\cdot)$ are absolutely continuous in θ.

P2. There exists a $\delta > 0$ such that, for any $|t| \leq \delta$,

$$E\left[\ell_{\theta_0+t}^{(i+1)}(X_1)\right]^2 \leq K_i \quad \text{for some finite constant } K_i, \tag{33}$$

where $i = 1, 2$.

We can easily prove Corollary 1 by following similar analysis in Theorem 3 and considering Lemma A.4. Thus, its proof is skipped.

COROLLARY 1. *Suppose that Conditions P1 & P2 hold. Also suppose that the parametric MLE $\widehat{\theta}_n$ is consistent and I_0 is nonsingular. Let $g = 1/2$. Then all the conclusions for $\widehat{\theta}_n$ and $\widehat{\theta}_n^{(k)}$ in Theorem 3 hold for the parametric estimation.*

The above corollary generalizes the one/two-step parametric estimation results in [25]. Comparing Theorem 3 with Corollary 1, we notice that $\widehat{\theta}_n^{(k)}$ converges to $\widehat{\theta}_n$ at a slower rate in semiparametric models. By comparing Lemmas A.3 and A.4, we know that this is due to the presence of an infinite dimensional η.

REMARK 3. *We would like to mention that the regularized $\widehat{S}_n(\theta)$ may not be differentiable in some semiparametric models, e.g., the penalized estimation of partly linear models under current status data studied in [15]. In such cases, we can take the discretization approach to construct $\widehat{\theta}_n^{(k)}$ as in the profile likelihood framework, i.e., (19), and obtain similar results as in Theorem 2 if we can prove that the non-smooth $\widehat{S}_n(\theta)$ share the same higher order quadratic expansion as $\log pl_n(\theta)$. Indeed, Cheng and Kosorok (2009) have proven such results for the non-smooth regularized $\widehat{S}_n(\theta)$ under weaker conditions. See [10] for more elaborations.*

5.1. Kernel Estimation in Semiparametric Models

In this subsection, we consider the kernel estimation in semiparametric models. Due to its simple form, the kernel estimate of η and the related iterative approach of estimating θ are widely used in semiparametric models, e.g., [2, 43]. In particular, the kernel approach is proven to be a powerful inferential tool for the class of conditionally parametric models (CPM), see [39, 40]. Thus we will focus on the class of CPM although our conclusions can be extended to more general class of semiparametric models by incorporating the results in [2]. Under kernel estimation, k^* is shown to depend on the order of bandwidth used in the kernel function.

The class of CPM was first introduced by Severini and Wong (1992) and further generalized to the quasi-likelihood framework by Severini and Staniswalis (1994). Specifically, we observe $X = (Y, W, Z)$ such that the distribution of Y conditional on partitioned covariates $W = w$ and $Z = z$ is parameterized by a finite dimensional parameter $\phi = (\theta, \lambda_z)$, where $\lambda_z \in H \subset \mathbb{R}$ depends on the value of z as a function $\eta(z)$. The joint distribution of (W, Z) is assumed to be independent of ϕ. Thus, this semiparametric model has the log-likelihood $\log lik(X; \theta, \eta(z))$ and is called conditionally parametric. The practical performance of the iterative estimation procedure (I)-(IV) for the CPM is extensively studied in [40].

We assume that $\eta(z) \in \mathcal{H} = \{h \in C^2(\mathcal{Z}) : h(z) \in \text{interior}(H) \text{ for all } z \in \mathcal{Z}\}$. An important feature of CPM is that its least favorable curve can be expressed as (see [39]

for details)

$$\eta_*(\theta)(z) = \arg \sup_{\eta \in C^2[0,1]} E[\log lik(X;\theta,\eta)|Z=z], \tag{34}$$

and thus its kernel estimate is written as

$$\widehat{\eta}(\theta)(z) = \arg \sup_{\eta \in C^2[0,1]} \sum_{i=1}^{n} \log lik(X_i;\theta,\eta(Z_i)) K\left(\frac{z-Z_i}{b_n}\right), \tag{35}$$

where $K(\cdot)$ is a kernel with the bandwidth $b_n \to 0$. For example, if $(Y|w=W, Z=z) \sim N(\theta' w, \eta(z))$, then we have

$$\begin{aligned} \widehat{\eta}(\theta)(z) &= \frac{\sum_{i=1}^n (Y_i - \theta' W_i)^2 K((z-Z_i)/b_n)}{\sum_{i=1}^n K((z-Z_i)/b_n)}, \\ \widehat{S}_n(\theta) &= -\frac{1}{2}\sum_{i=1}^{n} \frac{(Y_i - \theta' W_i)}{\widehat{\eta}(\theta)(Z_i)} - \frac{1}{2}\sum_{i=1}^{n} \log \widehat{\eta}(\theta)(Z_i). \end{aligned} \tag{36}$$

Although $\widehat{\eta}(\theta)$ (and thus $\widehat{S}_n(\theta)$) solved from (35) generally has no explicit form, based on (35) we can control the asymptotic behaviors of $\widehat{\eta}(\theta)$ (and thus $\widehat{S}_n(\theta)$) by assuming proper kernel conditions, see the below Example 3.

By exploiting the parametric structure of CPM, we will show $\widehat{S}_n(\theta)$ satisfies the general Condition G under the below Conditions K1-K2 and C1-C2.

K1. For arbitrary $\theta_1 \in \Theta$ and $\lambda_1 \in H$, if $\theta \neq \theta_1$, then $E_{\theta_1,\lambda_1} \log lik(X;\theta,\lambda) < E_{\theta_1,\lambda_1} \log lik(X;\theta_1,\lambda_1)$;

K2. Assume that

$$E\left\{ \sup_{(\theta,\lambda)\in\Theta\times H} \left| \frac{\partial^{r+s} \log lik(X;\theta,\lambda)}{\partial\theta^r \partial\lambda^s} \right|^2 \right\} < \infty \tag{37}$$

for all $r, s = 0, \ldots, 4$ and $r+s \leq 4$.

Similar identifiability Condition K1 and smoothness Condition K2 are also used in [39]. Our next conditions C1-C2 are concerned about the smoothness and convergence rate of $\eta_*(\theta)$ and $\widehat{\eta}(\theta)$. We denote the derivative of $\eta_*(\theta)$ ($\widehat{\eta}(\theta)$) w.r.t. θ as $\eta_*^{(s)}(\theta)$ ($\widehat{\eta}^{(s)}(\theta)$), and their values at θ_0 as $\eta_{*0}^{(s)}$ ($\widehat{\eta}_0^{(s)}$).

C1. Assume that, for all $r, s = 0, 1, 2, 3$ and $r+s \leq 3$,

$$\frac{\partial^{r+s}}{\partial z^r \partial\theta^s}\eta_*(\theta)(z) \text{ and } \frac{\partial^{r+s}}{\partial z^r \partial\theta^s}\widehat{\eta}(\theta)(z)$$

exist and $\sup_{\theta\in\mathcal{N}(\theta_0)} \|\eta_*^{(s)}(\theta)\|_\infty < \infty$.

C2. Assume that

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \|\widehat{\eta}^{(s)}(\theta) - \eta_*^{(s)}(\theta)\|_\infty = O_P(n^{-g}) \quad \text{for } s = 0, 1, 2, \tag{38}$$

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \|\widehat{\eta}^{(3)}(\theta) - \eta_*^{(3)}(\theta)\|_\infty = o_P(1), \tag{39}$$

$$\left\| \frac{\partial}{\partial z}\widehat{\eta}_0(z) - \frac{\partial}{\partial z}\eta_{*0}(z) \right\|_\infty = o_P(n^{-\delta}), \tag{40}$$

$$\left\| \frac{\partial}{\partial z}\widehat{\eta}_0^{(1)}(z) - \frac{\partial}{\partial z}\eta_{*0}^{(1)}(z) \right\|_\infty = o_P(n^{-\delta}). \tag{41}$$

for some $g \in (1/4, 1/2]$ and $(2g - 1/2) \leq \delta \leq g$.

In view of (34)-(35), we can verify C2 by applying the kernel theories under some proper kernel conditions and K1-K2. For example, in the below Lemma 2, we show that the convergence rate of the kernel estimate in (38), which determines the value of g in (25)-(26), relies on the order of bandwidth b_n used in (35). Note that Condition C2 implies (15) assumed for the NPMLE since

$$\begin{aligned} \|\widehat{\eta}(\widetilde{\theta}_n) - \eta_0\| &\leq \|\widehat{\eta}(\widetilde{\theta}_n) - \widehat{\eta}(\theta_0)\|_\infty + \|\widehat{\eta}(\theta_0) - \eta_*(\theta_0)\|_\infty \\ &\leq O_P(\|\widetilde{\theta}_n - \theta_0\| \vee n^{-g}) \end{aligned}$$

by the construction that $\eta_*(\theta_0) = \eta_0$, C1-C2 and (38). Thus, our conditions K1-K2 and C1-C2 are generally stronger than M1-M4 and (15) since the semiparametric models under consideration have the assumed parametric structure.

THEOREM 4. *Assuming that Conditions K1-K2 and C1-C2 hold, then the Condition G required in Theorem 3 is satisfied for the kernel estimation in conditionally parametric models.*

The consistency of $\widehat{\theta}_n$ required in Theorem 3 can be established if we further require the global condition $\sup_{\theta\in\Theta} \|\widehat{\eta}(\theta) - \eta_*(\theta)\|_\infty \to 0$, see Proposition 1 of [39]. In the third example, we apply Theorems 1 and 4 to a subclass of CPM, called conditionally exponential models (CEM), in which $\widehat{\eta}(\theta)$ has a closed-form. This makes the verifications of C1-C2 much easier. The relation between k^* and the order of b_n in (35) is clearly specified in the below example. We may also apply our theories to the more complicated semiparametric transformation model, i.e., [28].

Example 3. Conditionally Exponential Models

In CEM, there exists a function $\psi_\theta(\cdot)$ such that the conditional distribution of $\psi_\theta(Y,W)$ given $Z=z$ does not depend on θ and forms an exponential family. And its log-likelihood can be expressed as

$$\log lik(X;\theta,\eta) = \psi_\theta(Y,W)T(\eta(Z)) - A(\eta(Z)) + S(\psi_\theta(Y,W))$$

for some functions T, A and S. Some simple algebra gives that

$$\widehat{\eta}(\theta)(z) = \rho\left(\frac{\sum_{i=1}^n \psi_\theta(Y_i,W_i)K((z-Z_i)/b_n)}{\sum_{i=1}^n K((z-Z_i)/b_n)}\right), \tag{42}$$

where $\eta = \rho\{E_{\theta,\eta}(\psi_\theta(Y,W))\}$. In the previous conditional normal model, we have $\psi_\theta(Y,W) = (Y-\theta' W)^2$ and $\rho(t)=t$. Another example is that $(Y|W=w,Z=z) \sim \text{Exp}(0,\exp(\theta' w+\eta(z)))$ in which $\psi_\theta(Y,W) = Y\exp(-\theta' W)$ and $\rho(t)=\log t$.

We first apply Theorem 1 to obtain $\widehat{\theta}_n^{(0)}$. Condition I1 can be verified by adapting the consistency proof of $\widehat{\theta}_n$ in [39], see its Proposition 1. Condition I2 just follows from Condition G with v being g given in (43) as discussed in Section 3. We next discuss how to verify K1-K2 & C1-C2 in Theorem 4. Conditions K1-K2 are easily verified when $\Theta \times H$ is assumed to be compact. However, we need the following Lemma to verify Conditions C1-C2. Let $\psi_\theta^{(j)}(\cdot)$ be $(\partial^j/\partial\theta^j)\psi_\theta(\cdot)$ and $f_{\theta j}(\cdot|z)$ be its conditional density. Denote $f(z)$ as the marginal density of Z. Let M be a compact set so that $m_\theta(z) \equiv E[\psi_\theta(Y,W)|Z=z] \in \text{int}(M)$ for all z,θ.

