Econometric Theory, **14**, 1998, 423–462. Printed in the United States of America.

EFFICIENT SEMIPARAMETRIC SCORING ESTIMATION OF SAMPLE SELECTION MODELS

Songnian Chen and Lung-Fei Lee
Hong Kong University of Science and Technology

A semiparametric likelihood method is proposed for the estimation of sample selection models. The method is a two-step semiparametric scoring estimation procedure based on an index restriction and kernel estimation. Under some regularity conditions, the estimator is $\sqrt{n}$-consistent and asymptotically normal. The estimator is also asymptotically efficient in the sense that its asymptotic covariance matrix attains the semiparametric efficiency bound under the index restriction. For the binary choice sample selection model, it also attains the efficiency bound under the independence assumption. This method can be applied to the estimation of general sample selection models.

1. INTRODUCTION

The sample selection model was introduced in Gronau (1974) and Heckman (1974) to study the female labor supply problem. Subsequently, it has been extended to incorporate more general discrete choice settings (e.g., Lee, 1978; Dubin and McFadden, 1984). Under the normal distributional assumption, parametric estimation methods such as the maximum likelihood procedure and simple two-stage methods are available for the estimation of such models. Recently, alternative approaches that relax the normality assumption have been proposed. Semiparametric estimation of the binary choice equation has been considered by Cosslett (1983), Han (1987), Horowitz (1992), Horowitz and Hardle (1996), Ichimura (1993), Klein and Spady (1993), Manski (1985), Powell, Stock, and Stoker (1989), and Sherman (1993). For the binary choice sample selection model, based on the Edgeworth approximation of unknown distributions, Lee (1982) included additional terms in a two-stage estimation method. Cosslett (1991) has shown that the model can be consistently estimated without parametric distributional assumptions. Based on a series approximation to unknown density functions, Gallant and Nychka (1987) have proposed a consistent semi-nonparametric maximum likelihood method. The asymptotic distributions of both the estimators of Cosslett and Gallant and Nychka are unknown. Semiparametric estimators that are $\sqrt{n}$-

This research was supported by the Research Grants Council of Hong Kong under grants DAG94-95BM26 (Chen) and HKUST595/96H (Lee). We thank two anonymous referees and the co-editor Joel Horowitz for helpful comments that greatly improved the paper. Address correspondence to: Songnian Chen, Department of Economics, The Hong King University of Science and Technology, Kowloon, Hong Kong.

 0266-4666/98 $9.50

consistent and asymptotically normal are available in Ahn and Powell (1993), Andrews (1991), Donald (1995), Ichimura and Lee (1991), Newey (1988), Powell (1989), and Robinson (1988), among others. The approaches in Ahn and Powell and Powell and Robinson are single equation two-stage estimation methods by kernel methods, whereas the approaches in Andrews, Donald, and Newey are based on series approximation of conditional expectations. The approach in Ichimura and Lee is a semiparametric nonlinear least-squares method. The asymptotic efficiency issue for semiparametric estimation of a two-equation sample selection model has been investigated in Chamberlain (1986) under the restriction that the error terms are independent of the regressors. Chamberlain has derived an asymptotic efficiency bound for the variances of $\sqrt{n}$-consistent semiparametric estimators for the model.

In this paper, we propose a semiparametric likelihood estimation method based on an index restriction and kernel estimation. The estimation method is motivated by methods of adaptive estimation (Bickel, 1982; Manski, 1984; and Schick, 1986) and semiparametric estimation of index models (Ichimura, 1993; Ichimura and Lee, 1991; and Klein and Spady, 1993). The estimator is asymptotically normal. For the binary choice sample selection model, under some regularity conditions, the estimator is asymptotically efficient in the sense that its asymptotic variance attains the asymptotic efficiency bound of Chamberlain (1986). This approach involves the construction of effective score functions given an $\sqrt{n}$-consistent estimate and is a two-step maximum likelihood estimation procedure. This procedure can be applied to the estimation of general sample selection models with polychotomous choice and sequential choice models with selectivity. For the latter models, the estimator is semiparametric asymptotically efficient under index restrictions. In a recent paper, Ai (1997) proposed a one-step semiparametric likelihood-based estimator that is also semiparametric efficient. His estimator is defined by setting an estimated sample score equal to zero and thus involves solving a nonlinear equation. Given an initial $\sqrt{n}$-consistent estimator, our estimator has a closed form.

This paper is organized as follows. Section 2 introduces the semiparametric scoring estimator. Large sample properties are discussed in Section 3. In Section 4 a small Monte Carlo study is carried out to illustrate the usefulness of the estimator. Section 5 points out generalizations of the procedure to estimation of sequential choice models with selectivity. Section 6 concludes.

2. SAMPLE SELECTION MODELS AND SEMIPARAMETRIC ESTIMATION

Sample selection models consist of some discrete choice equations and outcome equations. The outcome equations are defined on the whole population, but their outcome values can be observed only selectively. Within a utility maximization framework (McFadden, 1974), suppose there are L different alternatives. Let $U_l = x\delta_{l,o} + u_l$ be the associated utility for alternative l, $l = 1,\dots,L$. Let I_l be a dichotomous indicator that alternative l will be chosen: $I_l = 1 \Leftrightarrow x\delta_{l,o} - x\delta_{j,o} \geq u_j - u_l, j \neq l, j = 1, \ldots, L$; and $I_l = 0$, otherwise. Associated with the first choice alternative, there is a vector of continuous outcomes: $y = x\gamma_o + \epsilon$ where y and ϵ are k-dimensional row vectors and γ_o is a matrix of coefficients. The value of y can be observed only when $I_1 = 1$. A more general model may allow continuous outcome equations associated with each of the alternatives. However, for notational simplicity, we will concentrate on the model with outcome equations available only for $I_1 = 1$.

Let $x\alpha_{l,o} = x\delta_{l,o} - x\delta_{L,o}$, $l = 1, \ldots, L-1$, and let $x\alpha_o = (x\alpha_{1,o}, \ldots, x\alpha_{L-1,o})$ be a vector of indices. As an index model, we assume that the joint density of ϵ and u, where $u = (u_1, \ldots, u_L)$ (and hence the marginal density of u) conditional on x can depend on x only through $x\alpha_o$. Let $\tilde{f}(\epsilon, u | x\alpha_o)$ be the conditional joint density of ϵ and u and $h(u|x\alpha_o)$ be the conditional marginal density of u conditional on x under the index restriction. Thus, the choice probability for the alternative l is

$$P(I_l = 1|x) = \int_{-\infty}^{\infty} \left(\prod_{s \neq l}^{L} \int_{-\infty}^{x\delta_{l,o} - x\delta_{s,o} - u_l} \right) h(u|x\alpha_o) \left(\prod_{s \neq l}^{L} du_s \right) du_l. \tag{2.1}$$

The density function of y conditional on $I_1 = 1$ and x is

$$f^*(y|I_1 = 1, x) = \int_{-\infty}^{\infty} \left(\prod_{s=2}^{L} \int_{-\infty}^{x\delta_{1,o} - x\delta_{s,o} + u_1} \right) \tilde{f}(y - x\gamma_o, u|x\alpha_o) \left(\prod_{s=2}^{L} du_s \right) du_1 / P(I_1 = 1|x). \tag{2.2}$$

From (2.1) and (2.2), the choice probabilities for the discrete alternatives are all functions of the indices $x\alpha_o$, and the conditional density function of y is a function of $x\alpha_o$ and $y - x\gamma_o$.

As a generalization, let us assume that the number of indices in the choice probabilities is m where m can be greater than, equal to, or less than the number of choices L. This generalization will include some models with multivariate qualitative dependent variables and models with partial observability. To capture possible constraints, let θ be the vector of parameters in Θ, a convex subset in a finite-dimensional Euclidean space. The α and γ are functions of θ. Let θ_o denote the true parameter vector. For any possible value θ of θ_o in Θ, denote $x\alpha(\theta) = (x\alpha_1(\theta), \ldots, x\alpha_m(\theta))$. Let $p(x_i\alpha|\theta)$ denote the density function of $x\alpha(\theta)$ evaluated at $x_i\alpha(\theta)$ and let $g(y_i - x_i\gamma, x_i\alpha|\theta)$ denote the density of $(y - x\gamma(\theta), x\alpha(\theta))$ evaluated at $(y_i - x_i\gamma(\theta), x_i\alpha(\theta))$. Furthermore, let $p(x_i\alpha|I_1 = 1, \theta)$ denote the density function of $x\alpha(\theta)$ conditional on $I_1 = 1$, evaluated at $x_i\alpha(\theta), g(y_i - x_i\gamma, x_i\alpha|I_1 = 1, \theta)$ denote the density of $(y - x\gamma(\theta), x\alpha(\theta))$ conditional on $I_1 = 1$, evaluated at $(y_i - x_i\gamma(\theta), x_i\alpha(\theta))$, and $f(y_i - x_i\gamma|I_1 = 1, x_i\alpha, \theta)$ be the conditional density $y - x\gamma(\theta)$ conditional on $I_1 = 1$ and $x\alpha(\theta)$ evaluated at $(y_i - x_i\gamma(\theta), x_i\alpha(\theta))$. If there are no confusions, we denote $x\alpha = x\alpha(\theta)$ and $x\gamma = x\gamma(\theta)$, with θ suppressed for simplicity. At θ_o, denote $\alpha_o = \alpha(\theta_o)$ and $\gamma_o = \gamma(\theta_o)$.

The choice probabilities in (2.1) are

$$P(I_l = 1|x) = E(I_l|x\alpha_o). \tag{2.3}$$

From (2.2), it is apparent that the density function of y conditional on $I_1 = 1$ and x is exactly the density function of $y - x\gamma_o$ conditional on $I_1 = 1$ and $x\alpha_o$, i.e.,

$$\begin{aligned} f^*(y|I_1 = 1, x) &= f(y - x\gamma_o|I_1 = 1, x\alpha_o, \theta_o) \\ &= g(y - x\gamma_o, x\alpha_o|I_1 = 1, \theta_o)/p(x\alpha_o|I_1 = 1, \theta_o). \end{aligned} \tag{2.4}$$

The sample selection model implies the preceding specific index structure. The index structure plays a useful role for semiparametric estimation of econometric models. Taking into account index structures, Powell (1989) proposed a two-stage method and Ichimura and Lee (1991) proposed a semiparametric nonlinear least-squares method for estimation of sample selection models. Several studies, including Horowitz and Hardle (1996), Ichimura (1993), Klein and Spady (1993), Powell et al. (1989), and Ruud (1986), have explored index structures for semiparametric estimation in different contexts. The estimation method proposed in this paper will also take advantage of the index structure. It is a kind of adaptive estimation procedure motivated by the works on adaptive estimation (Stone, 1975; Bickel, 1982; Manski, 1984; and Schick, 1986) for semiparametric models.

Given a random sample of size n, for any possible value θ, the probability function $P_l(x_i, \theta) = E(I_l|x_i\alpha, \theta)$ of I_l conditional on $x\alpha$ evaluated at a point $x_i\alpha$ can be estimated by nonparametric regression (Nadaraya, 1964; Watson, 1964):

$$P_{n,l}(x_i, \theta) = \frac{A_n(I_l|x_i, \theta)}{A_n(1|x_i, \theta)}, \tag{2.5}$$

where

$$A_n(v|x_i, \theta) = \frac{1}{n-1} \sum_{j \neq i}^{n} v_j \frac{1}{a_{1,n}^m} K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right), \tag{2.6}$$

$v = I_1, \ldots, I_L$, or 1, $K(\cdot)$ is a kernel function on R^m, and $a_{1,n} > 0$ is a bandwidth sequence. Let $z_i = (x_i, y_i)$. Then $f(y - x\gamma|I_1 = 1, x\alpha, \theta)$ evaluated at the point $(y_i - x_i\gamma, x_i\alpha)$ can be estimated by

$$f_n(y_i - x_i\gamma|I_{1i} = 1, x_i\alpha) = \frac{C_n(z_i, \theta)}{A_n(I_1|x_i, \theta)}, \tag{2.7}$$

where

$$C_n(z_i, \theta) = \frac{1}{n-1} \sum_{j \neq i}^{n} I_{1j} \frac{1}{a_{2,n}^{m+k}} J\left(\frac{(y_i - x_i\gamma) - (y_j - x_j\gamma)}{a_{2,n}}, \frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right) \tag{2.8}$$

and $J(\cdot)$ is a kernel function on R^{m+k} with a bandwidth $a_{2,n}$. These nonparametric functions can be used to formulate a semiparametric likelihood function:

$$\ln L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{I_{1i}\ln f_n(y_i - x_i\gamma|I_{1i}=1, x_i\alpha) + \sum_{l=1}^{L} I_{li}\ln P_{n,l}(x_i,\theta)\right\}.$$

However, instead of working with this function, we consider a simpler two-step estimation method.[1] In the existing semiparametric literature, various $\sqrt{n}$-consistent estimators exist (see the discussions after the assumptions in the next section for details) for the sample selection model that can be used as the initial estimate for the scoring method. Denote $\epsilon_i(\theta) = y_i - x_i\gamma$. Given an $\sqrt{n}$-consistent estimate $\bar{\theta}_n$ of θ_o, a two-step scoring (TSC) estimator can be defined as

$$\hat{\theta}_n = \bar{\theta}_n + I_n^{-1}(\bar{\theta}_n)\times S_n(\bar{\theta}_n), \qquad \textbf{(2.9)}$$

where

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\theta)\left\{I_{1i}\frac{\partial\ln f_n(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha)}{\partial\theta} + \sum_{l=1}^{L} I_{li}\frac{\partial\ln P_{n,l}(x_i,\theta)}{\partial\theta}\right\} \qquad \textbf{(2.10)}$$

and

$$I_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\theta)\left\{I_{1i}\frac{\partial\ln f_n(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha)}{\partial\theta}\frac{\partial\ln f_n(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha)}{\partial\theta'} + \sum_{l=1}^{L} I_{li}\frac{\partial\ln P_{n,l}(x_i,\theta)}{\partial\theta}\frac{\partial\ln P_{n,l}(x_i,\theta)}{\partial\theta'}\right\}, \qquad \textbf{(2.11)}$$

where $t_n(z_i,\theta)$ is a trimming function that is designed to trim the observations from the tails of the indices for a finite sample.