LEMMA 2. *Assume the following conditions hold:*

(a) $E\{\sup_\theta |\psi_\theta^{(j)}|\} < \infty$ *for* $j=0,1,2,3$*;*
(b) For some even integer $q \geq 10$, $\sup_\theta E\{|\psi_\theta^{(j)}|^q\} < \infty$ *for* $j=0,1,2,3$*;*
(c) $\sup_\theta \sup_x |f_{\theta j}^{(r)}(y,w|z)| < \infty$ *for* $j=0,1,2$ *and* $r=0,\ldots,4$*;*
(d) $\sup_z |f^{(r)}(z)| < \infty$ *for* $r=0,\ldots,4$*;*
(e) $0 < \inf_z f(z) \leq \sup_z f(z) < \infty$*;*
(f) $\sup_{m\in M} |\rho^{(j)}(m)| < \infty$ *for* $j=0,\ldots,4$.

Suppose that the kernel function $K(\cdot)$ *in (42) satisfies*

$$\int K(u)du = 1,\ \int uK(u)du = 0,\ \int u^2K(u)du < \infty \text{ and } \sup_u |K^{(r)}(u)| < \infty \text{ for } r=0,\ldots,4.$$

Condition C1 holds under the above conditions. If we choose $b_n \asymp n^{-\alpha}$ *for* $1/8 < \alpha <$

$(q-2)/(4q+16)$, *then Condition C2 is satisfied with*

$$g = 2\alpha \wedge \left(\frac{q}{2q+4} - \frac{\alpha(q+4)}{q+2} - \epsilon \right), \tag{43}$$

$$\delta = \frac{q}{2q+4} - \frac{\alpha(2q+6)}{q+2} - 2\epsilon \tag{44}$$

for any $\epsilon > 0$.

The above Lemma specifies the relation between the bandwidth order α in the kernel estimation (42) and k^* in Theorem 3. By some algebra, we can verify that $g \in (1/4, 1/2]$ and $(2g - 1/2) \leq \delta \leq g$ given the above range of α and q. We want to point out that the convergence rates of $\widehat{\eta}(\theta)$ (and its derivatives) may be improved, i.e., larger value of g, under more restrictive kernel conditions, see [2, 42].

Now we are ready to apply Theorem 4 and Lemma 2 to the previous conditional normal (exponential) example, in which q is shown to be arbitrarily large and M is chosen as a sufficiently large compact subset of $(0, \infty)$. For simplicity, in the below table, we assume that $q = 28$, $b_n \asymp n^{-1/5}$, $\epsilon = 1/600$ such that $g = 151/600 > 1/4$ and $\delta = 1/20$ according to (43)-(44).

Table 3. *Conditional Normal (Exponential) Model* ($g = 151/600$)

	$\psi = 1/2$	$\psi = 1/3$
Construction I	$r_1 = 1$	$r_1 = 2/3$
	$k^* = 1$	$k^* = 1$
Construction II	$r_1 = 451/600$	$r_1 = 251/600, r_2 = 353/600$
	$k^* = 1$	$k^* = 2$

	$\psi = 1/4$
Construction I	$r_1 = 1/2, r_2 = 1$
	$k^* = 2$
Construction II	$r_1 = 151/600, r_2 = 153/600, r_3 = 157/600, r_4 = 165/600$
	$r_5 = 181/600, r_6 = 213/600, r_7 = 277/600, r_8 = 405/600$
	$k^* = 8$

Remark: ψ: convergence rate of $\widehat{\theta}_n^{(0)}$; r_k: Define $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-r_k})$; Construction I: $\widehat{I}_n$ is constructed by (23); Construction II: $\widehat{I}_n$ is constructed by (24).

5.2. Penalized Estimation in Semiparametric Models

In many semiparametric models involving a smooth nuisance parameter, it is often convenient and beneficial to perform estimation using penalization, e.g., [29, 41]. Under regularity conditions, penalized semiparametric log-likelihood estimation can yield fully efficient estimates for θ, see (51). In penalized estimation framework, the value of k^* is shown to relate to the order of the smoothing parameter λ_n. A surprising result we have is that k^* iterations are also sufficient for recovering the estimation sparsity in high dimensional data, see the below partly linear example.

In this subsection, we assume that η belongs to the Sobolev class of functions $\mathcal{H}_k \equiv \{\eta : J^2(\eta) = \int_{\mathcal{Z}} (\eta^{(k)}(z))^2 dz < \infty\}$, where $\eta^{(j)}$ is the j-th derivative of η and $\mathcal{Z}$ is some compact set on the real line. The penalized log-likelihood in this context is defined as

$$\log lik_{\lambda_n}(\theta, \eta) = n\mathbb{P}_n \log lik(\theta, \eta) - n\lambda_n^2 J^2(\eta), \tag{45}$$

where λ_n is a smoothing parameter. We assume the following bounds for λ_n:

$$\lambda_n = o_P(n^{-1/4}) \text{ and } \lambda_n^{-1} = O_P(n^{k/(2k+1)}). \tag{46}$$

In practice, λ_n can be obtained by cross-validation [45]. Here, the regularized $\widehat{S}_n(\theta)$ becomes the log-profile penalized likelihood $\widehat{S}_{\lambda_n}(\theta)$:

$$\widehat{S}_{\lambda_n}(\theta) = \log_{\lambda_n}(\theta, \widehat{\eta}_{\lambda_n}(\theta)), \tag{47}$$

where $\widehat{\eta}_{\lambda_n}(\theta) = \arg\sup_{\eta \in \mathcal{H}_k} \log lik_{\lambda_n}(\theta, \eta)$ for any fixed θ and λ_n. We define the penalized estimate as $\widehat{\theta}_{\lambda_n}$.

The construction of the k-step penalized estimate $\widehat{\theta}^{(k)}_{\lambda_n}$ follows from (22) just with the change of $\widehat{S}_n(\cdot)$ to $\widehat{S}_{\lambda_n}(\cdot)$. For the penalized estimation, we need to slightly modify Condition G as follows:

G'. Assume that, for some constant c,

$$\frac{1}{n}\widehat{S}^{(1)}_{\lambda_n}(\theta_0) - c\mathbb{P}_n\widetilde{\ell}_0 = O_P(\lambda_n^2), \tag{48}$$

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left| \frac{1}{n}\widehat{S}^{(2)}_{\lambda_n}(\theta) + c\widetilde{I}_0 \right| = O_P(\lambda_n \vee \|\theta - \theta_0\|), \tag{49}$$

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left| \frac{1}{n}\widehat{S}^{(3)}_{\lambda_n}(\theta) \right| = O_P(1). \tag{50}$$

It is easy to verify Condition G' if $\widehat{\eta}_{\lambda_n}(\theta)$ has an explicit expression and $\log lik_{\lambda_n}(\theta,\eta)$ is smooth w.r.t. (θ,η), see the below example 4. We also want to point out that Condition G' is relaxable to a large extent, see Remark 3. For example, rather than the explicit form of $\widehat{\eta}_{\lambda_n}$, we may only require $\widehat{\eta}_{\lambda_n}$ satisfying $\|\widehat{\eta}_{\lambda_n}(\widetilde{\theta}_n)-\eta_0\| = O_P(\|\widetilde{\theta}_n-\theta_0\|\vee\lambda_n)$ for any consistent $\widetilde{\theta}_n$.

In view of (6) and (8), we can prove Theorem 5 similarly as Theorem 3.

THEOREM 5. *Suppose Condition G' holds, the penalized MLE $\widehat{\theta}_{\lambda_n}$ is consistent and $\widetilde{I}_0$ is nonsingular. We have*

$$\sqrt{n}(\widehat{\theta}_{\lambda_n}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{I}_0^{-1}\widetilde{\ell}_0(X_i) + O_P(\sqrt{n}\lambda_n^2). \tag{51}$$

Define $g=\max\{g': \lambda_n = O_P(n^{-g'})\}$, and thus $1/4<g\le k/(2k+1)$ based on Condition (46). Construct $\widehat{\theta}^{(k)}_{\lambda_n}$ as in (22) with the change of $\widehat{S}_n(\cdot)$ to $\widehat{S}_{\lambda_n}(\cdot)$. Then all the conclusions for $\widehat{\theta}^{(k)}_n$ in Theorem 3 also hold for $\widehat{\theta}^{(k)}_{\lambda_n}$.

The above asymptotic linear expansion (51) was also derived in [15] but under very different conditions. Theorem 5 implies that k^* depends on the order of the smoothing parameter λ_n, i.e., the value of g, see (32). Because of the duality between the penalized estimation and sieve estimation, we expect that the above conclusions also hold for the semiparametric sieve estimation, see [9]. For example, when η_0 is estimated in the form of B-spline (local polynomial) as in [23] ([8, 19]) , k^* may rely on the growth rate of the number of basis functions (the order of bandwidth in the kernel function). The detailed theoretical exploration towards this direction is beyond the scope of this article.

We next apply Theorem 5 to the following partly linear models under high dimensional data. Interestingly, we discover that one step iteration is sufficient for achieving the semiparametric estimation efficiency and recovering the estimation sparsity simultaneously.

Example 4. Sparse and Efficient Estimation of Partial Spline Model

The partial smoothing spline represents an important class of semiparametric models under penalized estimation. In particular, we consider

$$Y = W'\theta + \eta(Z) + \epsilon, \tag{52}$$

where $\eta\in\mathcal{H}_k$ and $0\le Z\le 1$. For simplicity, we assume that $\epsilon \overset{iid}{\sim} N(0,\sigma^2)$ and is independent of (W,Z). The normality of ϵ can be relaxed to the sub-exponential tail condition. In this example, we assume that some components of θ_0 are exactly zero

which is common for high dimensional data. It is well known that effective variable selection in semiparametric models could greatly improve their prediction accuracy and interpretability, e.g., [7, 16]. To achieve the estimation efficiency and recover sparsity of θ, Cheng and Zhang (2010) proposed the following double penalty estimation approach for (52). Specifically, they define $(\widehat{\theta}_{\lambda_n}, \widehat{\eta}_{\lambda_n})$ as the minimizer of

$$n\mathbb{P}_n(Y - W'\theta - \eta(Z))^2 + n\lambda_n^2 J^2(\eta) + n\tau_n^2 \sum_{j=1}^{d} \frac{|\theta_j|}{|\widetilde{\theta}_j|^\gamma}, \tag{53}$$

where γ is a fixed positive constant, $\widetilde{\theta} = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_d)'$ is the consistent initial estimate and τ_n is the second smoothing parameter for the purpose of sparsity, over $\Theta \times \mathcal{H}_k$.

We will show that $\widehat{\theta}^{(1)}_{\lambda_n}$ possesses the same *semiparametric oracle property*, whose definition is given below, as $\widehat{\theta}_{\lambda_n}$. The standard smoothing spline theory suggests that

$$\widehat{\eta}_{\lambda_n}(\theta)(\mathbf{z}) = A(\lambda_n)(\mathbf{y} - \mathbf{w}\theta), \tag{54}$$

where $\widehat{\eta}_{\lambda_n}(\theta)(\mathbf{z}) = (\widehat{\eta}_{\lambda_n}(\theta)(z_1), \ldots, \widehat{\eta}_{\lambda_n}(\theta)(z_n))'$, $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{w} = (w_1', \ldots, w_n')'$. The expression of the $n \times n$ influence matrix $A(\lambda_n)$ can be found in [22]. Therefore, $\widehat{\eta}_{\lambda_n}(\theta)$ is a natural spline of order $(2k-1)$ with knots on z_i's for any fixed θ. Plugging (54) back to (53), we have

$$\widehat{S}_{\lambda_n}(\theta) = \widetilde{S}_{\lambda_n}(\theta) + n\tau_n^2 \sum_{j=1}^{d} \frac{|\theta_j|}{|\widetilde{\theta}_j|^\gamma}, \tag{55}$$

where

$$\widetilde{S}_{\lambda_n}(\theta) = (\mathbf{y} - \mathbf{w}\theta)'[I - A(\lambda_n)](\mathbf{y} - \mathbf{w}\theta) \tag{56}$$

and I is the identity matrix of size n. When $\tau_n = 0$, the minimizer of (53) becomes the partial smoothing spline, and we denote it as $(\widetilde{\theta}_{\lambda_n}, \widetilde{\eta}_{\lambda_n})$. Note that $\widetilde{\theta}_{\lambda_n}$ has a simple analytic form as $\widetilde{\theta}_{\lambda_n} = [\mathbf{w}'(I - A(\lambda_n))\mathbf{w}]^{-1}\mathbf{w}'[I - A(\lambda_n)]\mathbf{y}$. However, $\widehat{\theta}_{\lambda_n}$ as the minimizer of $\widehat{S}_{\lambda_n}(\theta)$ does not have an explicit solution form, and has to be iteratively computed using software like Quadratic Programming or LARS [18], see Section 4 of [16]. Specifically, based on (22)-(23), we construct $\widehat{\theta}^{(1)}_{\lambda_n}$ as follows:

$$\widehat{\theta}^{(1)}_{\lambda_n} = \widehat{\theta}^{(0)}_{\lambda_n} + \left[\frac{\mathbf{w}'(I - A(\lambda_n))\mathbf{w}}{n}\right]^{-1} \left[\frac{\mathbf{w}'(I - A(\lambda_n))(\mathbf{y} - \mathbf{w}\widehat{\theta}^{(0)}_{\lambda_n})}{n} - \frac{\tau_n^2}{2}\delta_n(\widehat{\theta}^{(0)}_{\lambda_n})\right],$$

where $\delta_n(\theta) = (sign(\theta_1)/|\widetilde{\theta}_1|^\gamma, \ldots, sign(\theta_d)/|\widetilde{\theta}_d|^\gamma)'$.

Without loss of generality, we write $\theta_0 = (\theta_1', \theta_2')'$, where θ_1 consists of all q nonzero components and θ_2 consists of the rest $(d-q)$ zero elements, and define $\widehat{\theta}_{\lambda_n} = (\widehat{\theta}'_{\lambda_n,1}\widehat{\theta}'_{\lambda_n,2})'$ accordingly. We assume that W has zero mean, strictly positive definite covariance matrix Σ and finite fourth moment. The observations z_i's (real numbers) are sorted and satisfy

$$\int_0^{z_i} u(w)dw = \frac{i}{n} \qquad \text{for } i = 1, 2, \ldots, n, \tag{57}$$

where $u(\cdot)$ is a continuous and strictly positive function. The above regularity conditions are commonly used in the literature, e.g., [17, 22], and are relaxable. For example, Condition (57) can be weakened to the case in which z_i's are sufficiently close to a sequence satisfying (57). For simplicity, we assume that $\gamma = 1$ and $\widetilde{\theta}$ is $\sqrt{n}$-consistent. In this example, $\widetilde{\theta}_{\lambda_n}$ or the difference based estimate [46], which are both known to be $\sqrt{n}$ consistent, can serve as $\widetilde{\theta}$ or $\widehat{\theta}^{(0)}_{\lambda_n}$.

In this example, we say $\widehat{\theta}_{\lambda_n}$ satisfies the *semiparametric oracle property* if

O1. $\sqrt{n}(\widehat{\theta}_{\lambda_n,1} - \theta_1) \xrightarrow{d} N(0, \sigma^2\Sigma_{11}^{-1})$, where Σ_{11} is the $q \times q$ upper-left submatrix of Σ [Semiparametric Efficiency];

O2. $\widehat{\theta}_{\lambda_n,2} = 0$ with probability tending to one [Sparsity].

It is easily shown that $\sigma^2\Sigma_{11}^{-1}$ in O1 is the semiparametric efficiency bound for θ_1 since z is assumed to be fixed.

COROLLARY 2. *If $n^{k/(2k+1)}\lambda_n \to \lambda_0 > 0$ and $n^{k/(2k+1)}\tau_n \to \tau_0 > 0$, then $\widehat{\theta}_{\lambda_n}$ is $\sqrt{n}$-consistent and satisfies the semiparametric oracle property. Given that $\widehat{\theta}^{(0)}_{\lambda_n}$ is $\sqrt{n}$-consistent, then $\|\widehat{\theta}^{(1)}_{\lambda_n} - \widehat{\theta}_{\lambda_n}\| = O_P(n^{-1})$ and $\widehat{\theta}^{(1)}_{\lambda_n}$ also enjoys the semiparametric oracle property.*

The above Corollary is a simple but interesting application of Theorem 5. We can definitely relax its conditions to the general γ and non-$\sqrt{n}$ consistent $\widehat{\theta}^{(0)}_{\lambda_n}$ in which we may require more than one iteration. The conditions on λ_n and τ_n are also chosen for simplicity of expositions and are relaxable. It is also possible to extend the conclusions of Corollary 2 to the semiparametric quasi-likelihood framework proposed in [29] after more tedious algebra.

APPENDIX

A.1. Conditions M1-M4 on the Least Favorable Submodel

The LFS in Section 4 is constructed in the following manner. We first assume the existence of a smooth map from the neighborhood of θ into $\mathcal{H}$, of the form $t \mapsto \eta_*(t;\theta,\eta)$, such that the map $t \mapsto \ell(t,\theta,\eta)(x)$ can be defined as follows:

$$\ell(t,\theta,\eta)(x) \quad = \quad \log lik(t,\eta_*(t;\theta,\eta))(x), \tag{A.1}$$

where we require $\eta_*(\theta;\theta,\eta)=\eta$ for all $(\theta,\eta)\in\Theta\times\mathcal{H}$. Thus, $\log pl_n(\theta)=\sum_{i=1}^n \ell(X_i;\theta,\theta,\widehat{\eta}(\theta))$. See [14] for similar constructions. We define $\dot{\ell}(t,\theta,\eta)$, $\ddot{\ell}(t,\theta,\eta)$ and $\ell^{(3)}(t,\theta,\eta)$ as the first, second and third derivative of $\ell(t,\theta,\eta)$ with respect to t, respectively. Also denote $\ell_{t,\theta}(t,\theta,\eta)$ as $(\partial^2/\partial t\partial\theta)\ell(t,\theta,\eta)$.

M1. We assume that the derivatives $(\partial^{l+m}/\partial t^l\partial\theta^m)\ell(t,\theta,\eta)$ have integrable envelop functions in $L_1(P)$ for $(l+m)\leq 3$, and that the Fréchet derivatives of $\eta\mapsto\ddot{\ell}(\theta_0,\theta_0,\eta)$ and $\eta\mapsto\ell_{t,\theta}(\theta_0,\theta_0,\eta)$ are bounded around η_0;

M2. $E\dot{\ell}(\theta_0,\theta_0,\eta)=O(\|\eta-\eta_0\|^2)$ for all η around η_0;

M3. $\mathbb{G}_n(\dot{\ell}(\theta_0,\theta_0,\widehat{\eta}(\widetilde{\theta}_n))-\dot{\ell}(\theta_0,\theta_0,\eta_0))=O_P(n^{-2r+1/2}\vee n^{1/2-r}\|\widetilde{\theta}_n-\theta_0\|)$ for any $\widetilde{\theta}_n\xrightarrow{P}\theta_0$;

M4. The classes of functions $\{\ddot{\ell}(t,\theta,\eta)(x):(t,\theta,\eta)\in V\}$ and $\{\ell_{t,\theta}(t,\theta,\eta)(x):(t,\theta,\eta)\in V\}$ are P-Donsker, and $\{\ell^{(3)}(t,\theta,\eta)(x):(t,\theta,\eta)\in V\}$ is P-Glivenko-Cantelli, where V is some neighborhood of $(\theta_0,\theta_0,\eta_0)$.

See Section 2.2 of [14] for the discussions on M1-M4.

A.2. Useful Lemmas

The first two Lemmas are used in the proof of Lemma 1. The Lemmas A.3, A.4, A.5 and A.6 are used in the proofs of Theorem 3, Corollary 1, Theorem 4 and Corollary 2, respectively.

LEMMA A.1. *Suppose that Conditions M1-M4 and (15) hold. If $\widetilde{\theta}_n$ is n^{ψ}-consistent,*

then we have

$$\widehat{\ell}_n(\widetilde{\theta}_n, s_n) = \mathbb{P}_n\widetilde{\ell}_0 + O_P\left(n^{-\psi} \vee |s_n| \vee \frac{g_r(n^{-\psi} \vee |s_n|)}{n|s_n|}\right), \tag{A.2}$$

$$\widehat{\ell}_n(\widehat{\theta}_n + U_n, s_n) = \widehat{\ell}_n(\widehat{\theta}_n, s_n) - \widetilde{I}_0 U_n + O_P\left(\frac{g_r(|s_n| \vee \|U_n\|) \vee n^{1/2-2r}}{n|s_n|}\right), \tag{A.3}$$

$$\widehat{I}_n(\widetilde{\theta}_n, t_n) = \widetilde{I}_0 + O_P\left(\frac{g_r(\|\widetilde{\theta}_n - \widehat{\theta}_n\| \vee |t_n|) \vee nt_n\|\widetilde{\theta}_n - \widehat{\theta}_n\| \vee n^{1/2-2r}}{nt_n^2}\right), \tag{A.4}$$

where $g_r(t) = nt^3 \vee n^{1-2r}t$ *and* $U_n = O_P(n^{-s})$ *for some* $s > 0$.

PROOF: Under the assumptions M1-M4 and (15), [14] proved the following asymptotic expansion of $\log pl_n(\bar{\theta}_n)$, where $\bar{\theta}_n$ is consistent,

$$\log pl_n(\bar{\theta}_n) = \log pl_n(\theta_0) + (\bar{\theta}_n - \theta_0)' \sum_{i=1}^{n} \widetilde{\ell}_0(X_i) - \frac{n}{2}(\bar{\theta}_n - \theta_0)'\widetilde{I}_0(\bar{\theta}_n - \theta_0) + O_P\left(g_r(\|\bar{\theta}_n - \theta_0\|)\right), \tag{A.5}$$

$$\log pl_n(\bar{\theta}_n) = \log pl_n(\widehat{\theta}_n) - \frac{1}{2}n(\bar{\theta}_n - \widehat{\theta}_n)'\widetilde{I}_0(\bar{\theta}_n - \widehat{\theta}_n) + O_P\left(g_r(\|\bar{\theta}_n - \widehat{\theta}_n\|) \vee n^{1/2-2r}\right), \tag{A.6}$$

We first prove (A.3). (A.6) implies that

$$\log pl_n(\widehat{\theta}_n + V_n + s_n v_i) = \log pl_n(\widehat{\theta}_n) - \frac{n}{2}(V_n + s_n v_i)'\widetilde{I}_0(V_n + s_n v_i) + O_P(g_r(|s_n| \vee \|V_n\|) \vee n^{1/2-2r}),$$

$$\log pl_n(\widehat{\theta}_n + V_n) = \log pl_n(\widehat{\theta}_n) - \frac{n}{2}V_n'\widetilde{I}_0 V_n + O_P(g_r(\|V_n\|) \vee n^{1/2-2r}),$$

for any random vector $V_n = o_P(1)$ and $s_n \xrightarrow{P} 0$. Combining the above two expansions and (17), we have

$$[\widehat{\ell}_n(\widehat{\theta}_n + V_n, s_n)]_i = -\frac{s_n}{2}v_i'\widetilde{I}_0 v_i - v_i'\widetilde{I}_0 V_n + O_P\left(\frac{g_r(|s_n| \vee \|V_n\|) \vee n^{1/2-2r}}{n|s_n|}\right).$$

By taking $V_n = 0$ and U_n, respectively, in the above equation, we have proved (A.3). Following similar analysis in the above, (17) & (A.5) yield (A.2), and (18) & (A.6) yield (A.4). This completes the whole proof. □

LEMMA A.2. *Suppose that Conditions M1-M4 and (15) hold. If*

$$\widehat{I}_n(\widehat{\theta}_n^{(k-1)}, t_n) - \widetilde{I}_0 = O_P(r_n^{(k-1)}), \tag{A.7}$$

then we have $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| =$

$$O_P\left(|s_n| \vee \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\| r_n^{(k-1)} \vee \frac{g_r(|s_n| \vee \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|) \vee n^{1/2-2r}}{n|s_n|}\right) \tag{A.8}$$

for $k = 1, 2, \ldots$.