The function of trimming is to control for the erratic behavior of the nonparametric estimates that appear as denominators in (2.10) and (2.11). The trimming function is chosen to be continuously twice differentiable. Such a smooth trimming function can be defined as follows. Let G be a distribution function with support on the interval $[\frac{1}{2},1]$ such that the derivatives of G at the end points $\frac{1}{2}$ and 1 are zero. A simple G with such a smooth property is

$$G(c) = \int_{-1}^{4c-3}\frac{15}{16}(1-w^2)^2\,dw, \qquad \frac{1}{2}\le c\le 1, \qquad \textbf{(2.12)}$$

where $\frac{15}{16}(1-w^2)^2$ is a density function with support $[-1,1]$. This G is apparently a bounded polynomial function on $[\frac{1}{2},1]$. Define the following trimming function:

$$t^*(c) = \begin{cases} 1 & 1 \le c \\ G(c) & \frac{1}{2} < c < 1 \\ 0 & c \le \frac{1}{2}. \end{cases} \tag{2.13}$$

The denominators in (2.10) and (2.11) at $\bar{\theta}_n$ involve $A_n(I_l | x_i, \bar{\theta}_n)$, $l = 0, 1, \ldots, L$, and $C_n(z_i, \bar{\theta}_n)$. Let $\delta_{j,n} > 0$, $j = 1, 2$ be sequences of positive numbers converging to zero at certain rates to be specified subsequently. Let $I_0 \equiv 1$ to unify notation. The trimming function $t_n(z_i, \theta)$ is defined as

$$t_n(z_i, \theta) = t^*(C_n(z_i, \theta)/\delta_{2,n}) \prod_{l=0}^{L} t^*(A_n(I_l | x_i, \theta)/\delta_{1,n}). \tag{2.14}$$

The trimmed proportion of sample observations goes to zero as n goes to infinity. So the trimming becomes less severe as sample size increases, and, asymptotically, all the sample information can be preserved.[2]

3. LARGE SAMPLE PROPERTIES AND THE ASYMPTOTIC EFFICIENCY

3.1. The Asymptotic Normality

Our proposed estimator can be shown to be consistent and asymptotically normal. To justify these properties, we assume the following regularity conditions for the model.

Assumption 1.

(1) The samples $(I_{1i}, \ldots, I_{Li}, x_i, y_i)$, $i = 1, \ldots, n$ are independent and identically distributed (i.i.d.).
(2) Θ is a compact neighborhood of the true parameter vector θ_o in a finite-dimensional Euclidean space. The mappings $\alpha(\theta)$ and $\gamma(\theta)$ are twice continuously differentiable on Θ.
(3) The density function $f(\epsilon | I_1 = 1, t, \theta)$ and the probability functions $P(I_l = 1 | t, \theta)$, where $t = x\alpha(\theta)$ and $\epsilon = y - x\gamma(\theta)$, are twice continuously differentiable in t and ϵ.
(4) The functions $\sup_{\theta \in \Theta} \|\partial^2/\partial\theta\partial\theta' \ln f(y - x\gamma(\theta) | I_1 = 1, x\alpha(\theta), \theta)\|$ and $\sup_{\theta \in \Theta} \|\partial^2/\partial\theta\partial\theta' \ln P_l(x, \theta)\|$ for all $l = 1, \ldots, L$ have finite first moments.
(5) The information matrix $I(\theta_o)$ is nonsingular, where

$$I(\theta_o) = E\left\{ I_1 \frac{\partial \ln f(\epsilon(\theta_o) | I_1 = 1, x\alpha_o, \theta_o)}{\partial\theta} \frac{\partial \ln f(\epsilon(\theta_o) | I_1 = 1, x\alpha_o, \theta_o)}{\partial\theta'} + \sum_{l=1}^{L} I_l \frac{\partial \ln P_l(x, \theta_o)}{\partial\theta} \frac{\partial \ln P_l(x, \theta_o)}{\partial\theta'} \right\}.$$

(6) $\bar{\theta}_n$ is an $\sqrt{n}$-consistent estimator of θ_o.

Assumption 2.

(1) The support S_x of x is a compact set.
(2) The density functions $p(t|\theta)$, $g(\epsilon,t|\theta)$, the functions $E_\theta(I_l|t)p(t|\theta)$, $l = 1,\ldots,L, E_\theta(I_1|\epsilon,t)g(\epsilon,t|\theta)$, and their derivatives $(\partial/\partial t)p(t|\theta)$, $(\partial/\partial w)g(\epsilon,t)$, $(\partial/\partial t)[E_\theta(I_l|t)p(t|\theta)]$, $l = 1,\ldots,L,(\partial/\partial w)[E_\theta(I_1|\epsilon,t)g(\epsilon,t|\theta)]$ are uniformly continuous in t and ϵ, uniformly in $\theta\in\Theta$, and are uniformly bounded, where $w=(\epsilon,t)$.
(3) The functions $E_\theta(x|t)p(t|\theta)$, $E_\theta(I_l x|t)p(t|\theta)$, $l = 1,\ldots,L$, $E_\theta(I_1 x|\epsilon,t)g(\epsilon,t|\theta)$ and their derivatives $(\partial/\partial t)[E_\theta(x|t)p(t|\theta)]$, $(\partial/\partial t)[E_\theta(I_l x|t)p(t|\theta)]$, $l = 1,\ldots,L$, $(\partial/\partial w)E_\theta(I_1 x|\epsilon,t)g(\epsilon,t|\theta)$ are uniformly continuous in t and ϵ, uniformly in $\theta \in \Theta$, and are uniformly bounded.

Assumption 3.

(1) The kernel function $K(t)$ on R^m is a function with bounded support such that $\int_{R^m} K(t)\,dt = 1$ and $\int_{R^m} |K(t)|\,dt < \infty$. Similarly, $J(w)$ is a kernel function with bounded support on R^{k+m}.
(2) The kernel functions $K(t)$ and $J(w)$ are twice differentiable, and their first- and second-order derivatives are bounded. The terms K, J and their first-order derivatives have bounded variation.
(3) The $K(t)$ is a high-order kernel function with zero moments up to some finite order s_1^*, and $J(w)$ has zero moments up to some finite order s_2^*.
(4) The bandwidth sequence $\{a_{1,n}\}$ for $K(t)$ and $\{\delta_{1,n}\}$ are chosen such that $\lim_{n\to\infty}(n/\ln^4 n)a_{1,n}^{2(m+1)}\delta_{1,n}^4 = \infty$ and $\lim_{n\to\infty} na_{1,n}^{2s_1^*}/\delta_{1,n}^2 = 0$.
(5) The bandwidth sequence $\{a_{2,n}\}$ for $J(w)$ and $\{\delta_{2,n}\}$ are chosen such that $\lim_{n\to\infty}(n/\ln^4 n)a_{2,n}^{2(m+k+1)}\delta_{2,n}^4 = \infty$ and $\lim_{n\to\infty} na_{2,n}^{2s_2^*}/\delta_{2,n}^2 = 0$.

Assumption 4.

(1) The density functions $p(t|\theta)$ and $g(\epsilon,t|\theta)$ are differentiable in t and ϵ, respectively, to the orders s_1^* and s_2^*, and these derivatives are all bounded uniformly in $\theta\in\Theta$.
(2) The functions $E_\theta(I_l|t)p(t|\theta), E_\theta(I_l x|t)p(t|\theta), E_\theta(x|t)p(t|\theta)$ and their derivatives $(\partial/\partial t)[E_\theta(x|t)p(t|\theta)], (\partial/\partial t)[E_\theta(I_l|t)p(t|\theta)], (\partial/\partial t)[E_\theta(I_l x|t)p(t|\theta)], l=1,\ldots,L$, are differentiable in t to the order s_1^*+1, and these derivatives are bounded uniformly in $\theta\in\Theta$.
(3) The functions $E_\theta(I_1|\epsilon,t)g(\epsilon,t|\theta)$, $E_\theta(I_1 x|\epsilon,t)g(\epsilon,t|\theta)$ and their derivatives $(\partial/\partial w)[E_\theta(I_1|\epsilon,t)g(\epsilon,t|\theta)]$ and $(\partial/\partial w)[E_\theta(I_1 x|\epsilon,t)g(\epsilon,t|\theta)]$ are differentiable in (ϵ,t) to the order s_2^*+1, and these derivatives are bounded uniformly in $\theta\in\Theta$.

The conditions in Assumption 1 are some basic regularity conditions for the model that justify some of the fundamental relations on the likelihood functions. The initial $\sqrt{n}$-consistent estimate can be constructed in several ways. For example, for the model with binary choice, we can estimate the choice equation by Ichimura's semiparametric nonlinear least-squares method (Ichimura, 1993) and the maximum likelihood method of Klein and Spady (1993) or by the noniterative approaches of Horowitz and Hardle (1996) and Powell et al. (1989) and the outcome equation by two-stage methods of Andrews (1991), Newey (1988), and

Powell (1989). Alternatively, we can estimate all the parameters simultaneously by the semiparametric nonlinear least-squares method in Ichimura and Lee (1991). For models with polychotomous choices, the choice equations can be estimated by a semiparametric likelihood method in Lee (1995). Note that by choosing noniterative estimators as our initial estimates, implementation of our procedure could avoid difficult nonlinear optimization problems. The nonsingular information matrix assumption is an identification condition. Identification of the sample selection model has been considered in Chamberlain (1986), Powell (1989) and Ichimura and Lee (1991). Essentially, it requires that the regressors in the choice equations have variables that are not contained in the outcome equations. In addition, the indices in the polychotomous choice case need to be distinguishable from each other.

Assumption 2(1) can be relaxed at the cost of slower rates of convergence for some nonparametric estimates; thus stronger smoothness and boundedness conditions will be needed. It should be noted that compact support for $y - x\gamma(\theta)$ is not imposed by any of the assumptions. The rest of Assumption 2 and Assumption 4 is mainly boundedness and smoothness conditions.

The asymptotic properties of the estimator depend on some convergence properties and the orders of asymptotic biases of the kernel estimates. Proper bandwidth rates are chosen to guarantee convergence of the kernel estimates. Assumption 3(4) can be relaxed slightly but is made for convenience (see comment on this rate directly following equation (A.4) in Appendix A). High-order kernels are needed to obtain the desired small order of asymptotic biases. The uniform continuity and boundedness conditions guarantee that the convergences of the kernel functions are uniform. These regularity conditions can be implied by some basic conditions on the joint density function of the dependent and explanatory variables in the model. These conditions, however, are more direct. As the focus here is on efficient estimation and a first-step $\sqrt{n}$-consistent estimator is utilized to construct the score estimator, its consistency is not discussed separately. Actually, weaker smoothness conditions and regular kernels will be sufficient to ensure the consistency of the score estimator, as its proof only requires uniform convergence of the relevant nonparametric functionals in the construction of the score estimator and the consistency of the first-step estimator.

THEOREM 1. *Under Assumptions* 1–4, *the TSC estimator* $\hat{\theta}_n$ *is consistent and asymptotically normal*,

$$\sqrt{n}(\hat{\theta}_n - \theta_o) \xrightarrow{D} N(0, I^{-1}(\theta_o)).$$

Proof. The limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_o)$ can be derived from Taylor's expansion. By Taylor's expansion of $S_n(\bar{\theta}_n)$ at θ_o, $S_n(\bar{\theta}_n) = S_n(\theta_o) + (\partial S_n(\bar{\bar{\theta}}_n)/\partial\theta')(\bar{\theta}_n - \theta_o)$ where $\bar{\bar{\theta}}_n$ lies between $\bar{\theta}_n$ and θ_o. Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_o) = \sqrt{n}(\bar{\theta}_n - \theta_o) + \sqrt{n} I_n^{-1}(\bar{\theta}_n)\left[S_n(\theta_o) + \frac{\partial S_n(\bar{\bar{\theta}}_n)}{\partial \theta'}(\bar{\theta}_n - \theta_o)\right]$$
$$= \left[I + I_n^{-1}(\bar{\theta}_n)\frac{\partial S_n(\bar{\bar{\theta}}_n)}{\partial \theta'}\right] \times \sqrt{n}(\bar{\theta}_n - \theta_o) + I_n^{-1}(\bar{\theta}_n) \times \sqrt{n} S_n(\theta_o), \tag{3.1}$$

where I is an identity matrix. The asymptotic distribution of $\hat{\theta}_n$ can be derived in several steps. First of all, under the regularity conditions and the identification condition that $I(\theta_o)$ is nonsingular, we show in Sections B.2 and B.3 of Appendix B that

$$I_n^{-1}(\bar{\theta}_n) \xrightarrow{p} I^{-1}(\theta_o) \tag{3.2}$$

and

$$\frac{\partial S_n(\bar{\bar{\theta}}_n)}{\partial \theta'} \xrightarrow{p} -I(\theta_o), \tag{3.3}$$

and, therefore, $[I + I_n^{-1}(\bar{\theta}_n)(\partial S_n(\bar{\bar{\theta}}_n)/\partial \theta')] \times \sqrt{n}(\bar{\theta}_n - \theta_o) \xrightarrow{p} 0$ by Slutsky's lemma. From these, it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta_o) = I_n^{-1}(\theta_o)\sqrt{n} S_n(\theta_o) + o_p(1). \tag{3.4}$$

The final step is to show that $\sqrt{n} S_n(\theta_o) \xrightarrow{D} N(0, I(\theta_o))$, which is the most complicated part of the analysis. For that purpose, define

$$S_n^*(\theta_o) = \frac{1}{n}\sum_{i=1}^{n} t_n^{\infty}(z_i, \theta_o)\left\{ I_{1i}\frac{\partial \ln f(\epsilon_i(\theta_o) | I_{1i} = 1, x_i \alpha_o, \theta_o)}{\partial \theta} + \sum_{l=1}^{L} I_{li}\frac{\partial \ln P_l(x_i, \theta_o)}{\partial \theta}\right\} \tag{3.5}$$

with

$$t_n^{\infty}(z_i, \theta) = t^*(C(z_i, \theta)/\delta_{2,n}) \prod_{l=0}^{L} t^*(A(I_l | x_i, \theta)/\delta_{1,n}). \tag{3.6}$$

In Section B.1 of Appendix B we show that

$$\sqrt{n}[S_n(\theta_o) - S_n^*(\theta_o)] \xrightarrow{p} 0 \tag{3.7}$$

and

$$\sqrt{n} S_n^*(\theta_o) \xrightarrow{D} N(0, I(\theta_o)). \tag{3.8}$$

Combining these results,

$$\sqrt{n}(\hat{\theta}_n - \theta_o) \xrightarrow{D} N(0, I^{-1}(\theta_o)). \tag{3.9}$$

From the proof of the theorem we can see that $I_n^{-1}(\bar{\theta}_n)$ is a consistent estimator for $I^{-1}(\theta_o)$.