PROOF: Based on (19), we have

$$\begin{aligned}\widehat{I}_n(\widehat{\theta}_n^{(k-1)}, t_n)\sqrt{n}(\widehat{\theta}_n^{(k)} - \widehat{\theta}_n) &= \left[\sqrt{n}\widehat{I}_n(\widehat{\theta}_n^{(k-1)}, t_n)(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n)\right] + \sqrt{n}\widehat{\ell}_n(\widehat{\theta}_n, s_n) \\ &\quad + \left[\sqrt{n}(\widehat{\ell}_n(\widehat{\theta}_n^{(k-1)}, s_n) - \widehat{\ell}_n(\widehat{\theta}_n, s_n))\right]. \end{aligned} \tag{A.9}$$

The second term in (A.9) equals to

$$O_P\left(\sqrt{n}|s_n| \vee \frac{g_r(|s_n|) \vee n^{1/2-2r}}{\sqrt{n}|s_n|}\right)$$

according to (17) and (A.6). The third term in (A.9) can be written as

$$-\sqrt{n}\widetilde{I}_0(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n) + O_P\left(\frac{g_r(|s_n| \vee \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|) \vee n^{1/2-2r}}{\sqrt{n}|s_n|}\right).$$

for $k = 1, 2, \ldots$ by replacing U_n with $(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n)$ in (A.3). Combining the above analysis, the assumption (A.7) and nonsingularity of $\widetilde{I}_0$, we complete the proof of (A.8). □

LEMMA A.3. *Suppose that Condition G holds. If* $\widetilde{\theta}_n$ *is a* n^{ψ}*-consistent estimator for* $0 < \psi \leq 1/2$*, then we have*

$$\begin{aligned} & n^{-1}[\widehat{S}_n^{(1)}(\widetilde{\theta}_n) - \widehat{S}_n^{(1)}(\theta_0)] \\ = \; & -\widetilde{I}_0(\widetilde{\theta}_n - \theta_0) + O_P((n^{-g} \vee \|\widetilde{\theta}_n - \theta_0\|)\|\widetilde{\theta}_n - \theta_0\|), \qquad \text{(A.10)} \\ & n^{-1}[\widehat{S}_n^{(1)}(\widetilde{\theta}_n + U_n) - \widehat{S}_n^{(1)}(\widetilde{\theta}_n)] \\ = \; & -\widetilde{I}_0 U_n + O_P((n^{-g} \vee \|\widetilde{\theta}_n - \theta_0\|)\|U_n\|), \qquad \text{(A.11)} \end{aligned}$$

where U_n *a statistic of the order* $O_P(n^{-s})$ *for some* $s \geq \psi$.

PROOF: We first consider (A.10). Using a Taylor's expansion, we have

$$\begin{aligned}\frac{1}{n}\widehat{S}_n^{(1)}(\widetilde{\theta}_n) &= \frac{1}{n}\widehat{S}_n^{(1)}(\theta_0) + \frac{1}{n}\widehat{S}_n^{(2)}(\theta_0)(\widetilde{\theta}_n - \theta_0) + \frac{1}{2}(\widetilde{\theta}_n - \theta_0) \otimes \frac{\widehat{S}_n^{(3)}(\theta_1^*)}{n} \otimes (\widetilde{\theta}_n - \theta_0) \\ &= \frac{1}{n}\widehat{S}_n^{(1)}(\theta_0) + A + B,\end{aligned}$$

where θ_1^* lies between $\widetilde{\theta}_n$ and θ_0. In view of (8) and (26), we have $A = -\widetilde{I}_0(\widetilde{\theta}_n - \theta_0) + O_P(n^{-g}\|\widetilde{\theta}_n - \theta_0\|)$. Condition (27) implies that $B = O_P(\|\widetilde{\theta}_n - \theta_0\|^2)$. This completes the proof of (A.10). We next consider (A.11). Similarly, we have $[\widehat{S}_n^{(1)}(\widetilde{\theta}_n + U_n) - \widehat{S}_n^{(1)}(\widetilde{\theta}_n)]/n$

$$\begin{aligned}
&= \frac{1}{n}\widehat{S}_n^{(2)}(\widetilde{\theta}_n)U_n + O_P(\|U_n\|^2) \\
&= \frac{1}{n}S_n^{(2)}(\widetilde{\theta}_n)U_n + O_P(n^{-g}\|U_n\| \vee \|U_n\|^2), \\
&= \frac{1}{n}S_n^{(2)}(\theta_0)U_n + O_P(\|\widetilde{\theta}_n - \theta_0\|\|U_n\| \vee n^{-g}\|U_n\| \vee \|U_n\|^2), \\
&= -\widetilde{I}_0 U_n + O_P(n^{-1/2}\|U_n\| \vee \|\widetilde{\theta}_n - \theta_0\|\|U_n\| \vee n^{-g}\|U_n\| \vee \|U_n\|^2),
\end{aligned}$$

where the second equation follows from (26), the third equality follows from (27) and the last equation follows from CLT and (8). Considering that $1/4 < g \leq 1/2$ and $s \geq \psi$, we have proved (A.11). □

LEMMA A.4. *Let $\widehat{S}_n(\theta) = \sum_{i=1}^n \ell_\theta(X_i)$. Suppose that $\widetilde{\theta}_n$ is a n^ψ-consistent estimator for $0 < \psi \leq 1/2$. If $\ell_\theta(\cdot)$ satisfies P1 & P2, we have*

$$n^{-1}[\widehat{S}_n^{(1)}(\widetilde{\theta}_n) - \widehat{S}_n^{(1)}(\theta_0)] = -I_0(\widetilde{\theta}_n - \theta_0) + O_P(\|\widetilde{\theta}_n - \theta_0\|^2), \quad \text{(A.12)}$$

$$n^{-1}[\widehat{S}_n^{(1)}(\widetilde{\theta}_n + U_n) - \widehat{S}_n^{(1)}(\widetilde{\theta}_n)] = -I_0 U_n + O_P(\|\widetilde{\theta}_n - \theta_0\|\|U_n\|), \quad \text{(A.13)}$$

where U_n a statistic of the order $O_P(n^{-s})$ for any $s \geq \psi$.

PROOF: We only provide the proof of (A.13) since that of (A.12) is completely analogous and simpler. To show (A.13), it suffices to prove that, for every $C_1, C_2 > 0$ and $s \geq \psi$,

$$\begin{aligned}
&\sup_{|t|\leq C_1, |u|\leq C_2} \left| n^{-1}\left[\widehat{S}_n^{(1)}(\theta_0 + n^{-\psi}t + n^{-s}u) - \widehat{S}_n^{(1)}(\theta_0 + n^{-\psi}t)\right] + n^{-s}I_0 u \right| \\
= \; & O_P(n^{-s-\psi}).
\end{aligned}$$

Denote $Z_n(t,u) = n^{-1/2}[\widehat{S}_n^{(1)}(\theta_0 + n^{-\psi}t + n^{-s}u) - \widehat{S}_n^{(1)}(\theta_0 + n^{-\psi}t)]$ and $Z_n^0(t,u) = Z_n(t,u) - EZ_n(t,u)$. Then, it suffices to show that

$$\sup_{|t|\leq C_1, |u|\leq C_2} |Z_n^0(t,u)| = O_P(n^{1/2-\psi-s}), \quad \text{(A.14)}$$

$$\sup_{|t|\leq C_1, |u|\leq C_2} |EZ_n(t,u) + n^{1/2-s}I_0 u| = O_P(n^{1/2-\psi-s}). \quad \text{(A.15)}$$

The proofs of (A.14) and (A.15) are similar as those of (2.3) and (2.4) in Page 1224 of [25], and are thus skipped. □

LEMMA A.5. *Suppose Conditions K1-K2 & C1-C2 hold. Then we have*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}|_{\theta=\theta_0}A_{\theta,\eta_*(\theta)}\right)[\widehat{\eta}_0-\eta_{*0}](X_i) = O_P(n^{-\delta}), \tag{A.16}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}A_{\theta_0,\eta_0}[\widehat{\eta}_0^{(1)}-\eta_{*0}^{(1)}](X_i) = O_P(n^{-\delta}), \tag{A.17}$$

$$\frac{1}{\sqrt{n}}\dot{r}_n(\theta_0) = O_P(n^{1/2-2g}), \tag{A.18}$$

where $r_n(\theta)\equiv\widehat{S}_n(\theta)-S_n(\theta)-\sum_{i=1}^{n}A_{\theta,\eta_*(\theta)}[\widehat{\eta}(\theta)-\eta_*(\theta)]$.

PROOF: The proof of Lemma 2 in [39] directly implies (A.16) and (A.17). As for (A.18), by Taylor expansion, we first rewrite

$$\begin{aligned} r_n(\theta) &= \frac{1}{2}\sum_{i=1}^{n}\int_0^1\frac{\partial^2\log lik}{\partial\lambda^2}(X_i;\theta,\eta_t(\theta)(Z_i))dt\{\widehat{\eta}(\theta)(Z_i)-\eta_*(\theta)(Z_i)\}^2 \\ &\equiv \frac{1}{2}\sum_{i=1}^{n}Q_\theta(X_i)\{\widehat{\eta}(\theta)(Z_i)-\eta_*(\theta)(Z_i)\}^2, \end{aligned}$$

where $\eta_t(\theta)(Z_i)=\eta_*(\theta)(Z_i)+t(\widehat{\eta}(\theta)-\eta_*(\theta))(Z_i)$. To prove (A.18), it suffices to show that

$$\sup_{z\in\mathcal{Z}}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^j}{\partial\theta^j}|_{\theta=\theta_0}Q_\theta(X_i)\right| = O_P(1) \text{ for } j=0,1 \tag{A.19}$$

in view of (38). For $j=0$, we have

$$|Q_{\theta_0}(x)|\le\sup_{\lambda\in H}\left|\frac{\partial^2\log lik}{\partial\lambda^2}(x;\theta_0,\lambda)\right| = O_P(1) \text{ for all } z\in\mathcal{Z}$$

based on the smoothness Condition K2. The case $j=1$ can be established similarly. □

LEMMA A.6. *Let* $\eta_0(\mathbf{z})=(\eta_0(z_1),\ldots,\eta_0(z_n))'$ *and* $\epsilon=(\epsilon_1,\ldots,\epsilon_n)'$. *If* $\lambda_n\to 0$, *then we have*

$$\mathbf{w}'A(\lambda_n)\epsilon = O_P(\lambda_n^{-1/(2k)}), \tag{A.20}$$

$$\mathbf{w}'[I-A(\lambda_n)]\eta_0(\mathbf{z}) = O_P(n^{1/2}\lambda_n), \tag{A.21}$$

$$\mathbf{w}'(I-A(\lambda_n))\mathbf{w}/n = \Sigma+O_P(n^{-1/2}\vee n^{-1}\lambda_n^{-1/k}). \tag{A.22}$$

PROF: We first state the Lemmas 4.1 and 4.3 in [17]:

$$n^{-1}\sum_{l=1}^{n}[(I-A(\lambda_n))\eta_0(\mathbf{z})]_l^2 \leq \lambda_n^2 J^2(\eta_0), \tag{A.23}$$

$$tr(A(\lambda_n)) = O(\lambda_n^{-1/k}), \tag{A.24}$$

$$tr(A^2(\lambda_n)) = O(\lambda_n^{-1/k}). \tag{A.25}$$

Since $Var[(\mathbf{w}'A(\lambda_n)\epsilon)_i] = \sigma^2\Sigma_{ii}tr(A^2(\lambda_n))$, we can show that $[\mathbf{w}'A(\lambda_n)\epsilon]_i = O_P(\lambda_n^{-1/2k})$ based on (A.25), thus proved (A.20). We next consider (A.21) by establishing that $Var[\mathbf{w}'\{I-A(\lambda_n)\}\eta_0(\mathbf{z})]_i = \Sigma_{ii}\eta_0'(\mathbf{z})[I-A(\lambda_n)]^2\eta_0(\mathbf{z})$. Then, we can prove (A.21) by (A.23). As for (A.22), we first write (A.22) as the sum of

$$\Sigma + (\mathbf{w}'\mathbf{w}/n - \Sigma) - \mathbf{w}'A(\lambda_n)\mathbf{w}/n,$$

where the second term is $O_P(n^{-1/2})$ based on the central limit theorem. For the last term, we have $E\{[\mathbf{w}'A(\lambda_n)\mathbf{w}]_{ij}\}^2 =$

$$\begin{aligned}&(\Sigma_{ij})^2(tr(A(\lambda_n)))^2 + (\Sigma_{ii}\Sigma_{jj} + (\Sigma_{ij})^2)tr(A^2(\lambda_n))\\ &+(E(X_{1i}X_{1j})^2 - 2(\Sigma_{ij})^2 - \Sigma_{ii}\Sigma_{jj})\sum_r A_{rr}^2(\lambda_n)\end{aligned}$$

for $i \neq j$. When $i = j$, we have $E|(\mathbf{w}'A(\lambda_n)\mathbf{w})_{ii}| = \Sigma_{ii}tr(A(\lambda_n))$. By considering (A.24)-(A.25), we have proved (A.22). □