3.2. The Asymptotic Efficiency

In this subsection, we investigate the efficiency of the TSC estimator. For the binary choice sample selection model, Chamberlain (1986) has derived the asymptotic efficiency bound for semiparametric estimators of that model. From the asymptotic distribution in (3.9) one can see that the TSC estimator attains Chamberlain's efficiency bound. To investigate the efficiency of our estimator for the polychotomous choice sample selection model with index restrictions, we need to derive the semiparametric bound for such a model.[3] For an easy exposition, we will follow the discussion in Newey (1990). First, let us review some of the terminology as in Newey (1990). A parametric submodel of a semiparametric model is a parametric model that satisfies the semiparametric assumptions and contains the truth. A parametric submodel is said to be regular if the likelihood function satisfies some general differentiable conditions and its information matrix is nonsingular. The semiparametric asymptotic variance bound V for a semiparametric model is defined as the supremum of the Cramér–Rao bounds for all regular parametric submodels. The intuition behind this efficient criterion is based on Stein's idea (1956) that a parametric maximum likelihood estimator for any parametric submodel should be at least as efficient as any semiparametric estimator. An estimator $\tilde{\theta}$ is said to be regular in a parametric submodel if, for each θ_o, $\sqrt{n}(\tilde{\theta} - \theta_n)$ has a limiting distribution that does not depend on any local data-generating process where, for each sample of size n, the data are distributed according to θ_n where $\sqrt{n}(\theta_n - \theta_o)$ is bounded. An estimator $\tilde{\theta}$ is regular for a semiparametric model if it is regular in every regular parametric submodel and the limiting distribution does not depend on the parametric submodel. The semiparametric bound V is applied to the class of regular estimators. The class of regular estimators excludes superefficient estimators and estimators using more information that is not contained in the semiparametric model. Theorem 2.2 in Newey (1990) provides a powerful characterization for a semiparametric estimator to be regular. A way to derive the semiparametric bound for a model is to derive the relevant tangent set and its efficient score S. Let $\delta = (\theta', \eta')'$ be the parameters of a submodel and let $S_\delta = (S_\theta', S_\eta')'$ be the score vector. A tangent set $\mathcal{I}$ is defined as the mean-squares closure of all linear combinations of scores S_η for parametric submodels. The efficient score S is the (unique) vector such that

$$S_\theta - S \in \mathcal{I} \quad \text{and} \quad E(S't) = 0, \quad \forall t \in \mathcal{I}.$$

The semiparametric variance bound is shown to be $V = (E[SS'])^{-1}$ (Theorem 3.2 in Newey [1990]).

For the polychotomous choice sample selection model with index restrictions, a parametric submodel corresponds to a parametric family of conditional choice probabilities $P_l(x\alpha, \delta)$, $l = 1, \ldots, L$, and the conditional density $f(y - x\gamma | I_1 = 1, x\alpha, \delta)$, where $\delta = (\theta', \eta')'$ and η is an additional finite-dimensional parameter (column) vector. As an index model, x appears only through the index vectors $(x\alpha, x\gamma)$. As a convention, we may assume that it contains the true model at $\eta = 0$, i.e., $P_l(x\alpha, (\theta', 0)') = P_l(x, \theta)$, $\forall l$, and $f(y - x\gamma | I_1 = 1, x\alpha, (\theta', 0)') = f(y - x\gamma | I_1 = 1, x\alpha, \theta)$. The log likelihood function for a single observation for this parametric submodel is $\ln L(\delta | I, x) = \sum_{l=1}^{L} I_l \ln P_l(x\alpha, \delta) + I_1 \ln f(y - x\gamma | I_1 = 1, x\alpha, \delta) + g(x, \eta)$, where $g(x, \eta)$ refers to the marginal density (probability) function of x of the parametric submodel. Consider a set of submodels of the form $P_l(x\alpha, \delta) = P_l(x, \theta) + \eta q_l(x\alpha)$ and $f(y - x\gamma | I_1 = 1, x\alpha, \theta) + \eta q^*(y - x\gamma, x\alpha)$, where $q^*(y - x\gamma, x\alpha)$, $q_l(x\alpha)$, $l = 1, \ldots, L$ are continuously differentiable with bounded supports such that $\sum_{l=1}^{L} q_l(x\alpha) = 1$ and $\int q^*(y - x\gamma, x\alpha)\, dy = 1$ (in different contexts, see Chamberlain, 1986; Newey, 1990, p. 111). Therefore the score vector $S_\delta = (S'_\theta, S'_\eta)'$ is $S_\theta = \sum_{l=1}^{L} I_l (1/P_l(x, \theta_o))(\partial P_l(x, \theta_o)/\partial\theta) + I_1 (\partial/\partial\theta) \ln f(y - x\gamma_o | I_1 = 1, x\alpha_o, \theta_o)$ and $S_\eta = \sum_{l=1}^{L} I_l (1/P_l(x, \theta_o) q_l(x\alpha_o) + I_1 (1/f(y - x\gamma_o | I_1 = 1, x\alpha_o, \theta_o)) q^*(y - x\gamma_o, x\alpha_o) + (\partial \ln g(x, 0))/\partial\eta$. As the functions $q_l(x\alpha_o)$ and $q^*(y - x\gamma_o, x\alpha_o)$ are unrestricted except for certain normalization conditions and they are functions of indices, the tangent set for the model is

$$\mathcal{I} = \left\{ \sum_{l=1}^{L} I_l \frac{1}{P_l(x, \theta_o)} D_l(x\alpha_o) + I_1 \frac{1}{f(y - x\gamma_o | I_1 = 1, x\alpha_o, \theta_o)} C(y - x\gamma_o, x\alpha_o) \right.$$
$$\left. + D(x) : \sum_{l=1}^{L} D_l(x\alpha_o) = 0, \int C(y - x\gamma_o, x\alpha_o)\, dy = 0, E(D(x)) = 0 \right\}.$$

With $\partial/\partial\theta \ln f(y - x\gamma_o | I_1 = 1, x\alpha_o, \theta_o)$ and $(\partial \ln P_l(x, \theta_o))/\partial\theta$ in equations (B.13) and (B.14) of Appendix B, it is straightforward to verify that the efficient score for our model is $S = I_1(\partial \ln f(\epsilon(\theta_o) | I_1 = 1, x\alpha_o, \theta_o)/\partial\theta) + \sum_{l=1}^{L} I_l (\partial \ln P_l(x, \theta_o))/\partial\theta$ by checking $S_\theta - S \in \mathcal{I}$ and $E(S't) = 0$, $\forall t \in \mathcal{I}$. Therefore, the semiparametric asymptotic variance bound is

$$\begin{aligned} V &= [E(SS')]^{-1} \\ &= E\left\{ I_1 \frac{\partial \ln f(\epsilon(\theta_o) | I_1 = 1, x\alpha_o, \theta_o)}{\partial\theta} \frac{\partial \ln f(\epsilon(\theta_o) | I_1 = 1, x\alpha_o, \theta_o)}{\partial\theta'} \right. \\ &\qquad \left. + \sum_{l=1}^{L} I_l \frac{\partial \ln P_l(x, \theta_o)}{\partial\theta} \frac{\partial \ln P_l(x, \theta_o)}{\partial\theta'} \right\}^{-1}. \end{aligned}$$

Also note that S is the influence function for our TSC estimator by (B.15). The regularity of the TSC estimator follows trivially from Newey (1990, Theorem 2.2). Consequently the TSC estimator attains the semiparametric efficiency bound.

4. A SMALL MONTE CARLO STUDY

In this section we perform some Monte Carlo simulations to investigate the finite-sample performance of our estimator.

Simulated data are generated from the following model: $I_1 = 1_{\{x_1 + x_2\alpha + u_1 > 0\}}$ and $y = I_1(z_1\gamma_1 + x_2\gamma_2 + u_2)$, where $1_{\{\cdot\}}$ is the usual indicator function. The true parameters are $\alpha = 1$ and $\gamma_1 = \gamma_2 = 1$. The regressors x_1 and z_1 are randomly drawn from the standard normal distribution, and x_2 is drawn from a uniform $(-2,2)$, all of which are independent of each other. Three different designs are considered by varying the distribution of the error term (u_1, u_2). Correlation between u_1 and u_2 is introduced by setting $u_2 = \sqrt{0.5}u_1 + \sqrt{0.5}N(0,1)$. In the first design u_1 is drawn from an $N(0,1)$ (normal), u_1 is a standardized normal mixture in the second design (nonnormal), and u_1 is a heteroskedastic normal (HN) error term equal to an $N(0,1)(1 + 0.5(x_1 + x_2))$ in the third design.

Here we consider the results of estimating α and γ by comparing our estimator (CL) with the estimators proposed by Klein and Spady (1993, KL) and Powell (1989, PW). We use the standard normal density function as the kernel function for both KL and PW estimators as higher-order kernels do not lead to improved finite-sample performance for the sample sizes under consideration here. We use the product kernel based on the standard normal density function for our two-step estimator. It is difficult to propose optimal bandwidths for the estimators considered here. A generalized cross-validation criterion is used in selecting the bandwidths for KL (based on the likelihood criterion) and PW (based on the least-squares criterion) estimators. For the bandwidths a_{n1} and a_{n2} associated with our estimator, we use the bandwidth chosen for the KL estimator as the first bandwidth a_{n1}, whereas a_{n2} is selected based on likelihood cross-validation. Similar to the results reported in Klein and Spady (1993), trimming has almost no effect on the estimates, and thus untrimmed estimates are reported here.

The results from 300 replications with sample sizes 100 and 300 are presented in Tables 1 and 2. For each of the estimators we report the mean value (mean), the standard deviation (SD), and the root mean square error (RMSE).

In Table 1 we report the simulations when the sample size is equal to 100. Compared with the KL estimator, our efficient estimator improves in terms of bias and variance performance. However, the improvement over the PW estimator is almost negligible. Simulation results with sample size equal to 300 are reported in Table 2, which displays the same pattern reported in Table 1 with reduced RMSE's.

Table 1. Sample size equal to 100

Estimators	True	Mean	SD	RMSE
Normal				
α_{kl}	1.000	1.038	0.287	0.289
γ_{1pw}	1.000	1.002	0.141	0.141
γ_{2pw}	1.000	0.977	0.209	0.210
α_{cl}	1.000	1.017	0.279	0.279
γ_{1cl}	1.000	1.002	0.140	0.140
γ_{2cl}	1.000	0.975	0.208	0.209
Nonnormal				
α_{kl}	1.000	1.053	0.241	0.246
γ_{1pw}	1.000	0.993	0.157	0.157
γ_{2pw}	1.000	0.994	0.205	0.205
α_{cl}	1.000	1.033	0.234	0.236
γ_{1cl}	1.000	0.994	0.156	0.156
γ_{2cl}	1.000	0.993	0.204	0.204
HN				
α_{kl}	1.000	1.055	0.277	0.282
γ_{1pw}	1.000	1.007	0.198	0.197
γ_{2pw}	1.000	0.988	0.231	0.231
α_{cl}	1.000	1.035	0.270	0.272
γ_{1cl}	1.000	1.007	0.197	0.197
γ_{2cl}	1.000	0.987	0.230	0.230

5. ESTIMATION OF SEQUENTIAL CHOICE SAMPLE SELECTION MODELS

The method can be easily generalized to the estimation of sequential choice models with selectivity. Consider the model with two-stage decisions and a single continuous outcome equation. There are L_1 alternatives in the first-stage decision. If the first alternative was chosen in the first decision, the second-stage decision would involve L_2 alternatives. The disturbances in these decisions can be correlated. Define the L_2 mutually exclusive and exhaustive dichotomous indicators J_{li} for individual i where $J_{li} = 1$ if the l alternative in the second-stage decision is chosen and 0 otherwise. The continuous dependent variable y can be observed only when $J_1 = 1$. The observations will be $(I_i, J_i, J_{1i} y_i), i = 1, \ldots, n$ where $I_i = (I_{1i}, \ldots, I_{L_1 i})$ and $J_i = (J_{1i}, \ldots, J_{L_2 i})$. Suppose that $x\alpha$ is a vector of dimension m_1, $x\delta$ is a vector of dimension m_2, and $\epsilon(\theta) = y - x\gamma$ is a variable; let $K(\cdot), \overline{K}(\cdot)$, and $J(\cdot)$ be kernel functions on R^{m_1}, $R^{m_1+m_2}$, and $R^{m_1+m_2+1}$, respectively. Define