A.3. Proof of Theorem 1

Define $\mathcal{N}_n = \{\theta : \|\theta - \theta_0\| \leq Mn^{-\psi}\}$ and $\mathcal{N}_n^c$ as its complement for any $0 < M < \infty$. Note that $\mathcal{D}_n \cap \mathcal{N}_n \neq \emptyset$ for large enough M and $\mathcal{D}_n \cap \mathcal{N}_n^c \neq \emptyset$ for large enough n. We

first consider (12). For sufficiently large M and any $C_1 > 0$, we have

$$\begin{aligned}
P\left(\theta_n^D \in \mathcal{N}_n^c\right) &= P\left(\theta_n^D \in \mathcal{N}_n^c \text{ and } \theta_{iD} \in \mathcal{N}_n \text{ for some } i\right) \\
&\leq P\left(\max_{\mathcal{D}_n \cap \mathcal{N}_n} \widehat{S}_n(\theta) \leq \max_{\mathcal{D}_n \cap \mathcal{N}_n^c} \widehat{S}_n(\theta)\right) \\
&\leq P\left(\max_{\mathcal{D}_n \cap \mathcal{N}_n} \widehat{S}_n(\theta) < \widehat{S}_n(\theta_0) - C_1 n^{1-2\psi}\right) \\
&\quad + P\left(\left\{\max_{\mathcal{D}_n \cap \mathcal{N}_n} \widehat{S}_n(\theta) \leq \max_{\mathcal{D}_n \cap \mathcal{N}_n^c} \widehat{S}_n(\theta)\right\} \cap \left\{\max_{\mathcal{D}_n \cap \mathcal{N}_n} \widehat{S}_n(\theta) \geq \widehat{S}_n(\theta_0) - C_1 n^{1-2\psi}\right\}\right) \\
&\leq P\left(\max_{\mathcal{D}_n \cap \mathcal{N}_n} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) < -C_1 n^{1/2-2\psi} \cap \{\theta_n^o \text{ is consistent}\}\right) \\
&\quad + P\left(\max_{\mathcal{N}_n^c} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right) \\
&\quad + P\left(\theta_n^o \text{ is inconsistent}\right) \\
&\leq I + II + III,
\end{aligned}$$

where $\theta_n^o = \arg\max_{\mathcal{D}_n \cap \mathcal{N}_n} \widehat{S}_n(\theta)$.

The definition of $\mathcal{N}_n$ implies $III \to 0$ for any M as $n \to \infty$. We next analyze the term I as follows. In view of (11) and the definition of $\mathcal{N}_n$, we have that

$$\begin{aligned}
I &= P\left(\sqrt{n}(\theta_n^o - \theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - \frac{\sqrt{n}}{2}(\theta_n^o - \theta_0)'\widetilde{I}_0(\theta_n^o - \theta_0) + n^{-1/2}\Delta_n(\theta_n^o) < -C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|\|\theta_n^o - \theta_0\| + (\delta_{max}\sqrt{n}/2)\|\theta_n^o - \theta_0\|^2 + \|n^{-1/2}\Delta_n(\theta_n^o)\| > C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\| > \frac{C_1 - \delta_{max}M^2/2}{M} n^{1/2-\psi} + o_P(n^{1/2-\psi})\right) \\
&\leq \bar{I},
\end{aligned}$$

where δ_{max} is the largest eigenvalue of $\widetilde{I}_0$, and the second inequality follows from the definitions of $\mathcal{N}_n$ and Δ_n, and the range that $2v > 1/2 \geq \psi > 0$. Denote $\theta_n^* = \arg\max_{\mathcal{N}_n^c} \widehat{S}_n(\theta)$. We will show $II \to 0$ by first decomposing it as $II_1 + II_2$, where

$$\begin{aligned}
II_1 &= P\left(n^{-1/2}(\widehat{S}_n(\theta_n^*) - \widehat{S}_n(\theta_0)) \geq -C_1 n^{1/2-2\psi} \cap \{\theta_n^* \text{ is consistent}\}\right), \\
II_2 &= P\left((\widehat{S}_n(\theta_n^*) - \widehat{S}_n(\theta_0)) \geq -C_1 n^{1-2\psi} \cap \{\theta_n^* \text{ is inconsistent}\}\right).
\end{aligned}$$

Note that we can write $n^{-1/2}\Delta_n(\theta_n^*)$ as $\sqrt{n}\|\theta_n^* - \theta_0\|^2\epsilon_{1n} + \sqrt{n}\|\theta_n^* - \theta_0\|\epsilon_{2n}$, where $\epsilon_{1n} = o_P(1)$ and $\epsilon_{2n} = o_P(n^{-1/2})$, in the event that $\{\theta_n^* \text{ is consistent}\}$. Thus, according to (11),

we can write II_1 as

$$\begin{aligned}
& P\left((\theta_n^*-\theta_0)'\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0+\sqrt{n}\|\theta_n^*-\theta_0\|\epsilon_{2n}\geq\frac{\sqrt{n}}{2}(\theta_n^*-\theta_0)'\widetilde{I}_0(\theta_n^*-\theta_0)\right.\\
&\qquad\qquad\left.-\sqrt{n}\|\theta_n^*-\theta_0\|^2\epsilon_{1n}-C_1n^{1/2-2\psi}\right)\\
\leq\ & P\left(\|\theta_n^*-\theta_0\|\left[\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|+\sqrt{n}\epsilon_{2n}\right]\geq\frac{\sqrt{n}}{2}\|\theta_n^*-\theta_0\|^2\delta_{min}\right.\\
&\qquad\qquad\left.-\sqrt{n}\|\theta_n^*-\theta_0\|^2\epsilon_{1n}-C_1n^{1/2-2\psi}\right)\\
\leq\ & P\left(\left[\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|+\sqrt{n}\epsilon_{2n}\right]\geq\sqrt{n}\|\theta_n^*-\theta_0\|(\delta_{min}/2-\epsilon_{1n})-\frac{C_1n^{1/2-\psi}}{K}\right)\\
\leq\ & P\left(\left[\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|+\sqrt{n}\epsilon_{2n}\right]\geq\frac{\delta_{min}K^2/2-C_1}{K}n^{1/2-\psi}+o_P(n^{1/2-\psi})\right)\\
\leq\ & \bar{I}I_1,
\end{aligned}$$

where $\delta_{min}>0$ is the smalest eigenvalue of $\widetilde{I}_0$. All the above inequalities follow from the fact that $\|\theta_n^*-\theta_0\|\geq Kn^{-\psi}$ for some $K>M$ and $\epsilon_{1n}=o_P(1)$. The term II_2 is shown to converge to zero by the following contradiction arguments. By assuming that the event $\{(\widehat{S}_n(\theta_n^D)-\widehat{S}_n(\theta_0))\geq -C_1n^{1-2\psi}\}$ holds, we have $|\widehat{S}_n(\theta_n^D)-\widehat{S}_n(\widehat{\theta}_n)|=\widehat{S}_n(\widehat{\theta}_n)-\widehat{S}_n(\theta_n^D)\leq\widehat{S}_n(\widehat{\theta}_n)-\widehat{S}_n(\theta_0)+C_1n^{1-2\psi}$. Note that (11) and the consistency of $\widehat{\theta}_n$ imply $\widehat{S}_n(\theta_0)-\widehat{S}_n(\widehat{\theta}_n)=o_P(n)$. Then, we can show that $|\widehat{S}_n(\theta_n^D)-\widehat{S}_n(\widehat{\theta}_n)|/n=o_P(1)$ which implies that θ_n^D is consistent by (10). This implication contradicts with another event in II_2, i.e., $\{\theta_n^D$ is inconsistent$\}$. Therefore we can claim that $II_2\to 0$.

In view of the above discussions, it remains to show that $\bar{I}$ and $\bar{I}I_1$ converge to zero. Note that $\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|$ in $\bar{I}$ is $O_P(1)$, and so is $(\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\|+\sqrt{n}\epsilon_{2n})$ in $\bar{I}I_1$. Therefore, by choosing sufficiently large C_1 and $K>M$, meanwhile keeping the inequality $\delta_{max}M^2<2C_1<\delta_{min}K^2$ valid, we show that $\bar{I}$ and $\bar{I}I_1$ can be arbitrarily close to zero. For example, we can take $K=M+B$ and $C_1=(\delta_{max}M^2+\delta_{min}(M+B)^2)/4$ for some fixed $B>0$ and sufficiently large M. This completes the proof of (12).

Our proof of (13) is similar as that of (12). Denote θ_{iS} as an element in $\mathcal{S}_n$. Similarly,

we have

$$\begin{aligned}
P(\theta_n^S \in \mathcal{N}_n^c) &\leq E\left\{P\left(\theta_n^S \in \mathcal{N}_n^c \text{ and } \theta_{iS} \in \mathcal{N}_n \text{ for some } i|\mathcal{S}_n\right)\right\} \\
&\quad + E\left\{P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i|\mathcal{S}_n\right)\right\} \\
&\leq P\left(\max_{\mathcal{S}_n\cap\mathcal{N}_n} \widehat{S}_n(\theta) \leq \max_{\mathcal{S}_n\cap\mathcal{N}_n^c} \widehat{S}_n(\theta)\right) + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
&\leq P\left(\max_{\mathcal{S}_n\cap\mathcal{N}_n} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) < -C_2 n^{1/2-2\psi}\right) \\
&\quad + P\left(\max_{\mathcal{S}_n\cap\mathcal{N}_n^c} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) \geq -C_2 n^{1/2-2\psi}\right) \\
&\quad + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
&\leq P\left(\max_{\mathcal{S}_n\cap\mathcal{N}_n} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) < -C_2 n^{1/2-2\psi} \cap \{\theta_n^\dagger \text{ is consistent}\}\right) \\
&\quad + P\left(\max_{\mathcal{N}_n^c} n^{-1/2}(\widehat{S}_n(\theta) - \widehat{S}_n(\theta_0)) \geq -C_2 n^{1/2-2\psi}\right) \\
&\quad + P(\theta_n^\dagger \text{ is inconsistent}) + P\left(\theta_{iS} \in \mathcal{N}_n^c \text{ for all } i\right) \\
&\leq I' + II' + III' + IV',
\end{aligned}$$

where C_2 is an arbitrary positive constant and $\theta_n^\dagger = \arg\max_{\mathcal{S}_n\cap\mathcal{N}_n} \widehat{S}_n(\theta)$.

We first consider the terms III' & IV'. Since $\theta_n^\dagger \in \mathcal{N}_n$, we have $III' \to 0$ for any M as $n \to \infty$. The term IV' is computed as

$$\left(1 - P(\bar{\theta} \in \mathcal{N}_n)\right)^{card(\mathcal{S}_n)}. \tag{A.26}$$

Since the density of $\bar{\theta}$ is assumed to be bounded away from zero around θ_0 and $card(\mathcal{S}_n) \geq \widetilde{C} n^\psi$, (A.26) is bounded above by

$$\left(1 - \rho n^{-\psi} M\right)^{card(\mathcal{S}_n)} \leq \left(1 - \rho M \widetilde{C}/card(\mathcal{S}_n)\right)^{card(\mathcal{S}_n)} \longrightarrow \exp(-\rho M \widetilde{C}), \tag{A.27}$$

for some $\rho > 0$.

We next consider I'. According to (11), we can show

$$\begin{aligned}
& n^{-1/2}(\widehat{S}_n(\theta_n^\dagger) - \widehat{S}_n(\theta_0)) \\
\geq\ & \max_{\mathcal{S}_n\cap\mathcal{N}_n} \left\{-\frac{\sqrt{n}}{2}(\theta - \theta_0)'\widetilde{I}_0(\theta - \theta_0)\right\} - \max_{\mathcal{S}_n\cap\mathcal{N}_n} \{-\sqrt{n}(\theta - \theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - \Delta_n(\theta)/\sqrt{n}\} \\
\geq\ & -\min_{\mathcal{S}_n\cap\mathcal{N}_n} \left\{\frac{\sqrt{n}}{2}(\theta - \theta_0)'\widetilde{I}_0(\theta - \theta_0)\right\} - \max_{\mathcal{S}_n\cap\mathcal{N}_n} \{-\sqrt{n}(\theta - \theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - \Delta_n(\theta)/\sqrt{n}\}.
\end{aligned}$$

Therefore, we can bound I' by $I'_1 + I'_2$, where

$$\begin{aligned} I'_1 &= P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} \{-\sqrt{n}(\theta-\theta_0)'\mathbb{P}_n\widetilde{\ell}_0 - n^{-1/2}\Delta_n(\theta)\} > (C_2/2)n^{1/2-2\psi}\right), \\ I'_2 &= P\left(\min_{\mathcal{S}_n \cap \mathcal{N}_n} \{\sqrt{n}(\theta-\theta_0)'\widetilde{I}_0(\theta-\theta_0)\} > C_2 n^{1/2-2\psi}\right). \end{aligned}$$

Given sufficiently large C_2/M, I'_1 can be arbitrarily close to zero since

$$\begin{aligned} I'_1 &\leq P\left(\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\| > \frac{C_2}{2M}n^{1/2-\psi} + O_P(n^{1/2-2\psi} \vee n^{1/2-2v})\right) \\ &\leq P\left(\|\sqrt{n}\mathbb{P}_n\widetilde{\ell}_0\| > \frac{C_2}{2M}n^{1/2-\psi} + o_P(n^{1/2-\psi})\right), \end{aligned} \tag{A.28}$$

where the last inequality follows from the assumption that $2v > 1/2 \geq \psi$. Since $\min_{\mathcal{N}_n^c}\{\sqrt{n}(\theta-\theta_0)'\widetilde{I}_0(\theta-\theta_0)\} > C_2 n^{1/2-2\psi}$ by choosing $\delta_{\min}M^2 > C_2$, I'_2 is bounded above by

$$\begin{aligned} & P\left(\min_{\mathcal{S}_n}\{\sqrt{n}(\theta-\theta_0)'\widetilde{I}_0(\theta-\theta_0)\} > C_2 n^{1/2-2\psi}\right) \\ \leq\ & \left[P(\sqrt{n}(\bar{\theta}-\theta_0)'\widetilde{I}_0(\bar{\theta}-\theta_0) > C_2 n^{1/2-2\psi})\right]^{card(\mathcal{S}_n)} \\ \leq\ & \left[1 - P(\|\bar{\theta}-\theta_0\| \leq (C_2/\delta_{max})^{1/2}n^{-\psi})\right]^{card(\mathcal{S}_n)} \\ \leq\ & \left[1 - P\left(\|\bar{\theta}-\theta_0\| \leq (C_2/\delta_{max})^{1/2}\widetilde{C}/card(\mathcal{S}_n)\right)\right]^{card(\mathcal{S}_n)} \\ \leq\ & (1 - \rho\widetilde{C}(C_2/\delta_{max})^{1/2}/card(\mathcal{S}_n))^{card(\mathcal{S}_n)} \\ \longrightarrow\ & \exp(-\rho\widetilde{C}\sqrt{C_2/\delta_{max}}) \end{aligned} \tag{A.29}$$

for some $\rho > 0$. In the above, the third and fourth inequality follows from the assumptions that $card(\mathcal{S}_n) \geq \widetilde{C}n^{\psi}$ and the density for $\bar{\theta}$ is bounded away from zero around θ_0, respectively. By assuming that $2C_2 < K^2\delta_{min}$ for some $K > M$, we can prove that $II' \to 0$ in the same manner as we show $II \to 0$.

Let $L = \min\{K^2/2, M^2\}$. In view of (A.27), (A.28), (A.29) and the above discussions on II', by choosing sufficiently large C_2, $K > M$ and C_2/M, meanwhile keeping the inequality $C_2 < L\delta_{min}$ valid, we can make $P(\theta_n^S \in \mathcal{N}_n^c)$ arbitrarily small. For example, we can take $C_2 = M^{3/2}\delta_{min}$ and $K = M + B$, for some fixed $B > 0$ and sufficiently large M. This completes the whole proof. □

A.4. Proof of Lemma 1

By (A.4) in Lemma A.1 and (A.8) in Lemma A.2, we obtain that $(\widehat{\theta}_n^{(k)} - \widehat{\theta}_n)$

$$\begin{aligned}
&= O_P\left(\frac{g_r\left(|t_n^{(k-1)}| \vee \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|\right) \vee n t_n^{(k-1)} \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\| \vee n^{1/2-2r}}{n\{t_n^{(k-1)}\}^2} \times \right. \\
&\qquad \left. \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\| \vee |s_n^{(k-1)}| \vee \frac{g_r\left(|s_n^{(k-1)}| \vee \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|\right) \vee n^{1/2-2r}}{n|s_n^{(k-1)}|}\right) \\
&= O_P\left(\left(|t_n^{(k-1)}| \vee \frac{n^{-2r} \vee n^{-r_{k-1}}}{|t_n^{(k-1)}|} \vee \frac{n^{-3r_{k-1}} \vee n^{-2r-r_{k-1}} \vee n^{-1/2-2r}}{\{t_n^{(k-1)}\}^2}\right)\right. \\
&\qquad \left. \times n^{-r_{k-1}} \vee \frac{n^{-3r_{k-1}} \vee n^{-2r-r_{k-1}} \vee n^{-1/2-2r}}{|s_n^{(k-1)}|} \vee |s_n^{(k-1)}| \vee n^{-2r}\right) \\
&= O_P\left(f_{k-1}(|t_n^{(k-1)}|) \vee h_{k-1}(|s_n^{(k-1)}|) \vee n^{-2r}\right).
\end{aligned}$$

To analyze the above order, we have to consider three different stages: (i) $r_{k-1} < r$; (ii) $r \le r_{k-1} < 1/2$; (iii) $r_{k-1} \ge 1/2$. For the stage (i), the smallest order of f_{k-1}, i.e., $n^{-3r_{k-1}/2}$, is achieved by taking $|t_n^{(k-1)}| \asymp n^{-r_{k-1}/2}$, and the smallest order of h_{k-1}, i.e., $n^{-3r_{k-1}/2}$, is achieved by taking $|s_n| \asymp n^{-3r_{k-1}/2}$. For the stage (ii), the smallest order of f_{k-1}, i.e., $n^{-3r_{k-1}/2}$, is achieved by taking $|t_n^{(k-1)}| \asymp n^{-r_{k-1}/2}$, and the smallest order of h_{k-1}, i.e., $n^{-(2r+r_{k-1})/2}$, is achieved by taking $|s_n^{(k-1)}| \asymp n^{-(2r+r_{k-1})/2}$. For the last stage (iii), the smallest order of f_{k-1}, i.e., $n^{-3r_{k-1}/2}$, is achieved by taking $|t_n^{(k-1)}| \asymp n^{-r_{k-1}/2}$, and the smallest order of h_{k-1}, i.e., $n^{-r-1/4}$, is achieved by taking $|s_n^{(k-1)}| \asymp n^{-r-1/4}$. This completes the whole proof. □

A.5. Proof of Theorem 2

According to the proof in Lemma 1, we also need to consider the stochastic order of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$ in terms of three stages: (i) $r_{k-1} < r$; (ii) $r \le r_{k-1} < 1/2$; (iii) $r_{k-1} \ge 1/2$. In stage (i), we have $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^{3/2}) = O_P(n^{-S_1(\psi,k)})$ if $k \le K_1(\psi, r)$. In stage (ii), we have $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^{1/2} n^{-r})$, which implies that $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-S_2(\psi,r,k)})$ if $r \le \psi < 1/2$. It is easy to show that $S_2(\psi, r, k) \ge 1/2$ if $k \ge K_2(\psi, r, 1/2)$. In the last stage (iii), we obtain the the smallest order of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$, i.e., $O_P(n^{-r-1/4})$. Combining the above analysis of (i)-(iii), we can conclude that the stochastic order of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$ is continuously improving till the optimal bound

$O_P(n^{-r-1/4})$ and can be expressed as $O_P(n^{-S(\psi,r,k)})$. (21) also follows from the above analysis. □

A.6. Proof of Theorem 3

We first show (30) by applying Lemma A.3. In (A.10), we replace $\widetilde{\theta}_n$ by $\widehat{\theta}_n$. Since $\widehat{\theta}_n$ is assumed to be consistent and θ_0 is an interior point of Θ, we have $\widehat{S}_n^{(1)}(\widehat{\theta}_n) = 0$. By (6) and (25), we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \sqrt{n}\widetilde{I}_0^{-1}\mathbb{P}_n\widetilde{\ell}_0 + O_P(n^{1/2-2g} \vee n^{1/2}\|\widehat{\theta}_n - \theta_0\|^2) \tag{A.30}$$

given that $\widehat{\theta}_n$ is consistent and $\widetilde{I}_0$ is nonsingular. Considering the range of g, we can show $\widehat{\theta}_n$ is actually $\sqrt{n}$-consistent, and thus simplify (A.30) to (30).

We next show (31). By (22), we can write $\sqrt{n}\widehat{I}_n(\widehat{\theta}_n^{(0)})(\widehat{\theta}_n^{(1)} - \widehat{\theta}_n)$ as

$$\begin{aligned} & \sqrt{n}\widehat{I}_n(\widehat{\theta}_n^{(0)})(\widehat{\theta}_n^{(0)} - \widehat{\theta}_n) + n^{1/2}(\widehat{\ell}_n(\widehat{\theta}_n^{(0)}) - \widehat{\ell}_n(\widehat{\theta}_n)) \\ = \; & \sqrt{n}\widehat{I}_n(\widehat{\theta}_n^{(0)})(\widehat{\theta}_n^{(0)} - \widehat{\theta}_n) + n^{-1/2}\widehat{S}_n^{(2)}(\widehat{\theta}_n^{(0)})(\widehat{\theta}_n^{(0)} - \widehat{\theta}_n) + O_P(\sqrt{n}\|\widehat{\theta}_n - \widehat{\theta}_n^{(0)}\|^2) \\ = \; & O_P(\sqrt{n}\|\widehat{\theta}_n - \widehat{\theta}_n^{(0)}\|^2) \end{aligned}$$

under Condition G. Further, by (26) and (27), we have the invertibility of $\widehat{I}_n(\widehat{\theta}_n^{(0)})$ based on that of $\widetilde{I}_0$. This implies $\widehat{\theta}_n^{(1)} - \widehat{\theta}_n = O_P(\|\widehat{\theta}_n^{(0)} - \widehat{\theta}_n\|^2)$. By the induction principal, we can thus show

$$\widehat{\theta}_n^{(k)} - \widehat{\theta}_n = O_P(\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^2) \;\text{ for any } k \geq 1. \tag{A.31}$$

(31) follows from (A.31) trivially.