Table 2. Sample size equal to 300

Estimators	True	Mean	SD	RMSE
Normal				
α_{kl}	1.000	1.010	0.152	0.153
γ_{1pw}	1.000	0.997	0.085	0.084
γ_{2pw}	1.000	0.998	0.125	0.125
α_{cl}	1.000	1.004	0.151	0.151
γ_{1cl}	1.000	0.997	0.084	0.084
γ_{2cl}	1.000	0.997	0.125	0.125
Nonnormal				
α_{kl}	1.000	1.021	0.121	0.123
γ_{1pw}	1.000	0.993	0.084	0.084
γ_{2pw}	1.000	0.992	0.120	0.120
α_{cl}	1.000	1.015	0.120	0.121
γ_{1cl}	1.000	0.993	0.084	0.084
γ_{2cl}	1.000	0.991	0.119	0.119
HN				
α_{kl}	1.000	1.020	0.136	0.137
γ_{1pw}	1.000	1.007	0.099	0.099
γ_{2pw}	1.000	1.002	0.124	0.124
α_{cl}	1.000	1.014	0.135	0.135
γ_{1cl}	1.000	1.007	0.099	0.099
γ_{2cl}	1.000	1.001	0.124	0.123

$$A_n(I_l|x_i,\theta) = \frac{1}{n-1}\sum_{j\neq i}^{n} I_{lj}\frac{1}{a_{1,n}^{m_1}}K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right), \qquad l=1,\ldots,L_1,$$

$$A_n(1|x_i,\theta) = \frac{1}{n-1}\sum_{j\neq i}^{n} \frac{1}{a_{1,n}^{m_1}}K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right),$$

$$\bar{A}_n(J_l|x_i,\theta) = \frac{1}{n-1}\sum_{j\neq i}^{n} I_{1j}J_{lj}\frac{1}{a_{2,n}^{m_1+m_2}}\bar{K}\left(\frac{x_i\alpha - x_j\alpha}{a_{2,n}}, \frac{x_i\delta - x_j\delta}{a_{2,n}}\right),$$

$$l=1,\ldots,L_2,$$

$$\bar{A}_n(1|x_i,\theta) = \frac{1}{n-1}\sum_{j\neq i}^{n} I_{1j}\frac{1}{a_{2,n}^{m_1+m_2}}\bar{K}\left(\frac{x_i\alpha - x_j\alpha}{a_{2,n}}, \frac{x_i\delta - x_j\delta}{a_{2,n}}\right),$$

and

$$C_n(z_i,\theta) = \frac{1}{n-1}\sum_{j\neq i}^{n} I_{1j}J_{1j}\frac{1}{a_{3,n}^{m_1+m_2+1}} \times J\left(\frac{\epsilon_i(\theta)-\epsilon_j(\theta)}{a_{3,n}}, \frac{x_i\alpha - x_j\alpha}{a_{3,n}}, \frac{x_i\delta - x_j\delta}{a_{3,n}}\right).$$

The probabilities $P(I_l = 1|x_i\alpha)$ and $P(J_j = 1|I_{1i} = 1, x_i\alpha, x_i\delta)$ can be estimated by $P_{n,l}(x_i,\theta) = (A_n(I_l|x_i,\theta))/(A_n(1|x_i,\theta))$ and $\bar{P}_{n,l}(x_i,\theta) = (\bar{A}_n(J_l|x_i,\theta))/(\bar{A}_n(1|x_i,\theta))$, respectively. The conditional density function $f(\epsilon_i(\theta)|I_{1i}J_{1i} = 1, x_i\alpha, x_i\delta, \theta)$ can be estimated by $f_n(\epsilon_i(\theta)|I_{1i}J_{1i} = 1, x_i\alpha, x_i\delta) = (C_n(z_i,\theta))/(\bar{A}_n(J_1|x_i,\theta))$. The log (pseudo) likelihood function for the sample observations will be

$$\ln L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{l=1}^{L_1} I_{li}\ln P_{n,l}(x_i,\theta) + I_{1i}\sum_{j=1}^{L_2} J_{ji}\ln \bar{P}_{n,j}(x_i,\theta) + I_{1i}J_{1i}\ln f_n(\epsilon_i(\theta)|I_{1i}J_{1i} = 1, x_i\alpha, x_i\delta, \theta)\right].$$

The corresponding TSC estimator will be $\hat{\theta}_n = \bar{\theta}_n + I_n^{-1}(\bar{\theta}_n)S_n(\bar{\theta}_n)$, where

$$S_n(\bar{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\bar{\theta}_n)\left\{I_{1i}J_{1i}\frac{\partial \ln f_n(\epsilon_i(\bar{\theta}_n)|I_{1i}J_{1i} = 1, x_i\alpha(\bar{\theta}_n), x_i\delta(\bar{\theta}_n))}{\partial\theta} + I_{1i}\sum_{l=1}^{L_2} J_{li}\frac{\partial \ln \bar{P}_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta} + \sum_{l=1}^{L_1} I_{li}\frac{\partial \ln P_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta}\right\},$$

$$I_n(\bar{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\bar{\theta}_n)\left\{I_{1i}J_{1i}\frac{\partial \ln f_n(\epsilon_i(\bar{\theta}_n)|I_{1i}J_{1i} = 1, x_i\alpha(\bar{\theta}_n), x_i\delta(\bar{\theta}_n))}{\partial\theta} \times \frac{\partial \ln f_n(\epsilon_i(\bar{\theta}_n)|I_{1i}J_{1i} = 1, x_i\alpha(\bar{\theta}_n), x_i\delta(\bar{\theta}_n))}{\partial\theta'} + I_{1i}\sum_{l=1}^{L_2} J_{li}\frac{\partial \ln \bar{P}_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta}\frac{\partial \ln \bar{P}_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta'} + \sum_{l=1}^{L_1} I_{li}\frac{\partial \ln P_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta}\frac{\partial \ln P_{n,l}(x_i,\bar{\theta}_n)}{\partial\theta'}\right\},$$

with $t_n(z_i,\bar{\theta}_n)$ equal to $t^*(C_n(z_i,\bar{\theta}_n)/\delta_{3,n})\prod_{l=0}^{L_1} t^*(A_n(I_l|x_i,\bar{\theta}_n)/\delta_{1,n})\prod_{l=0}^{L_2} t^*(\bar{A}_n(J_l|x_i,\bar{\theta}_n)/\delta_{2,n})$. As

$$\begin{aligned}
&\frac{\partial \ln f(\epsilon(\theta_o)|I_1J_1 = 1, x\alpha_o, x\delta_o,\theta_o)}{\partial\theta}\\
&\quad = \left(E\left(\left[\frac{\partial x\gamma_o}{\partial\theta}, -\frac{\partial x\alpha_o}{\partial\theta}, -\frac{\partial x\delta_o}{\partial\theta}\right]\middle| x\alpha_o, x\delta_o\right)\right.\\
&\qquad \left. - \left[\frac{\partial x\gamma_o}{\partial\theta}, -\frac{\partial x\alpha_o}{\partial\theta}, -\frac{\partial x\delta_o}{\partial\theta}\right]\right)\\
&\qquad \times \bar{\bar{\nabla}} \ln f(\epsilon(\theta_o)|I_1J_1 = 1, x\alpha_o, x\delta_o,\theta_o),\\
&\frac{\partial \ln \bar{P}_l(x,\theta_o)}{\partial\theta} = \left(\left[\frac{\partial x\alpha_o}{\partial\theta}, \frac{\partial x\delta_o}{\partial\theta}\right] - E\left(\left[\frac{\partial x\alpha_o}{\partial\theta}, \frac{\partial x\delta_o}{\partial\theta}\right]\middle| x\alpha_o, x\delta_o\right)\right)\\
&\qquad \times \nabla \ln P(J_l = 1|x\alpha_o, x\delta_o),
\end{aligned}$$

and $(\partial \ln P_l(x,\theta_o))/\partial\theta = [(\partial x\alpha_o/\partial\theta) - E((\partial x\alpha_o/\partial\theta)|x\alpha_o)]\nabla \ln P(I_l = 1|x\alpha_o)$, from the appendixes, under similar regularity conditions, $\sqrt{n}(\hat{\theta}_n - \theta_o) \xrightarrow{D} N(0, I^{-1}(\theta_o))$, where

$$\begin{aligned}
I(\theta_o) = E\Bigg\{ & I_1J_1 \frac{\partial \ln f(\epsilon(\theta_o)|I_1J_1 = 1, x\alpha_o, x\delta_o,\theta_o)}{\partial\theta}\\
&\times \frac{\partial \ln f(\epsilon(\theta_o)|I_1J_1 = 1, x\alpha_o, x\delta_o,\theta_o)}{\partial\theta'}\\
&+ I_1\sum_{l=1}^{L_2} J_l \frac{\partial \ln \bar{P}_l(x,\theta_o)}{\partial\theta}\frac{\partial \ln \bar{P}_l(x,\theta_o)}{\partial\theta'}\\
&+ \sum_{l=1}^{L_1} I_l \frac{\partial \ln P_l(x,\theta_o)}{\partial\theta}\frac{\partial \ln P_l(x,\theta_o)}{\partial\theta'}\Bigg\}.
\end{aligned}$$

6. CONCLUSION

In this article, we have considered semiparametric efficient estimation of sample selection models under index restrictions. A two-step semiparametric scoring procedure is proposed. The estimator attains the semiparametric efficiency bound for the polychotomous choice sample selection model under the index restriction. Further, it also attains the efficiency bound under the independence assumption for the binary choice sample selection model. There are complications for the semiparametric likelihood-based method when the densities of the error term and the linear index in the selection equations approach zero. A carefully designed trimming scheme is introduced, combined with special structures of the index restriction to ensure the efficiency of the estimator. A small Monte Carlo study is carried out to illustrate the usefulness of the estimator. The method can be extended to sequential choice sample selection models.

As the independence restriction is slightly stronger than the index restriction, it will be desirable to determine the semiparametric efficiency bound for the polychotomous choice sample selection model under the independence restriction and propose an estimator that attains the efficiency bound. A less ambitious goal is to determine whether there exists an estimator under the independence restriction that is more efficient than the estimator proposed in the paper. It is a topic for future research.

NOTES

1. There are technical issues on trimming indices that make it difficult to work directly with this semiparametric likelihood function.

2. In an earlier version of this paper (Lee, 1990), the density functions of indexes are assumed to be bounded away from zero. In that case, trimming is not needed and the proofs can be simpler. However, such an assumption is too strong to be practical.

3. General semiparametric efficiency bounds for polychotomous choice models under independence restrictions are difficult to find (see Thompson, 1993).

REFERENCES

Ahn, H. & J.L. Powell (1993) Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–29.

Ai, C. (1997) A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–963.

Andrews, D.W.K. (1991) Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–345.

Bickel, P.J. (1982) On adaptive estimation. *Annals of Statistics* 10, 647–671.

Chamberlain, G. (1986) Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics* 32, 189–218.

Cosslett, S. (1983) Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51, 765–782.

Cosslett, S. (1991) Distribution-free estimation of a regression model with sample selectivity. In W.A. Barnett, J. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge: Cambridge University Press.

Donald, S.G. (1995) Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* 65, 347–380.

Dubin, J.A. & D.L. McFadden (1984) An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52, 345–362.

Gallant, A.R. & D.W. Nychka (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica* 55, 363–390.

Gronau, R. (1974) Wage comparisons—a selectivity bias. *Journal of Political Economy* 82, 1119–1143.

Han, A.K. (1987) Non-parametric analysis of a generalized regression model. *Journal of Econometrics* 35, 303–316.

Heckman, J. (1974) Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–694.

Horowitz, J.L. (1992) A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–531.

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

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

Ichimura, H. & L.F. Lee (1991) Semiparametric estimation of multiple index models: Single equation estimation. In W.A. Barnett, J. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge: Cambridge University Press.

Klein, R.W. & R.H. Spady (1993) An efficient semiparametric estimator for discrete choice models. *Econometrica* 61, 387–421.

Lee, L.F. (1978) Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19, 415–433.

Lee, L.F. (1982) Some approaches to the correction of selectivity bias. *Review of Economic Studies* 49, 355–372.

Lee, L.F. (1990) Efficient Semiparametric Scoring Estimation of Sample Selection Models. Discussion paper 253, University of Minnesota.

Lee, L.F. (1995) Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics* 65, 381–428.

Manski, C.F. (1984) Adaptive estimation of non-linear regression models. *Econometric Reviews* 3, 145–194.

Manski, C.F. (1985) Semiparametric analysis of discrete responses: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–334.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics.* New York: Academic.

Nadaraya, E.A. (1964) On estimating regression. *Theory of Probability and Its Applications* 10, 186–190.

Newey, W.K. (1988) Two Step Estimation of Sample Selection Models. Manuscript, Department of Economics, Princeton University.

Newey, W.K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1058.

Pollard, D. (1984) *Convergence of Stochastic Process.* New York: Springer-Verlag.

Powell, J.L. (1989) Semiparametric estimation of bivariate latent variable models. Manuscript, University of Wisconsin.

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Robinson, P.M. (1988) Root-n-consistent semiparametric regression. *Econometrica* 56, 931–954.

Ruud, P.A. (1986) Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics Annals* 32, 157–187.

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

Sherman, R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123–138.

Stein, C. (1956) Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 187–195. Berkeley: University of California Press.

Stone, C. (1975) Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 3, 267–284.

Thompson, T.S. (1993) Some efficiency bounds for semiparametric discrete choice models. *Journal of Econometrics* 58, 257–274.

Watson, G.S. (1964) Smooth regression analysis. *Sankhya, Series A* 26, 359–372.

APPENDIX A: SOME USEFUL LEMMAS

The asymptotic properties of the TSC estimator depend on some convergence properties of the nonparametric functions $A_n(I_l|x_i,\theta), C_n(z_i,\theta)$ and their derivatives with respect to θ. In this section we discuss three propositions and one lemma in preparation for the proof of the main theorem. The proofs of Propositions 1 and 2 can be found in Ichimura and Lee (1991).