To show (32), we first prove $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| =$

$$O_P\left(n^{1/2-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^2 \vee n^{-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|\right). \tag{A.32}$$

By replacing $\widetilde{\theta}_n$ and U_n with $\widehat{\theta}_n$ and $(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n)$ in (A.11), respectively, we establish that $n^{-1/2}[\widehat{S}_n^{(1)}(\widehat{\theta}_n^{(k-1)}) - \widehat{S}_n^{(1)}(\widehat{\theta}_n)] =$

$$-\sqrt{n}\widetilde{I}_0(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n) + O_P(n^{1/2-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|). \tag{A.33}$$

Similarly, by setting $\widetilde{\theta}_n$ as $\widehat{\theta}_n$, and then setting U_n as $(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n + n^{-1/2}t_1v_j)$ and $(\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n + n^{-1/2}t_2v_j)$ in (A.11), respectively, we have that

$$[\widehat{I}_n(\widehat{\theta}_n^{(k-1)})]_{ij} = [\widetilde{I}_0]_{ij} + O_P(n^{1/2-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\| \vee n^{-g}) \tag{A.34}$$

when $\widehat{I}_n^{(k-1)}$ is defined in (24). Following similar logic in analyzing (A.31), we can obtain (A.32) by considering (A.33)-(A.34). Next we will show that (A.32) implies (32) by the following analysis. Based on (A.32) we have

$$\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = \begin{cases} O_P(n^{-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|) & \text{if } \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\| = O_P(n^{-1/2}), \\ O_P(n^{1/2-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^2) & \text{if } \|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^{-1} = O_P(n^{1/2}). \end{cases} \tag{A.35}$$

It is easy to show that $\|\widehat{\theta}_n^{(L_1(\psi,g))} - \widehat{\theta}_n\| = O_P(n^{-1/2})$ and $\|\widehat{\theta}_n^{(L_1(\psi,g)-1)} - \widehat{\theta}_n\|^{-1} = O_P(n^{1/2})$. In other words, if $k \leq L_1(\psi, g)$, then we have the relation that $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{1/2-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|^2)$ based on (A.35). This implies the form of $R_1(\psi, g, k)$ in (29). Note that $R_1(\psi, g, k)$ is an increasing function of k under the condition that $\psi + g > 1/2$. After $L_1(\psi, g)$ iterations, we have

$$\|\widehat{\theta}_n^{(L_1(\psi,g))} - \widehat{\theta}_n\| = O_P(n^{-R_1(\psi,g,L_1(\psi,g))}) = O_P(n^{-1/2}). \tag{A.36}$$

Thus, we have the relation that $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\| = O_P(n^{-g}\|\widehat{\theta}_n^{(k-1)} - \widehat{\theta}_n\|)$ for $k \geq (L_1(\psi, g)+1)$ based on (A.35). Combining this relation with (A.36), we can show the form of $R_2(\psi, g, k)$ when $k > L_1(\psi, g)$. Since $R(\psi, g, k)$ is an increasing function of k given that $1/2 - g < \psi \leq 1/2$, the stochastic order of $\|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\|$ is continuously decreasing as $k \to \infty$. The calculation of k^* also follows from the above analysis. □

A.7. Proof of Theorem 4

We first consider (25) by rewriting its LHS as

$$\frac{1}{n}\frac{\partial}{\partial\theta}|_{\theta=\theta_0}\left[\sum_{i=1}^{n} A_{\theta,\eta_*(\theta)}[\widehat{\eta}(\theta) - \eta_*(\theta)](X_i) + r_n(\theta)\right],$$

where $r_n(\theta)$ is defined in Lemma A.5. Therefore, we have

$$\begin{aligned} & n^{-1}[\widehat{S}_n^{(1)}(\theta_0) - S_n^{(1)}(\theta_0)] \\ = \; & \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}|_{\theta=\theta_0} A_{\theta,\eta_*(\theta)}\right)(\widehat{\eta}_0 - \eta_{*0}) + \frac{1}{n}\sum_{i=1}^{n} A_{\theta_0,\eta_0}(\widehat{\eta}_0^{(1)} - \eta_{*0}^{(1)}) + \frac{1}{n}\dot{r}_n(\theta_0) \\ = \; & O_P(n^{-2g}) \end{aligned}$$

by Lemma A.5 and the condition that $\delta \geq (2g - 1/2)$. As discussed previously, we will show (26) together with (28). By Taylor expansion, we have

$$\begin{aligned}\widehat{S}_n(\theta) - S_n(\theta) &= \sum_{i=1}^{n} \int_0^1 \frac{\partial \log lik}{\partial \lambda}(X_i; \theta, \eta_t(\theta)(Z_i)) dt [\widehat{\eta}(\theta)(Z_i) - \eta_*(\theta)(Z_i)] \\ &\equiv \sum_{i=1}^{n} R_\theta(X_i)[\widehat{\eta}(\theta)(Z_i) - \eta_*(\theta)(Z_i)],\end{aligned}$$

where $\eta_t(\theta) = \eta_*(\theta) + t(\widehat{\eta}(\theta) - \eta_*(\theta))$. Hence, to prove (26) and (28), it suffices to show that

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \sup_{z \in \mathcal{Z}} \left| n^{-1} \sum_{i=1}^{n} \frac{\partial^j}{\partial \theta^j} R_\theta(X_i) \right| = O_P(1) \quad \text{for } j = 0, 1, 2, 3 \tag{A.37}$$

in view of (38) and (39). Considering the smoothness Condition K2, we can prove (A.37) using the same approach as in the proof of (A.19).

In the end, it remains to show that the class of functions

$$\{(\partial^3/\partial\theta^3) \log lik(x; \theta, \eta_*(\theta)) : \theta \in \mathcal{N}(\theta_0)\}$$

is P-Glivenko-Cantelli and that

$$\sup_{\theta \in \mathcal{N}(\theta_0)} E \left| (\partial^3/\partial\theta^3) \log lik(X; \theta, \eta_*(\theta)) \right| < \infty. \tag{A.38}$$

Let $\ell^{(3)}(\theta, \eta(\theta)) = (\partial^3/\partial\theta^3) \log lik(x; \theta, \eta_*(\theta))$. For any $\theta_1, \theta_2 \in \mathcal{N}(\theta_0)$, we have $|\ell^{(3)}(\theta_1, \eta_*(\theta_1)) - \ell^{(3)}(\theta_2, \eta_*(\theta_2))|$

$$\begin{aligned}&\leq \sup_{\theta,\lambda} \left| \frac{\partial \ell^{(3)}}{\partial \theta}(\theta, \lambda) \right| \|\theta_1 - \theta_2\| + \sup_{\theta,\lambda} \left| \frac{\partial \ell^{(3)}}{\partial \lambda}(\theta, \lambda) \right| \|\eta_*(\theta_1) - \eta_*(\theta_2)\|_\infty \\ &\leq \sup_{\theta,\lambda} \left| \frac{\partial \ell^{(3)}}{\partial \theta}(\theta, \lambda) \right| \|\theta_1 - \theta_2\| + \sup_{\theta,\lambda} \left| \frac{\partial \ell^{(3)}}{\partial \lambda}(\theta, \lambda) \right| \sup_{\theta \in \mathcal{N}(\theta_0)} \|\eta_*^{(1)}(\theta)\|_\infty \times \|\theta_1 - \theta_2\| \\ &\leq A\|\theta_1 - \theta_2\|.\end{aligned}$$

By Condition K2 and $\sup_{\theta \in \mathcal{N}(\theta_0)} \|\eta_*^{(1)}(\theta)\|_\infty < \infty$ in Condition C1, we know that $EA^2 < \infty$. Thus, by the P-G-C preservation Theorem 9.23 of [26] and compactness of $\mathcal{N}(\theta_0)$, we know that

$$\{(\partial^3/\partial\theta^3) \log lik(x; \theta, \eta_*(\theta)) : \theta \in \mathcal{N}(\theta_0)\}$$

is P-Glivenko-Cantelli. The last condition (A.38) follows from the Conditions K2 and C1 by some algebra. □

A.8. Proof of Lemma 2

Let

$$\widehat{m}_\theta(z) = \frac{\sum_{i=1}^n \psi_\theta(Y_i, W_i) K((z - Z_i)/b_n)}{\sum_{i=1}^n K((z - Z_i)/b_n)}.$$

Note that $\widehat{\eta}(\theta)(z) = \rho(\widehat{m}_\theta(z))$ by (42). Correspondingly, we have $\eta_*(\theta)(z) = \rho(m_\theta(z))$ based on Lemma 7 of [39]. Following the proof of Lemma 8 in [39], we can derive that

$$\begin{aligned} & \sup_{\theta\in\Theta} \left\| \frac{\partial^{k+j}}{\partial z^k \partial\theta^j} \widehat{m}_\theta(z) - \frac{\partial^{k+j}}{\partial z^k \partial\theta^j} m_\theta(z) \right\|_\infty \\ = \; & O_P\left(n^{-\frac{q}{2q+4}} b_n^{-k-\frac{q+4}{q+2}} n^\epsilon \vee b_n^2 \right) \end{aligned} \tag{A.39}$$

for any $\epsilon > 0$, $k = 0, 1$ and $j = 0, 1, 2, 3$. Considering (42), (A.39) and Condition (f), we can show that

$$\sup_{\theta\in\mathcal{N}(\theta_0)} \|\widehat{\eta}^{(s)}(\theta) - \eta_*^{(s)}(\theta)\|_\infty = O_P\left(n^{-\frac{q}{2q+4}} b_n^{-\frac{q+4}{q+2}} n^\epsilon \vee b_n^2 \right) \tag{A.40}$$

for $s = 0, 1, 2, 3$ after some algebra. Following similarly logic, we show that

$$\left\| \frac{\partial}{\partial z}\widehat{\eta}_0(z) - \frac{\partial}{\partial z}\eta_{*0}(z) \right\|_\infty = O_P\left(n^{-\frac{q}{2q+4}} b_n^{-\frac{2q+6}{q+2}} n^\epsilon \vee b_n^2 \right) \tag{A.41}$$

$$\left\| \frac{\partial}{\partial z}\widehat{\eta}_0^{(1)}(z) - \frac{\partial}{\partial z}\eta_{*0}^{(1)}(z) \right\|_\infty = O_P\left(n^{-\frac{q}{2q+4}} b_n^{-\frac{2q+6}{q+2}} n^\epsilon \vee b_n^2 \right) \tag{A.42}$$

Considering (A.40)-(A.42), we complete the whole proof. □

A.9. Proof of Corollary 2

For the $\sqrt{n}$ consistency of $\widehat{\theta}_{\lambda_n}$, it suffices to show that, for any given $\epsilon > 0$, there exists a large constant M such that

$$P\left\{ \inf_{\|s\|=M} \Delta_n(s) > 0 \right\} \geq 1 - \epsilon, \tag{A.43}$$

where $\Delta_n(s) \equiv [\widehat{S}_{\lambda_n}(\theta_0 + n^{-1/2}s) - \widehat{S}_{\lambda_n}(\theta_0)]$. According to (55), we have

$$\Delta_n(s) \geq \widetilde{S}_{\lambda_n}(\theta_0 + n^{-1/2}s) - \widetilde{S}_{\lambda_n}(\theta_0) + n\tau_n^2 \sum_{j=1}^q \frac{|\theta_{0j} + n^{-1/2}s_j| - |\theta_{0j}|}{|\widetilde{\theta}_j|},$$

where s_j is the j-th element of s. The Taylor expansion further gives

$$\Delta_n(s) \geq n^{-1/2} s' \widetilde{S}^{(1)}_{\lambda_n}(\theta_0) + \frac{1}{2} s'[\widetilde{S}^{(2)}_{\lambda_n}(\theta_0)/n]s + n\tau_n^2 \sum_{j=1}^{q} \frac{|\theta_{0j} + n^{-1/2}s_j| - |\theta_{0j}|}{|\widetilde{\theta}_j|}, \quad \text{(A.44)}$$

where $\widetilde{S}^{(j)}_{\lambda_n}(\theta_0)$ represents the j-th derivative of $\widetilde{S}_{\lambda_n}(\theta)$ at θ_0. Based on (56), we have

$$\widetilde{S}^{(1)}_{\lambda_n}(\theta_0) = -2\mathbf{w}'[I - A(\lambda_n)](\mathbf{y} - \mathbf{w}\theta_0), \quad \text{(A.45)}$$
$$\widetilde{S}^{(2)}_{\lambda_n}(\theta_0) = 2\mathbf{w}'[I - A(\lambda_n)]\mathbf{w}. \quad \text{(A.46)}$$

Lemma A.6 implies that

$$\widetilde{S}^{(1)}_{\lambda_n}(\theta_0) = O_P(n^{1/2}), \quad \text{(A.47)}$$
$$\widetilde{S}^{(2)}_{\lambda_n}(\theta_0) = O_P(n) \quad \text{(A.48)}$$

since λ_n is required to converge to zero. Hence, we know the first two terms in the right hand side of (A.44) have the same order, i.e. $O_P(1)$. And the second term, which converges to some positive constant, dominates the first one by choosing sufficiently large M. The third term is bounded by $n^{1/2}\tau_n^2 M_0$ for some positive constant M_0 since $\widetilde{\beta}_j$ is the consistent estimate for the nonzero coefficient. Considering that $\sqrt{n}\tau_n^2 \to 0$, we have shown the $\sqrt{n}$-consistency of $\widehat{\theta}_{\lambda_n}$.