PROPOSITION 1. *Let $K(v)$ be a function on R^m with a bounded support D such that $\int_D |K(v)|\,dv < \infty$. Let $t(z,\theta)$ be a continuous m-dimensional random vector. Suppose that $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$, where $g(t|\theta)$ is the density function of $t(z,\theta)$, is uniformly continuous in t, uniformly in (θ,z_i). (A function $g(t,z_i,\theta)$ is said to be uniformly continuous in t uniformly in (θ,z_i) if $\forall\epsilon > 0$ and there exists a $\delta > 0$ (may depend on ϵ only) such that whenever $\|t_1 - t_2\| \le \delta$, $\|g(t_1,z_i,\theta) - g(t_2,z_i,\theta)\| \le \epsilon$ for all (z_i,θ).) Then*

$$\lim_{n\to\infty} \sup_{z_i,\theta} \left\| E\left[c(z,z_i,\theta)\,\frac{1}{a_n^m}\,K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)\middle| z_i\right] - E[c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta]g(t(z_i,\theta)|\theta)\right\| = 0.$$

Furthermore, if $K(v)$ is a function with zero moments up to the order s^*, i.e., $\int_D v_1^{i_1}\cdots v_m^{i_m}K(v)\,dv = 0$, for all $i_j \ge 0, j = 1,\dots,m$, $i_1 + \cdots + i_m < s^*$ and $\int_D \|v\|^{s^*}|K(v)|dv < \infty$, and $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$ is differentiable on R^m with respect to t to the order s^*, and all the derivatives are uniformly bounded, then

$$\sup_{z_i,\theta} \left\| E\left[c(z,z_i,\theta)\,\frac{1}{a_n^m}\,K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)\middle| z_i\right] - E[c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta]g(t(z_i,\theta)|\theta)\right\| = O(a_n^{s^*}).$$

PROPOSITION 2. *Let $K(v)$ be a function on R^m with a bounded support D such that $K(v)$ goes to zero at the boundary of D and its gradient $(\partial K(v))/\partial v$ is bounded. Suppose that $(\partial/\partial t)[E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)]$, where $g(t|\theta)$ is the density function of $t(z,\theta)$, are uniformly continuous in t, uniformly in (z_i,θ).* Then

$$\lim_{n\to\infty} \sup_{z_i,\theta} \left\| E\left[c(z,z_i,\theta)\,\frac{1}{a_n^{m+1}}\,\frac{\partial K\left(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)}{\partial v}\middle| z_i\right] - \frac{\partial}{\partial t}\left[E(c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta)g(t(z_i,\theta)|\theta)\right]\right\| = 0.$$

Furthermore, if $K(v)$ has zero moments up to the order s^*, $E(c(z,z_i,\theta)|t,z_i,\theta)g(t|\theta)$ is differentiable at t everywhere to the order $s^* + 1$, and these derivatives are uniformly bounded, then

$$\sup_{z_i,\theta}\left\| E\left[c(z,z_i,\theta)\,\frac{1}{a_n^{m+1}}\,\frac{\partial K\left(\dfrac{t(z_i,\theta)-t(z,\theta)}{a_n}\right)}{\partial v}\right| z_i\right] - \frac{\partial}{\partial t}\left[E(c(z,z_i,\theta)|t(z_i,\theta),z_i,\theta)g(t(z_i,\theta)|\theta)\right]\right\| = O(a_n^{s^*}).$$

Denote $A(v|x,\theta) = E(v|x\alpha,\theta)p(x\alpha|\theta)$, *and*

$C(z,\theta) = E(I_1|\epsilon(\theta),x\alpha,\theta)g(\epsilon(\theta),x\alpha|\theta) = f(\epsilon(\theta)|I_1 = 1,x\alpha,\theta)E(I_1|x\alpha,\theta)p(x\alpha|\theta)$.

PROPOSITION 3. *Under Assumptions* 1–4, *we have*

$$\sup_{S_x\times\Theta} |A_n(v|x_i,\theta) - A(v|x_i,\theta)| = O_P[(na_{1,n}^m/\ln^2 n)^{-1/2}] + O(a_{1,n}^{s_1^*}) \qquad \textbf{(A.1)}$$

and

$$\sup_{S_z\times\Theta} |C_n(z_i,\theta) - C(z_i,\theta)| = O_P[(na_{2,n}^{m+k}/\ln^2 n)^{-1/2}] + O(a_{2,n}^{s_2^*}). \qquad \textbf{(A.2)}$$

Proof. Define a Euclidean class of functions as in Pakes and Pollard (1989). By Example 2.10 of Pakes and Pollard, the class of functions of (v,x) indexed by (t,a,α) of the form $vK(t - x\alpha/a)$ is uniformly bounded and Euclidean for $v = I_1 \dots, I_L$ or 1, by Assumption 3. The class of functions of (z,I_1) indexed by $(s,t,a_1,a_2,\alpha,\gamma)$ of the form $I_1 J((s - (y - x\gamma))/a_1,(t - x\alpha)/a_2)$ is also uniformly bounded and Euclidean. In addition, $E[vK(t - x\alpha/a)]^2 \le Ma^m$ *and*

$$E\left[I_1 J\left(\frac{s-(y-x\gamma)}{a_2},\frac{t-x\alpha}{a_2}\right)\right]^2 \le Ma_2^{m+k}$$

for some finite M. By Theorem 2.37 of Pollard (1984)

$$\sup_{S_x\times\Theta} |A_n(v|x_i,\theta) - E[A_n(v|x_i,\theta)|x_i]| = O_P[(na_{1,n}^m/\ln^2 n)^{-1/2}] \qquad \textbf{(A.3)}$$

and

$$\sup_{S_z\times\Theta} |C_n(z_i,\theta) - E[C_n(z_i,\theta)|z_i]| = O_P[(na_{2,n}^{m+k}/\ln^2 n)^{-1/2}]. \qquad \textbf{(A.4)}$$

Actually the rate in (A.3) can be improved to $O_P[(na_{1,n}^m/\ln n)^{-1/2}]$ as a result of the compact nature of $S_x \times \Theta$ by using the results in Ichimura and Lee (1991).

As $K(\cdot)$ has zero moments up to order s_1^* and $J(\cdot)$ has zero moments up to order s_2^*, Proposition 1 guarantees that $\sup_{S_x\times\Theta}|E[A_n(v|x_i,\theta)|x_i] - A(v|x_i,\theta)| = O(a_{1,n}^{s_1^*})$ and $\sup_{z_i}|E[C_n(z_i,\theta)|z_i] - C(z_i,\theta)| = O(a_{2,n}^{s_2^*})$. In summary, $\sup_{S_x\times\Theta}|A_n(v|x_i,\theta) - A(v|x_i,\theta)| = O_P[(na_{1,n}^m/\ln^2 n)^{-1/2}] + O(a_{1,n}^{s_1^*})$ and $\sup_{S_z\times\Theta}|C_n(z_i,\theta) - C(z_i,\theta)| = O_P[(na_{2,n}^{m+k}/\ln^2 n)^{-1/2}] + O(a_{2,n}^{s_2^*})$. ■

Define $\rho_\theta(x,\theta) = (\partial x\gamma/\partial\theta, -\partial x\alpha/\partial\theta)$, $\eta_\theta(x,\theta) = \partial x\alpha/\partial\theta$, $r_\theta(x_i,x_j,\theta) = \rho_\theta(x_j,\theta) - \rho_\theta(x_i,\theta)$, and $s_\theta(x_i,x_j,\theta) = \eta_\theta(x_i,\theta) - \eta_\theta(x_j,\theta)$. The first-order derivatives of $A_n(I_l|x_i,\theta)$ and $C_n(z_i,\theta)$ with respect to θ are

$$\frac{\partial A_n(v|x_i,\theta)}{\partial\theta} = \frac{1}{n-1}\sum_{j\neq i}^{n} v_j s_\theta(x_i,x_j,\theta)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)$$

and

$$\frac{\partial C_n(z_i,\theta)}{\partial\theta} = \frac{1}{n-1}\sum_{j\neq i}^{n} I_{1j}\, r_\theta(x_i,x_j,\theta)\,\frac{1}{a_{2,n}^{m+k+1}}\,\nabla J\left(\frac{\epsilon_i(\theta)-\epsilon_j(\theta)}{a_{2,n}},\frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right),$$

where $\nabla K(w) = (\partial/\partial w')K(w)$ and $\nabla J(v) = (\partial/\partial v')J(v)$ denote gradient (column) vectors of the kernel functions. Similarly to the arguments leading to (A.1) and (A.2), Theorem 2.37 of Pollard (1984) and Proposition 2 imply that

$$\sup_{S_x\times\Theta}\left|\frac{\partial A_n(v|x_i,\theta)}{\partial\theta} - \frac{\partial A(v|x_i,\theta)}{\partial\theta}\right| = O_P[(na_{1,n}^{m+2}/\ln^2 n)^{-1/2}] + O(a_{1,n}^{s_1^*}).$$

For any component θ_k of θ, by Lemma 2 in Appendix D,

$$\begin{aligned}\frac{\partial A(v|x_i,\theta)}{\partial\theta_k} &= \operatorname{tr}\nabla[E(v_i|x_i\alpha,\theta)\eta_{\theta_k}(x_i,\theta) - E(v_i\eta_{\theta_k}(x_i,\theta)|x_i\alpha,\theta)]p(x_i\alpha|\theta)\\ &\quad + [E(v_i|x_i\alpha,\theta)\eta_{\theta_k}(x_i,\theta) - E(v_i\eta_{\theta_k}(x_i,\theta)|x_i\alpha,\theta)]\nabla p(x_i\alpha|\theta),\end{aligned}$$

where $\nabla E(\cdot|t,\theta) = (\partial/\partial t')E(\cdot|t,\theta)$ and $\nabla p(t,\theta) = (\partial/\partial t')p(t|\theta)$. Also,

$$\sup_{S_z\times\Theta}\left|\frac{\partial C_n(z_i,\theta)}{\partial\theta} - \frac{\partial C(z_i,\theta)}{\partial\theta}\right| = O_P[(na_{2,n}^{m+k+2}/\ln^2 n)^{-1/2}] + O(a_{2,n}^{s_2^*}),$$

where

$$\begin{aligned}\frac{\partial C(z_i,\theta)}{\partial\theta_k} &= \operatorname{tr}\nabla[E(I_{1i}\rho_{\theta_k}(x_i,\theta)|\epsilon_i(\theta),x_i\alpha,\theta) - E(I_{1i}|\epsilon_i(\theta),x_i\alpha,\theta)\rho_{\theta_k}(x_i,\theta)]\\ &\quad\times g(\epsilon_i(\theta),x_i\alpha|\theta)\\ &\quad + [E(I_{1i}\rho_{\theta_k}(x_i,\theta)|\epsilon_i(\theta),x_i\alpha,\theta) - E(I_{1i}|\epsilon_i(\theta),x_i\alpha,\theta)\rho_{\theta_k}(x_i,\theta)]\\ &\quad\times\nabla g(\epsilon_i(\theta),x_i\alpha|\theta),\end{aligned}$$

by Lemma 2 in Appendix D, where $\nabla g(w|\theta) = (\partial/\partial w')g(w|\theta)$.

Define $t_d^*(c) = 1$ for $\frac{1}{2}\leq c$ and $= 0$ otherwise, $t_{nd}(z_i,\theta) = t_d^*(C_n(z_i,\theta)/\delta_{2,n})\prod_{l=0}^{L} t_d^*(A_n(I_l|x_i,\theta)/\delta_{1,n})$, and $t_{nd}^{\infty}(z_i,\theta) = t_d^*(C(z_i,\theta)/\delta_{2,n})\prod_{l=0}^{L} t_d^*(A(I_l|x_i,\theta)/\delta_{1,n})$. These trimming functions are introduced to simplify notations in subsequent proofs. The following lemma is useful in controlling denominators with small value by trimming functions in nonparametric estimation.

LEMMA 1. *Under Assumptions* 1–3,

(1)

$$\frac{t_n(z_i,\theta)}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)} \leq \frac{M}{\delta_{2,n}^{(m_1+m_2)}} \qquad \textbf{(A.5)}$$

and

$$\frac{t_n(z_i,\theta)}{A_n^{m_1}(z_i,\theta)A^{m_2}(z_i,\theta)} \le \frac{M}{\delta_{1,n}^{(m_1+m_2)}} \quad \textbf{(A.6)}$$

in probability uniformly in (z_i,θ), *where* M *is a generic constant, and* $m_1, m_2 \ge 0$.

(2)

$$\frac{\partial t_n(z_i,\theta)}{\partial\theta'}\frac{1}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)} \le \frac{M}{\delta_{2,n}^{(m_1+m_2+1)}} \quad \textbf{(A.7)}$$

and

$$\frac{\partial t_n(z_i,\theta)}{\partial\theta'}\frac{1}{A_n^{m_1}(z_i,\theta)A^{m_2}(z_i,\theta)} \le \frac{M}{\delta_{1,n}^{(m_1+m_2+1)}} \quad \textbf{(A.8)}$$

in probability uniformly in (z_i,θ), *for* $m_1, m_2 \ge 0$.

(3) *The inequalities in* (A.5) *and* (A.6) *still hold by replacing* $t_n(z_i,\theta)$ *by* $t_{nd}(z_i,\theta)$, $t_n^\infty(z_i,\theta)$, *or* $t_{nd}^\infty(z_i,\theta)$.

Proof. We will prove (A.5) only. (A.6)–(A.8) can be proved similarly. Now we rewrite $(t_n(z_i,\theta))/(C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta))$ as

$$\frac{t_n(z_i,\theta)}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)} = \frac{t_n(z_i,\theta)}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)}\big[1_{\{|C_n(z_i,\theta)-C(z_i,\theta)|\ge\delta_{2,n}/4,\, C_n(z_i,\theta)>\delta_{2,n}/2\}} + 1_{\{|C_n(z_i,\theta)-C(z_i,\theta)|\le\delta_{2,n}/4,\, C_n(z_i,\theta)\ge\delta_{2,n}/2\}}\big].$$

First, note that

$$\frac{t_n(z_i,\theta)}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)}1_{\{|C_n(z_i,\theta)-C(z_i,\theta)|\le\delta_{2,n}/4,\, C_n(z_i,\theta)>\delta_{2,n}/2\}} \le \frac{M}{\delta_{2,n}^{(m_1+m_2)}}$$

uniformly in (z_i,θ). Second,

$$P\left(\frac{t_n(z_i,\theta)}{C_n^{m_1}(z_i,\theta)C^{m_2}(z_i,\theta)}1_{\{|C_n(z_i,\theta)-C(z_i,\theta)|\ge\delta_{2,n}/4,\, C_n(z_i,\theta)>\delta_{2,n}/2\}} \ge \frac{M}{\delta_{2,n}^{(m_1+m_2)}}, \quad \text{for some } z_i,\theta\right) \le P\left(\sup_{z_i,\theta}|C_n(z_i,\theta)-C(z_i,\theta)| \ge \frac{\delta_{2,n}}{4}\right),$$

where the term on the right converges to zero. Therefore (A.5) holds. ■

APPENDIX B: ASYMPTOTIC DISTRIBUTION

In this appendix, we will derive the asymptotic distribution of the TSC estimator. The derivations are divided into several steps.

B.1: $\sqrt{n}(S_n(\theta_o) - S_n^*(\theta_o)) = o_p(1)$ **and** $\sqrt{n}S_n^*(\theta_o) \xrightarrow{D} N(0, I(\theta_o))$. In this step, we investigate the limiting property of $\sqrt{n}(S_n(\theta_o) - S_n^*(\theta_o))$. We would like to show that this difference converges in probability to zero and, hence, $\sqrt{n}S_n(\theta_o)$ has the same limiting distribution as $\sqrt{n}S_n^*(\theta_o)$.

To simplify notation, define $\Delta A_n(I_l|x_i,\theta) = A_n(I_l|x_i,\theta) - A(I_l|x_i,\theta)$, $\Delta A_{n,\theta}(I_l|x_i,\theta) = (\partial A_n(I_l|x_i,\theta))/\partial\theta - (\partial A(I_l|x_i,\theta))/\partial\theta$ for $l = 0,\ldots,L$, $\Delta C_n(z_i,\theta) = C_n(z_i,\theta) - C(z_i,\theta)$, and $\Delta C_{n,\theta}(z_i,\theta) = ((\partial C_n(z_i,\theta))/\partial\theta) - ((\partial C(z_i,\theta))/\partial\theta)$, where $I_0 = 1$. By an expansion,

$$\begin{aligned}
&\frac{1}{C_n(z_i,\theta)}\frac{\partial C_n(z_i,\theta)}{\partial\theta} - \frac{1}{C(z_i,\theta)}\frac{\partial C(z_i,\theta)}{\partial\theta} \\
&\quad = \frac{1}{C(z_i,\theta)}\left(\Delta C_{n,\theta}(z_i,\theta) - \Delta C_n(z_i,\theta)\frac{\partial \ln C(z_i,\theta)}{\partial\theta}\right) \\
&\qquad + \frac{1}{C_n(z_i,\theta)C(z_i,\theta)}\left([\Delta C_n(z_i,\theta)]^2\frac{\partial \ln C(z_i,\theta)}{\partial\theta} - \Delta C_n(z_i,\theta)\Delta C_{n,\theta}(z_i,\theta)\right),
\end{aligned} \tag{B.1}$$

and a similar expansion is valid for the $(1/A_n(I_l|x_i,\theta))(\partial A_n(I_l|x_i,\theta)/\partial\theta)$. Define

$$\begin{aligned}
\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta} &= I_{1i}\left[\frac{\partial \ln C(z_i,\theta)}{\partial\theta} - \frac{\partial \ln A(I_1|x_i,\theta)}{\partial\theta}\right] \\
&\quad + \sum_{l=1}^{L} I_{li}\left[\frac{\partial \ln A(I_l|x_i,\theta)}{\partial\theta} - \frac{\partial \ln A(1|x_i,\theta)}{\partial\theta}\right] \\
&= I_{1i}\frac{\partial \ln f(\epsilon_i(\theta)|I_{1i} = 1, x_i\alpha, \theta)}{\partial\theta} + \sum_{l=1}^{L} I_{li}\frac{\partial \ln P_l(x_i,\theta)}{\partial\theta}.
\end{aligned} \tag{B.2}$$

It follows that $\sqrt{n}S_n(\theta_o)$ can be rewritten as

$$\begin{aligned}
\sqrt{n}S_n(\theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t_n(z_i,\theta_o)\Bigg\{&\frac{\partial \ln l(I_i,z_i,\theta_o)}{\partial\theta} + I_{1i}[L_{c,n}(z_i) - L_{1,n}(x_i)] \\
&+ \sum_{l=1}^{L} I_{li}[L_{l,n}(x_i) - L_{o,n}(x_i)] \\
&+ I_{1i}[Q_{c,n}(z_i) - Q_{1,n}(x_i)] \\
&+ \sum_{l=1}^{L} I_{li}[Q_{l,n}(x_i) - Q_{o,n}(x_i)]\Bigg\},
\end{aligned}$$

where

$$L_{c,n}(z_i) = \frac{\Delta C_{n,\theta}(z_i,\theta_o)}{C(z_i,\theta_o)} - \frac{\Delta C_n(z_i,\theta_o)}{C(z_i,\theta_o)} \frac{\partial \ln C(z_i,\theta_o)}{\partial \theta},$$

$$L_{l,n}(x_i) = \frac{\Delta A_{n,\theta}(I_l|x_i,\theta_o)}{A(I_l|x_i,\theta_o)} - \frac{\Delta A_n(I_l|x_i,\theta_o)}{A(I_l|x_i,\theta_o)} \frac{\partial \ln A(I_l|x_i,\theta_o)}{\partial \theta},$$

$$Q_{c,n}(z_i) = \frac{[\Delta C_n(z_i,\theta_o)]^2}{C_n(z_i,\theta_o)C(z_i,\theta_o)} \frac{\partial \ln C(z_i,\theta_o)}{\partial \theta} - \frac{\Delta C_n(z_i,\theta_o)\Delta C_{n,\theta}(z_i,\theta_o)}{C_n(z_i,\theta_o)C(z_i,\theta_o)},$$

and

$$Q_{l,n}(x_i) = \frac{[\Delta A_n(I_l|x_i,\theta_o)]^2}{A_n(I_l|x_i,\theta_o)A(I_l|x_i,\theta_o)} \frac{\partial \ln A(I_l|x_i,\theta_o)}{\partial \theta} - \frac{\Delta A_n(I_l|x_i,\theta_o)\Delta A_{n,\theta}(I_l|x_i,\theta_o)}{A_n(I_l|x_i,\theta_o)A(I_l|x_i,\theta_o)}.$$

The trimming function $t_n(z_i,\theta_o)$ depends on nonparametric estimates. The difference $t_n(z_i,\theta_o) - t_n^\infty(z_i,\theta_o)$ can be analyzed by a Taylor series expansion. Define $d_{c,n}(z_i) = \prod_{l=0}^{L} t^*(A(I_{li}|x_i,\theta)/\delta_{1,n})\nabla t^*(C(z_i,\theta)/\delta_{2,n})$ and $d_{j,n}(z_i) = t^*(C(z_i,\theta)/\delta_{2,n})\prod_{l=0,l\neq j}^{L} t^* (A(I_{li}|x_i,\theta)/\delta_{1,n})\nabla t^*(A(I_{ji}|x_i,\theta)/\delta_{1,n})$. As the function t^* is twice continuously differentiable in its arguments and these derivatives are bounded, $t_n(z_i,\theta_o) - t_n^\infty(z_i,\theta_o)$ satisfies the Lipschitz condition of order 1:

$$|t_n(z_i,\theta_o) - t_n^\infty(z_i,\theta_o)| \\ \leq c_1 \left\{ \|\delta_{2,n}^{-1}\Delta C_n(z_i,\theta_o)\| + \sum_{l=0}^{L} \|\delta_{1,n}^{-1}\Delta A_n(I_l|x_i,\theta_o)\| \right\} (t_{nd}(z_i,\theta_o) + t_{nd}^\infty(z_i,\theta_o)) \quad \textbf{(B.3)}$$

for some constant c_1. Also,

$$\left| t_n(z_i,\theta_o) - t_n^\infty(z_i,\theta_o) - d_{c,n}(z_i)\delta_{2,n}^{-1}\Delta C_n(z_i,\theta_o) - \sum_{l=0}^{L} d_{l,n}(z_i)\delta_{1,n}^{-1}\Delta A_n(I_l|x_i,\theta_o) \right| \\ \leq c_2 \left\{ \|\delta_{2,n}^{-1}\Delta C_n(z_i,\theta_o)\|^2 + \sum_{l=0}^{L} \|\delta_{1,n}^{-1}\Delta A_n(I_l|x_i,\theta_o)\|^2 \right\}$$

for some constant c_2. With these expansions, it follows that

$$\sqrt{n}S_n(\theta_o) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n^\infty(z_i,\theta_o) \frac{\partial \ln l(I_i,z_i,\theta_o)}{\partial \theta} + L_n^{(1)} + L_n^{(2)} + R_n, \quad \textbf{(B.4)}$$

where

$$L_n^{(1)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ d_{c,n}(z_i)\delta_{2,n}^{-1}\Delta C_n(z_i,\theta_o) + \sum_{l=0}^{L} d_{l,n}(z_i)\delta_{1n}^{-1}\Delta A_n(I_l|x_i,\theta_o) \right\} \frac{\partial \ln l(I_i,z_i,\theta_o)}{\partial \theta},$$

$$L_n^{(2)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n^\infty(z_i,\theta_o) \left\{ I_{1i}[L_{c,n}(z_i) - L_{1,n}(x_i)] + \sum_{l=1}^{L} I_{li}[L_{l,n}(x_i) - L_{0,n}(x_i)] \right\},$$

and

$$|R_n| \le \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(c_2 \left\{ \|\delta_{2,n}^{-1} \Delta C_n(z_i,\theta_o)\|^2 + \sum_{l=0}^{L} \|\delta_{1,n}^{-1} \Delta A(I_l|x_i,\theta_o)\|^2 \right\} \right) \left\| \frac{\partial \ln l(I_i,z_i,\theta_o)}{\partial\theta} \right\|$$
$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(c_1 \left\{ \|\delta_{2,n}^{-1} \Delta C_n(z_i,\theta_o)\| + \sum_{l=0}^{L} \|\delta_{1,n}^{-1} \Delta A(I_l|x_i,\theta_o)\| \right\} \right)$$
$$\times \left\{ \|L_{c,n}(z_i) - L_{1,n}(z_i)\| + \sum_{l=1}^{L} \|L_{l,n}(x_i) - L_{0,n}(x_i)\| \right\} (t_{nd}(z_i,\theta_o) + t_{nd}^{\infty}(z_i,\theta_o))$$
$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n(z_i,\theta_o) \left\{ \|Q_{c,n}(z_i) - Q_{1,n}(x_i)\| + \sum_{l=1}^{L} \|Q_{l,n}(x_i) - Q_{0,n}(x_i)\| \right\}$$
$$= R_{n1} + R_{n2} + R_{n3} \quad \text{(say)}.$$

B.1.1: Show $R_n = o_p(1)$. We first consider R_{n1}. By (A.1), (A.2), and Assumptions 1 and 3,

$$R_{n1} \le n^{1/2}\{[O_P[(na_{1,n}^{m}/\ln^2 n)^{-1}] + O(a_{1,n}^{2s_1^*})]\delta_{1,n}^{-2} + [O_P[(na_{2,n}^{m+k}/\ln^2 n)^{-1}]$$
$$+ O(a_{2,n}^{2s_2^*})]\delta_{2,n}^{-2}\} = o_p(1).$$

Similarly, we can show that $R_{n2} = o_p(1)$ and $R_{n3} = o_p(1)$.

B.1.2: Show $L_n^{(2)} = o_p(1)$. Denote $L_{n,1}^{(2)} = 1/\sqrt{n}\sum_{i=1}^{n} t_n^{\infty}(z_i,\theta_o) I_{1i}[L_{c,n}(z_i) - L_{1,n}(x_i)]$ and

$$L_{n,2}^{(2)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n^{\infty}(z_i,\theta_o) \sum_{l=1}^{L} I_{li}[L_{l,n}(x_i) - L_{0,n}(x_i)].$$

It follows that $L_n^{(2)} = L_{n,1}^{(2)} + L_{n,2}^{(2)}$. The $L_{n,1}^{(2)}$ and $L_{n,2}^{(2)}$ can be simplified to

$$L_{n,1}^{(2)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n^{\infty}(z_i,\theta_o) I_{1i} \left\{ \frac{1}{C(z_i,\theta_o)} \left(\frac{\partial C_n(z_i,\theta_o)}{\partial\theta} - \frac{\partial \ln C(z_i,\theta_o)}{\partial\theta} C_n(z_i,\theta_o) \right) \right.$$
$$- \frac{1}{A(I_1|x_i,\theta_o)}$$
$$\left. \times \left(\frac{\partial A_n(I_1|x_i,\theta_o)}{\partial\theta} - \frac{\partial \ln A(I_1|x_i,\theta_o)}{\partial\theta} A_n(I_1|x_i,\theta_o) \right) \right\}$$

and

$$L_{n,2}^{(2)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_n^{\infty}(z_i,\theta_o) \sum_{l=1}^{L} I_{li}$$
$$\times \left\{ \frac{1}{A(I_l|x_i,\theta_o)} \left(\frac{\partial A_n(I_l|x_i,\theta_o)}{\partial\theta} - \frac{\partial \ln A(I_l|x_i,\theta_o)}{\partial\theta} A_n(I_l|x_i,\theta_o) \right) \right.$$
$$\left. - \frac{1}{A(1|x_i,\theta_0)} \left(\frac{\partial A_n(1|x_i,\theta_o)}{\partial\theta} - \frac{\partial \ln A(1|x_i,\theta_o)}{\partial\theta} A_n(1|x_i,\theta_o) \right) \right\}.$$