To complete the proof of other parts, we first need to show

$$\|\widehat{\theta}^{(1)}_{\lambda_n} - \widehat{\theta}_{\lambda_n}\| = O_P(n^{-1}) \quad \text{(A.49)}$$

based on Theorem 5. And then we will verify Condition G' for the case $c = -2$. It is easy to show that $\mathbb{P}_n\widetilde{\ell}_0 = \mathbf{w}'\epsilon/n$ and $\widetilde{I}_0 = \Sigma$ in this example. To verify (48), we have

$$\begin{aligned} & \frac{1}{n}\widehat{S}^{(1)}_{\lambda_n}(\theta_0) + 2\mathbb{P}_n\widetilde{\ell}_0 \\ = & -\frac{2}{n}\mathbf{w}'(I - A(\lambda_n))\eta_0(\mathbf{z}) + \frac{2}{n}\mathbf{w}'A(\lambda_n)\epsilon + \tau_n^2\delta_n(\theta_0) \\ = & O_P(n^{-1/2}\lambda_n \vee n^{-1}\lambda_n^{-1/(2k)} \vee \tau_n^2), \end{aligned}$$

where the second equality follows from Lemma A.6 and the fact that $\delta_n(\theta_0) = O_P(1)$. Considering the conditions on τ_n and λ_n, we have proved (48). (49) follows from (A.46) and (A.22), and (50) trivially holds. Having shown the consistency of $\widehat{\theta}_{\lambda_n}$ and verified G', we are able to show (A.49).

For any sequence of estimate θ_n, the below arguments show that $\theta_n = 0$ with probability tending to one if it is $\sqrt{n}$-consistent. For any $\sqrt{n}$-consistent estimator, it suffices to show that

$$\widehat{S}_{\lambda_n}\{(\bar{\theta}_1, 0)\} = \min_{\|\bar{\theta}_2\| \leq Cn^{-1/2}} \widehat{S}_{\lambda_n}\{(\bar{\theta}_1, \bar{\theta}_2)\} \tag{A.50}$$

for any $\bar{\theta}_1$ satisfying $\|\bar{\theta}_1 - \theta_1\| = O_P(n^{-1/2})$ with probability approaching to 1. In order to show (A.50), we need to show that $\partial \widehat{S}_{\lambda_n}(\theta)/\partial\theta_j < 0$ for $\theta_j \in (-Cn^{-1/2}, 0)$ and $\partial \widehat{S}_{\lambda_n}(\theta)/\partial\theta_j > 0$ for $\theta_j \in (0, Cn^{-1/2})$ holds when $j = q+1, \ldots, d$ with probability tending to 1. By two term Taylor expansion of $\widetilde{S}_{\lambda_n}(\theta)$ at θ_0, $\partial \widehat{S}_{\lambda_n}(\theta)/\partial\theta_j$ can be expressed in the following form:

$$\frac{\partial \widehat{S}_{\lambda_n}(\theta)}{\partial\theta_j} = \frac{\partial \widetilde{S}_{\lambda_n}(\theta_0)}{\partial\theta_j} + \sum_{k=1}^{d} \frac{\partial^2 \widetilde{S}_{\lambda_n}(\theta_0)}{\partial\theta_j \partial\theta_k}(\theta_k - \theta_{0k}) + n\tau_n^2 \frac{1 \times sign(\theta_j)}{|\widetilde{\theta}_j|},$$

for $j = q+1, \ldots, d$. Note that $\|\bar{\theta} - \theta_0\| = O_P(n^{-1/2})$ by the above construction. Hence, we have

$$\frac{\partial \widehat{S}_{\lambda_n}(\theta)}{\partial\theta_j} = O_P(n^{1/2}) + sign(\theta_j)\frac{n\tau_n^2}{|\widetilde{\theta}_j|}$$

by (A.47) and (A.48). We assume that $n^{k/(2k+1)}\tau_n \to \tau_0 > 0$ which implies that $\sqrt{n}\tau_n^2/|\widetilde{\theta}_j| \to \infty$ for $\sqrt{n}$ consistent $\widetilde{\theta}_j$ and $j = q+1, \ldots, d$. Thus, we show that the sign of θ_j determines that of $\partial \widehat{S}_{\lambda_n}(\theta)/\partial\theta_j$. The above arguments apply to $\widehat{\theta}_{\lambda_n,2}$ and $\widehat{\theta}^{(1)}_{\lambda_n,2}$ since both of them are proven to be $\sqrt{n}$ consistent in view of the previous discussions, i.e., (A.49).

Now it remains to show the semiparametric efficiency of $\widehat{\theta}_{\lambda_n,1}$, which immediately implies that of $\widehat{\theta}^{(1)}_{\lambda_n,1}$ based on (A.49). Since we have shown $\widehat{\theta}_{\lambda_n,2} = 0$, we can establish that

$$\frac{\partial \widehat{S}_{\lambda_n}(\theta)}{\partial\theta_j}|_{\theta=(\widehat{\theta}_{\lambda_n,1},0)} = 0 \quad \text{for any } j = 1, \ldots, q \tag{A.51}$$

with probability tending to one. Let $\mathbf{w}_1$ denote the first q columns of $\mathbf{w}$. Applying Taylor

expansion to (A.51) around θ_0, we obtain

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_{\lambda_n,1}-\theta_1) &= \sqrt{n}\left\{\frac{1}{n}\mathbf{w}_1'[I-A(\lambda_n)]\mathbf{w}_1\right\}^{-1}\frac{1}{n}\mathbf{w}_1'[I-A(\lambda_n)](\eta_0(\mathbf{z})+\epsilon)\\ &\quad +O_P(\sqrt{n}\tau_n^2)\\ &= \left\{\Sigma_{11}+O_P(n^{-1/2}\vee n^{-1}\lambda_n^{-1/k})\right\}^{-1}\frac{1}{\sqrt{n}}\mathbf{w}_1'\epsilon\\ &\quad +O_P(\sqrt{n}\tau_n^2\vee n^{-1/2}\lambda_n^{-1/(2k)}\vee\lambda_n)\\ &= \frac{1}{\sqrt{n}}\Sigma_{11}^{-1}\sum_{i=1}^{n}W_{1i}\epsilon_i+O_P(\sqrt{n}\lambda_n^2\vee\sqrt{n}\tau_n^2)\end{aligned}$$

based on (A.45) & (A.46). This completes the whole proof. □

References

[1] Ai, C. and Chen, X.H. (2003) Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions, *Econometrika* 71 1795-1843

[2] Andrews, D. (1995) Nonparametric Kernel Estimation for Semiparametric Models, *Econometric Theory* 11 560-596

[3] Andrews, D. (2002) Higher-order Improvements of a Computationally Attractive k-step Bootstrap for Extremum Estimators, *Econometrica* 70 1 119-162

[4] Banerjee, M., Mukherjee, D. and Mishra, S. (2009), Semiparametric Binary Regression Models under Shape Constraints with an Application to Indian Schooling Data. *Journal of Econometrics* 149 101-117.

[5] Bickel, P. (1982). *On Adaptive Estimation. Annals of Statistics* 10 647-671.

[6] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

[7] Bunea, F. and McKeague, I. (2005). *Covariate Selection for Semiparametric Hazard Function Regression Models. Journal of Multivariate Analysis* 92 186-204.

[8] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997) Generalized Partially Linear Single Index Models, *Journal of American Statistical Association*, 92, 477-489.

[9] Chen, X. (2007) Large Sample Sieve Estimation of Semi-Nonparametric Models, Chp. 76 in Handbook of Econometrics, Vol. 6B, eds. J.J. Heckman and E.E. Leamer. North-Holland, Amsterdam.

[10] Cheng, G. (2010) Semiparametric Inferences based on K-Step Profile Maximum Likelihood Estimate, *Technical Report*.

[11] Cheng, G. (2009) Semiparametric Additive Isotonic Regression, *Journal of Statistical Planning and Inference*, 100, 345-362.

[12] Cheng, G. and Huang, J.Z. (2009) Bootstrap Consistency for General Semiparametric M-estimation, *Annals of Statistics*, To Appear.

[13] Cheng, G. and Kosorok, M.R. (2008a). Higher order semiparametric frequentist inference with the profile sampler. *Annals of Statistics*, 36, 1786-1818

[14] Cheng, G. and Kosorok, M.R. (2008b). General Frequentist Properties of the Posterior Profile Distribution. *Annals of Statistics*, 36, 1819-1853

[15] Cheng, G. and Kosorok, M.R. (2009) The Penalized Profile Sampler. *Journal of Multivariate Analysis*, **100** 345–362.

[16] Cheng, G. and Zhang, H.H. (2010) Sparse and Efficient Estimation for Partial Spline Models with Increasing Dimension. *Unpublished Manuscript*.

[17] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik* **31**, 377-403.

[18] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression, *Annals of Statistics* 32, 407-451.

[19] Fan, J., Heckman, N.E. and Wand, W.P. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions. *Journal of American Statistical Association* 90 141-150. 196–216.

[20] Fletcher, R. (1980). Practical methods of optimization. vol. 1. New York: John Wiley.

[21] Forrester, W., Hooper, H., Peng, A. and Schick, S. (2003) On the Construction of Efficient Estimators in Semiparametric Models, *Statist. & Decisions* 21 109-138.

[22] Heckman, N. (1986). Spline smoothing in a partly linear models. *Journal of Royal Statistical Society, Series B* **48** 244–248.

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

[24] Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics* **24,** 540–568.

[25] Jassen, P., Jureckova, J. and Veraverbeke, N. (1985). Rate of Convergence of One- and Two-step M-estimators with Applications to Maximum Likelihood and Pitman Esitmators. *Annals of Statistics* **25** 1471–1509.

[26] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

[27] Lee, B. L., Kosorok, M. R. and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association* **100** 960–969.

[28] Linton, O., Sperlich, S. and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *Annals of Statistics* 36 686-718.

[29] Mammen, E. and van de Geer, S. (1997) Penalized Quasi-Likelihood Estimation in Partial Linear Models *Annals of Statistics* 25 1014-1035

[30] Murphy, S. A., Rossini, A.J. and van der Vaart, A. W. (1997) MLE in the proportional odds model. *J. of Amer. Statisti. Assoc.* **92** 968–976.

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

[32] Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **93** 1461–1474.

[33] Ortega, J.M. and Rheinboldt, W.C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press.

[34] Robinson, P.M. (1988). The Stochastic Difference Between Econometric Statistics. *Econometrica* **56** 531–548.

[35] Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996) A Semiparametric Mixture Approach to Case-Control Studies with Errors in Covariables, *Journal of the American Statistical Association* 91 722-732.

[36] Schick, A. (1986) On asymptotically efficient estimation in semiparametric models, *Annals of Statistics* 14 1139-1151.

[37] Schick, A. (1987) A note on the construction of asymptotically linear estimators. *Journal of Statistical Planning and Inferences* 16 89-106. Correction (1989) 262-270.

[38] Schick, A. (1996) Root-n Consistent and Efficient Estimation in Semiparametric Additive Regression Models, *Statist. and Probab. Lett.* 30 45-51.

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

[40] Severini, T.A. and Staniswalis, J.G. (1994) Quasi-likelihood estimatio in semiparmetric models, *Journal of American Statistical Association* 89 501-511.

[41] Silverman, B. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society, Series B* 47 1-52

[42] Staniswalis, J.G. (1989) On the Kernel Estimate of a Regression Function in Likelihood Based Models, *JASA* 84 276-283.

[43] Speckman, P. (1988) Kernel Smoothing in Partial Linear Models *JRSS-B* 50 413-436

[44] Swann, W.H. (1972) Discrete search methods in *Numerical Methods for Unconstrained Optimization*, ed. by Murray, W., New York: Academic Press, 13-28.

[45] G. Wahba (1998) Spline Models for Observational Data. SIAM, Philadelphia.

[46] Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters* 57 135-143.