Let

$$
\begin{aligned}
&\Phi_1(z_i, I_i, z_j, I_j, a_n, \theta)\\
&\quad = t_n^\infty(z_i,\theta_o) I_{1i} I_{1j}\Bigg[\frac{1}{C(z_i,\theta)}\, r_\theta(x_i,x_j,\theta)\,\frac{1}{a_{2,n}^{m+k+1}}\,\nabla J\left(\frac{\epsilon_i(\theta)-\epsilon_j(\theta)}{a_{2,n}}, \frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right)\\
&\qquad - \frac{\partial \ln C(z_i,\theta)}{\partial\theta}\,\frac{1}{C(z_i,\theta)}\,\frac{1}{a_{2,n}^{m+k}}\, J\left(\frac{\epsilon_i(\theta)-\epsilon_j(\theta)}{a_{2,n}}, \frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right)\\
&\qquad - \frac{1}{A(I_1|x_i,\theta)}\, s_\theta(x_i,x_j,\theta)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\\
&\qquad + \frac{\partial \ln A(I_1|x_i,\theta)}{\partial\theta}\,\frac{1}{A(I_1|x_i,\theta)}\,\frac{1}{a_{1,n}^{m}}\, K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\Bigg].
\end{aligned}
$$

It follows that $L_{n,1}^{(2)} = 1/(\sqrt{n}(n-1))\sum_{i=1}^n\sum_{j\neq i}^n \Phi_1(z_i, I_i, z_j, I_j, a_n, \theta_o)$, which can be analyzed as a U-statistic. Similarly, define

$$
\begin{aligned}
&\Phi_2(x_i, I_i, x_j, I_j, a_n, \theta)\\
&\quad = t_n^\infty(z_i,\theta_o)\sum_{l=1}^{L} I_{li}\Bigg[I_{lj}\,\frac{1}{A(I_l|x_i,\theta)}\, s_\theta(x_i,x_j,\theta)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\\
&\qquad - I_{lj}\,\frac{\partial \ln A(I_l|x_i,\theta)}{\partial\theta}\,\frac{1}{A(I_l|x_i,\theta)}\,\frac{1}{a_{1,n}^{m}}\, K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\\
&\qquad - \frac{1}{A(1|x_i,\theta)}\, s_\theta(x_i,x_j,\theta)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\\
&\qquad + \frac{\partial \ln A(1|x_i,\theta)}{\partial\theta}\,\frac{1}{A(1|x_i,\theta)}\,\frac{1}{a_{1,n}^{m}}\, K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\Bigg].
\end{aligned}
$$

It follows that $L_{n,2}^{(2)} = 1(\sqrt{n}(n-1))\sum_{i=1}^n\sum_{j\neq i}^n \Phi_2(x_i, I_i, x_j, I_j, a_n, \theta_o)$.

The next step is to show that both $L_{n,1}^{(2)}$ and $L_{n,2}^{(2)}$ converge in probability to zero. The following result on the U-statistic will be useful for this purpose.

PROPOSITION 4. *Let $z_1,\ldots,z_n$ be i.i.d. observations and $\Psi_n(z_1,z_2,a_n)$ be a sequence of measurable functions. Suppose that the following conditions hold.*

(1) *There exist Lebesgue square integrable functions $h_j(z)$, $j = 1,2$ such that*

$$|E(\Psi_n(z_1,z_2,a_n)|z_1)| \le h_1(z_1), \qquad |E(\Psi_n(z_1,z_2,a_n)|z_2)| \le h_2(z_2);$$

(2) $E(\Psi_n(z_1,z_2,a_n)) = O(a_n^s)$ *and* $\operatorname{var}(\Psi_n(z_1,z_2,a_n)) = O(1/a_n^r)$; *and*
(3) $\lim_{n\to\infty} E(\Psi_n(z_1,z_2,a_n)|z_j) = 0$, *almost everywhere* (*a.e.*), $j = 1,2$.

If $\lim_{n\to\infty}\sqrt{n}a_n^s = 0$ *and* $\lim_{n\to\infty} na_n^r = \infty$, *then*

$$\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Psi_n(z_i,z_j,a_n) \xrightarrow{p} 0.$$

Proof. See Ichimura and Lee (1991).

As the kernel functions $K(\cdot)$ and $J(\cdot)$ are, respectively, of orders s_1^* and s_2^*, Proposition 1, Proposition 2, and Lemma 1 imply that

$$|E(\Phi_1(z_i,I_i,z_j,I_j,a_n)|z_i,I_i)| = O(\max\{a_{1,n}^{s_1^*}\delta_{1,n}^{-1}, a_{2,n}^{s_2^*}\delta_{2,n}^{-1}\})\left\{\left\|\frac{\partial \ln C(z_i,\theta)}{\partial\theta}\right\|^2 + \left\|\frac{\partial \ln A(I_l|x_i,\theta)}{\partial\theta}\right\|^2\right\}$$

and

$$|E(\Phi_2(x_i,I_i,x_j,I_j,a_n)|x_i,I_i)| = O(a_{1,n}^{s_1^*}\delta_{1,n}^{-1})\left\{\left\|\frac{\partial \ln A(I_l|x_i,\theta)}{\partial\theta}\right\|^2\right\}$$

uniformly in z_i, x_i, and I_i, which in turn imply that $E(\Phi_1(z_i,I_i,z_j,I_j,a_n)) = O(\max\{a_{1,n}^{s_1^*}\delta_{1,n}^{-1}, a_{2,n}^{s_2^*}\delta_{2,n}^{-1}\})$ and $E(\Phi_2(x_i,I_i,x_j,I_j,a_n)) = O(a_{1,n}^{s_1^*}\delta_{1,n}^{-1})$. It is straightforward to verify that $\mathrm{var}(\Phi_1(z_i,I_i,z_j,I_j,a_n)) = O(\max\{1/(a_{1,n}^{m+2}\delta_{1,n}^2), 1/(a_{2,n}^{m+k+2}\delta_{2,n}^2)\})$ and $\mathrm{var}(\Phi_2(x_i,I_i,x_j,I_j,a_n)) = O(1/a_{1,n}^{m+2}\delta_{1,n}^2)$. Because

$$\begin{aligned}
&E(\Phi_1(z_i,I_i,z_j,I_j,a_n)|z_i,I_i)\\
&\quad = t_n^\infty(z_i,\theta_o)I_{1i}\\
&\qquad \times \left\{\frac{1}{C(z_i,\theta_o)}E\left[I_{1j}r_\theta(x_i,x_j,\theta_o)\frac{1}{a_{2,n}^{m+k+1}}\nabla J\left(\frac{\epsilon_i(\theta_o)-\epsilon_j(\theta_o)}{a_{2,n}}, \frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right)\middle| z_i\right]\right.\\
&\qquad - \frac{\partial \ln C(z_i,\theta_o)}{\partial\theta}\frac{1}{C(z_i,\theta_o)}E\left[I_{1j}\frac{1}{a_{2,n}^{m+k}}J\left(\frac{\epsilon_i(\theta_o)-\epsilon_j(\theta_o)}{a_{2,n}}, \frac{x_i\alpha - x_j\alpha}{a_{2,n}}\right)\middle| z_i\right]\\
&\qquad - \frac{1}{A(I_1|x_i,\theta_o)}E\left[I_{1j}s_\theta(x_i,x_j,\theta_o)\frac{1}{a_{1,n}^{m+1}}\nabla K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\middle| x_i\right]\\
&\qquad \left. + \frac{\partial \ln A(I_1|x_i,\theta_o)}{\partial\theta}\frac{1}{A(I_1|x_i,\theta_o)}E\left[I_{1j}\frac{1}{a_{1,n}^{m}}K\left(\frac{x_i\alpha - x_j\alpha}{a_{1,n}}\right)\middle| x_i\right]\right\}
\end{aligned}$$

and $\lim_{n\to\infty} t_n^\infty(z_i,\theta_o) = 1$, it follows that

$$\begin{aligned}
&\lim_{n\to\infty} E(\Phi_1(z_i,I_i,z_j,I_j,a_n)|z_i,I_i)\\
&\quad = I_{1i}\left[\frac{1}{C(z_i,\theta_o)}\frac{\partial C(z_i,\theta_o)}{\partial\theta} - \left(\frac{1}{C(z_i,\theta_o)}\right)^2\frac{\partial C(z_i,\theta_o)}{\partial\theta}C(z_i,\theta_o)\right.\\
&\qquad \left. - \frac{1}{A(I_1|x_i,\theta_o)}\frac{\partial A(I_1|x_i,\theta_o)}{\partial\theta} + \left(\frac{1}{A(I_1|x_i,\theta_o)}\right)^2\frac{\partial A(I_1|x_i,\theta_o)}{\partial\theta}A(I_1|x_i,\theta_o)\right]\\
&\quad = 0.
\end{aligned}$$

Similarly,

$$
\begin{aligned}
&E(\Phi_2(x_i,I_i,x_j,I_j,a_n,\theta_o)|x_i,I_i)\\
&\quad=\sum_{l=1}^{L} I_{li}\left\{\frac{1}{A(I_l|x_i,\theta_o)}\,E\left[I_{lj}\,s_\theta(x_i,x_j,\theta_o)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha_o-x_j\alpha_o}{a_{1,n}}\right)\middle|\,x_i\right]\right.\\
&\qquad\qquad-\frac{\partial\ln A(I_l|x_i,\theta_o)}{\partial\theta}\,\frac{1}{A(I_l|x_i,\theta_o)}\,E\left[I_{lj}\,\frac{1}{a_{1,n}^{m}}\,K\left(\frac{x_i\alpha_o-x_j\alpha_o}{a_{1,n}}\right)\middle|\,x_i\right]\\
&\qquad\qquad-\frac{1}{A(1|x_i,\theta_o)}\,E\left[s_\theta(x_i,x_j,\theta_o)\,\frac{1}{a_{1,n}^{m+1}}\,\nabla K\left(\frac{x_i\alpha_o-x_j\alpha_o}{a_{1,n}}\right)\middle|\,x_i\right]\\
&\qquad\qquad\left.+\frac{\partial\ln A(1|x_i,\theta_o)}{\partial\theta}\,\frac{1}{A(1|x_i,\theta_o)}\,E\left[\frac{1}{a_{1,n}^{m}}\,K\left(\frac{x_i\alpha_o-x_j\alpha_o}{a_{1,n}}\right)\middle|\,x_i\right]\right\},
\end{aligned}
$$

and hence

$$
\begin{aligned}
&\lim_{n\to\infty} E(\Phi_2(x_i,I_i,x_j,I_j,a_n)|x_i,I_i)\\
&\quad=\sum_{l=1}^{L} I_{li}\left\{\frac{1}{A(I_l|x_i,\theta_o)}\,\frac{\partial A(I_l|x_i,\theta_o)}{\partial\theta}-\frac{\partial\ln A(I_l|x_i,\theta_o)}{\partial\theta}\,\frac{1}{A(I_l|x_i,\theta_o)}\,A(I_l|x_i,\theta_o)\right.\\
&\qquad\qquad\left.-\frac{1}{A(1|x_i,\theta_o)}\,\frac{\partial A(1|x_i,\theta_o)}{\partial\theta}+\frac{\partial\ln A(1|x_i,\theta_o)}{\partial\theta}\,\frac{1}{A(1|x_i,\theta_o)}\,A(1|x_i,\theta_o)\right\}\\
&\quad=0.
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
E(\Phi_1(z_j,I_j,z_i,I_i,a_n)|z_i,I_i)&=I_{1i}\{E[T_1(z_j,I_j,z_i,a_n)|z_i]-E[T_2(z_j,I_j,z_i,a_n)|z_i]\\
&\qquad-E[T_3(z_j,I_j,z_i,a_n)|z_i]+E[T_4(z_j,I_j,z_i,a_n)|z_i]\},
\end{aligned}
$$

where

$$
\begin{aligned}
T_1(z_j,I_j,z_i,a_n)&=t_n^\infty(z_j,\theta_o)E[I_{1j}\,r_\theta(x_j,x_i,\theta_o)|\epsilon_j(\theta_o),x_j\alpha_o,z_i]\\
&\qquad\times\frac{1}{a_n^{m+k+1}}\,\nabla J\left(\frac{\epsilon_j(\theta_o)-\epsilon_i(\theta_o)}{a_n},\frac{x_j\alpha_o-x_i\alpha_o}{a_n}\right)\frac{1}{C(z_j,\theta_o)},\\
T_2(z_j,I_j,z_i,a_n)&=t_n^\infty(z_j,\theta_o)E\left[I_{1j}\,\frac{\partial C(z_j,\theta_o)}{\partial\theta}\middle|\,\epsilon_j(\theta_o),x_j\alpha_o\right]\\
&\qquad\times\frac{1}{a_n^{m+k}}\,J\left(\frac{\epsilon_j(\theta)-\epsilon_i(\theta)}{a_n},\frac{x_j\alpha_o-x_i\alpha_o}{a_n}\right)\left(\frac{1}{C(z_j,\theta_o)}\right)^2,\\
T_3(z_j,I_j,z_i,a_n)&=t_n^\infty(z_j,\theta_o)E[I_{1j}\,s_\theta(x_j,x_i,\theta_o)|x_j\alpha_o,z_i]\\
&\qquad\times\frac{1}{a_n^{m+1}}\,\nabla K\left(\frac{x_j\alpha_o-x_i\alpha_o}{a_n}\right)\frac{1}{A_1(x_j,\theta_o)},
\end{aligned}
$$

and

$$T_4(z_j,I_j,z_i,a_n) = t_n^\infty(z_j,\theta_o)E\left[I_{1j}\,\frac{\partial A_1(x_j,\theta_o)}{\partial\theta}\bigg|x_j\alpha_o\right]\frac{1}{a_n^m}K\left(\frac{x_j\alpha_o - x_i\alpha_o}{a_n}\right)\left(\frac{1}{A_1(x_j,\theta_o)}\right)^2.$$

As shown in Appendix C, for each component k,

$$\lim_{n\to\infty} E(T_{1,k}(z_j,I_j,z_i,a_n)|z_i) = -\text{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg|x_i\alpha_o\right)\quad\text{a.e.,}\tag{B.5}$$

$$\lim_{n\to\infty} E(T_{2,k}(z_j,I_j,z_i,a_n)|z_i) = -\text{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg|x_i\alpha_o\right)\quad\text{a.e.,}\tag{B.6}$$

$$\lim_{n\to\infty} E(T_{3,k}(x_j,I_j,x_i,a_n)|z_i) = -\text{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg|x_i\alpha_o\right)\quad\text{a.e.,}\tag{B.7}$$

and

$$\lim_{n\to\infty} E(T_{4,k}(x_j,I_j,x_i,a_n)|z_i) = -\text{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg|x_i\alpha_o\right)\quad\text{a.e.}\tag{B.8}$$

Hence it follows that $\lim_{n\to\infty}E(\Phi_1(z_j,I_j,z_i,I_i,a_n)|z_i,I_i)=0$ a.e. It remains to analyze the following term:

$$E(\Phi_2(z_j,I_j,z_i,I_i,a_n)|z_i,I_i) = \sum_{l=1}^{L}\{I_{li}E[W_1(z_j,I_{lj},z_i,a_n)|z_i] - I_{li}E[W_2(z_j,I_{lj},z_i,a_n)|z_i]$$
$$- E[W_3(z_j,I_{lj},z_i,a_n)|z_i] + E[W_4(z_j,I_{lj},z_i,a_n)|z_i]\},$$

where

$$W_1(z_j,I_{lj},z_i,a_n) = t_n^\infty(z_j,\theta_o)E[I_{lj}s_\theta(x_j,x_i,\theta_o)|x_j\alpha_o,x_i]$$
$$\times\frac{1}{a_n^{m+1}}\nabla K\left(\frac{x_j\alpha_o - x_i\alpha_o}{a_n}\right)\frac{1}{A(I_l|x_j,\theta_o)},$$

$$W_2(z_j,I_{lj},z_i,a_n) = t_n^\infty(z_j,\theta_o)E\left[I_{lj}\,\frac{\partial A(I_l|x_j,\theta_o)}{\partial\theta}\bigg|x_j\alpha_o\right]$$
$$\times\frac{1}{a_n^m}K\left(\frac{x_j\alpha_o - x_i\alpha_o}{a_n}\right)\left(\frac{1}{A(I_l|x_j,\theta_o)}\right)^2,$$

$$W_3(z_j,I_{lj},z_i,a_n) = t_n^\infty(z_j,\theta_o)E[I_{lj}s_\theta(x_j,x_i,\theta_o)|x_j\alpha_o,x_i]$$
$$\times\frac{1}{a_n^{m+1}}\nabla K\left(\frac{x_j\alpha_o - x_i\alpha_o}{a_n}\right)\frac{1}{A(1|x_j,\theta_o)},$$

and

$$W_4(z_j,I_{lj},z_i,a_n) = t_n^\infty(z_j,\theta_o)E\left[I_{lj}\,\frac{\partial A(1|x_j,\theta_o)}{\partial\theta}\bigg|x_j\alpha_o\right]\frac{1}{a_n^m}K\left(\frac{x_j\alpha_o - x_i\alpha_o}{a_n}\right)\frac{1}{A^2(1|x_j,\theta_o)}.$$

As shown in Appendix C, we have

$$\lim_{n\to\infty} E(W_{1,k}(z_j,I_{lj},z_i,a_n)|z_i) = -\mathrm{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg| x_i\alpha_o\right), \quad \text{a.e.,} \tag{B.9}$$

$$\lim_{n\to\infty} E(W_{2,k}(z_j,I_{lj},z_i,a_n)|z_i) = -\mathrm{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg| x_i\alpha_o\right), \quad \text{a.e.,} \tag{B.10}$$

$$\begin{aligned}\lim_{n\to\infty} E(W_{3,k}(z_j,I_{lj},z_i,a_n)|z_i) &= -\mathrm{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg| x_i\alpha_o\right)E(I_{li}|x_i\alpha_o)\\ &\quad + [\eta_{\theta_k}(x_i,\theta_o) - E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\\ &\quad \times \nabla E(I_{li}|x_i\alpha_o), \quad \text{a.e.,}\end{aligned} \tag{B.11}$$

and

$$\lim_{n\to\infty} E(W_{4,k}(z_j,I_{lj},z_i,a_n)|z_i) = -\mathrm{tr}\,\nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\bigg| x_i\alpha_o\right)\times E(I_{li}|x_i\alpha_o), \quad \text{a.e.} \tag{B.12}$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} E(\Phi_2(z_j,I_j,z_i,I_i,a_n)|z_i,I_i) &= -[\eta_\theta(x_i,\theta_o) - E(\eta_\theta(x_i,\theta_o)|x_i\alpha_o)]\sum_{l=1}^{L}\nabla E(I_{li}|x_i\alpha_o)\\ &= 0 \quad \text{a.e.,}\end{aligned}$$

because $\sum_{l=1}^{L} E(I_l|x\alpha_o) = 1$ implies $\sum_{l=1}^{L}\nabla E(I_l|x\alpha_o) = 0$. Finally, Proposition 4 implies that both $L_{n,1}^{(2)}$ and $L_{n,2}^{(2)}$ go to zero in probability as $\lim_{n\to\infty} na_{j,n}^{2s_j^*}/\delta_{j,n}^2 = 0$, $j = 1,2$, $\lim_{n\to\infty} na_{1,n}^{m+2}\delta_{1,n} = \infty$, and $\lim_{n\to\infty} na_{2,n}^{m+k+2}\delta_{2,n} = \infty$.

B.1.3: Show $L_n^{(1)} = o_p(1)$. The $L_n^{(1)}$ is also a U-statistic. Define

$$\begin{aligned}&\Psi_n(z_i,I_i,z_j,I_j)\\ &\quad = \Bigg\{d_{c,n}(z_i)\delta_{2,n}^{-1}\left[I_{1j}\frac{1}{a_{2,n}^{m+k}}J\left(\frac{\epsilon_i(\theta_o)-\epsilon_j(\theta_o)}{a_{2,n}},\frac{x_i\alpha_o - x_j\alpha_o}{a_{2,n}}\right) - C(z_i,\theta_o)\right]\\ &\qquad + \sum_{l=0}^{L} d_{ln}(z_i)\delta_{1,n}^{-1}\left[I_{lj}\frac{1}{a_{1,n}^{m}}K\left(\frac{x_i\alpha_o - x_j\alpha_o}{a_{1,n}}\right) - A(I_l|x_i,\theta_o)\right]\Bigg\}\\ &\qquad \times \frac{\partial \ln l(I_i,z_i,\theta_o)}{\partial\theta}.\end{aligned}$$

The $L_n^{(1)}$ can be rewritten as $L_n^{(1)} = (1/(\sqrt{n}(n-1)))\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Psi_n(z_i,I_i,z_j,I_j)$. By the asymptotic unbiasedness of nonparametric density estimates, it is apparent that $\lim_{n\to\infty} E[\Psi_n(z_i,I_i,z_j,I_j)|z_i,I_i] = 0$. It remains to investigate $E[\Psi_n(z_j,I_j,z_i,I_i)|z_i,I_i]$. By the index property in the model, $E(x|\epsilon(\theta_o),x\alpha_o) = E(x|x\alpha_o)$ that implies $E([\partial \ln f(\epsilon(\theta_o)|I_1 = 1, x\alpha_o\theta_o)]/\partial\theta|\epsilon(\theta_o),x\alpha_o) = 0$, and $E([\partial \ln P_l(x,\theta_o)]/\partial\theta|\epsilon(\theta_o),x\alpha_o) = 0$ because

$$\frac{\partial \ln f(\epsilon_i(\theta_o)|I_{1i}=1, x_i\alpha_o, \theta_o)}{\partial\theta}$$
$$= [E(\rho_\theta(x_i,\theta_o)|x_i\alpha_o) - \rho_\theta(x_i,\theta_o)]\bar{\nabla}\ln f(\epsilon_i(\theta_o)|I_{1i}=1, x_i\alpha_o, \theta_o) \quad \textbf{(B.13)}$$

and

$$\frac{\partial \ln P_l(x_i,\theta_o)}{\partial\theta} = [\eta_\theta(x_i,\theta_o) - E(\eta_\theta(x_i,\theta_o)|x_i\alpha_o)]\nabla \ln P(I_{li}|x_i\alpha_o), \quad \textbf{(B.14)}$$

where $\bar{\nabla}\ln f(\epsilon|I_1 = 1, t, \theta) = \partial/\partial(\epsilon, t)' \ln f(\epsilon|I_1 = 1, t, \theta)$ and $\nabla \ln P(I_l = 1|t) = \partial/\partial t' \ln P(I_l = 1|t)$ (see Appendix C for the derivation of (B.13) and (B.14)). It follows that

$$E\left(\frac{\partial \ln l(I_i, z_i, \theta_o)}{\partial\theta}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right)$$
$$= E\left\{E[I_{1i}|\epsilon_i(\theta_o), x_i]\frac{\partial \ln f(\epsilon_i(\theta_o)|I_{1i}=1, x_i\alpha_o, \theta_o)}{\partial\theta}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right\}$$
$$+ \sum_{l=1}^{L} E\left\{E[I_{li}|\epsilon_i(\theta_o), x_i]\frac{\partial \ln P_l(x_i,\theta_o)}{\partial\theta}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right\}$$
$$= E\left\{E[I_{1i}|\epsilon_i(\theta_o), x_i\alpha_o]\frac{\partial \ln f(\epsilon_i(\theta_o)|I_{1i}=1, x_i\alpha_o, \theta_o)}{\partial\theta}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right\}$$
$$+ \sum_{l=1}^{L} E\left\{E[I_{li}|\epsilon_i(\theta_o), x_i\alpha_o]\frac{\partial \ln P_l(x_i,\theta_o)}{\partial\theta}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right\}$$
$$=0$$

by using the index property that $E(I_l|\epsilon(\theta_o), x) = E(I_l|\epsilon(\theta_o), x\alpha_o)$. Hence $L_n^{(1)} = o_p(1)$ by Proposition 4.

B.1.4: Asymptotic Distribution of $\sqrt{n}S_n^*(\theta_o)$. It follows from (B.4) and the results in (B.1.1)–(B.1.3) that

$$\sqrt{n}S_n(\theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} t_n^\infty(z_i,\theta_o)\frac{\partial \ln l(I_i, z_i, \theta_o)}{\partial\theta} + o_p(1),$$

i.e., $\sqrt{n}S_n(\theta_o)$ is asymptotically equivalent to $\sqrt{n}S_n^*(\theta_o)$. Because

$$E\left[(t_n^\infty(z,\theta_o) - 1)\frac{\partial \ln l(I, z, \theta_o)}{\partial\theta}\right] = E\left\{E\left[(t_n^\infty(z,\theta_o) - 1)\frac{\partial \ln l(I, z, \theta_o)}{\partial\theta}\middle|\epsilon(\theta_o), x\alpha_o\right]\right\}$$
$$= E\left\{(t_n^\infty(z,\theta_o) - 1)E\left(\frac{\partial \ln l(I, z, \theta_o)}{\partial\theta}\middle|\epsilon(\theta_o), x\alpha_o\right)\right\}$$
$$= 0$$

and $E[(t_n^\infty(z,\theta_o) - 1)^2\|(\partial \ln l(I, z, \theta_o))/\partial\theta\|^2] \to 0$ by the dominated convergence theorem, then

$$\sqrt{n}S_n(\theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln l(I_i, z_i, \theta_o)}{\partial\theta} + o_p(1). \quad \textbf{(B.15)}$$

By the central limit theorem, $\sqrt{n}S_n^*(\theta_o) \xrightarrow{D} N(0, I(\theta_o))$.

B.2: Covariance Matrix. The following uniform law of large numbers is useful here, whose proof can be found in Ichimura and Lee (1991).

PROPOSITION 5 (A UNIFORM LAW OF LARGE NUMBERS). *Let $\{z_i\}$ be a sequence of i.i.d. random vectors. The measurable function $h(z,\theta,a_n)$ takes the form $h(z,\theta,a_n) = (1/a_n^d)h_1(z,\theta)h_2(z,\theta,(s(z,\theta))/a_n)$, where $a_n = O(1/n^p)$, $p > 0$, $d \geq 0$, $\theta \in \Theta$, and $s(z,\theta)$ is a finite-dimensional vector-valued function. Suppose that the following conditions are satisfied.*

(1) *Θ is a compact subset of a finite-dimensional Euclidean space.*
(2) *The function $h_1(z,\theta)$ is differentiable with respect to θ. The rth-order moment, where $r \geq 2$, of $\sup_{\theta\in\Theta}|h_1(z,\theta)|$ is finite. The first moment of $\sup_{\theta\in\Theta}|\partial h_1(z,\theta)/\partial\theta|$ exists and is finite.*
(3) *$|h_2| \leq c$ for some constant c.*
(4) *$E(h_1^2 h_2^2) = O(a_n^{\bar{d}})$ uniformly in $\theta \in \Theta$, for some $\bar{d}$.*
(5) *The functions $h_2(z,\theta,u)$ and $s(z,\theta)$ satisfy the bounded Lipschitz condition of order 1 with respect to θ and u.*

If $\lim_{n\to\infty}(n/\ln n)a_n^{2(1+1/r)d-\bar{d}} = \infty$, then $\sup_{\theta\in\Theta}|1/n\sum_{i=1}^n h(z_i,\theta,a_n) - E(h(z,\theta,a_n))| \xrightarrow{p} 0$. Furthermore, in addition to the preceding conditions, if $E(h(z,\theta,a_n))$ converges to a limit function $h_\infty(\theta)$ uniformly in $\theta \in \Theta$, then $\sup_{\theta\in\Theta}|1/n\sum_{i=1}^n h(z_i,\theta,a_n) - h_\infty(\theta)| \xrightarrow{p} 0$.

Define

$$I_n^*(\theta) = \frac{1}{n}\sum_{i=1}^n t_n(z_i,\theta)\Bigg\{I_{1i}\frac{\partial \ln f(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha,\theta)}{\partial\theta}\frac{\partial \ln f(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha,\theta)}{\partial\theta'} + \sum_{l=1}^L I_{li}\frac{\partial \ln P_l(x_i,\theta)}{\partial\theta}\frac{\partial \ln P_l(x_i,\theta)}{\partial\theta'}\Bigg\}$$

and

$$I_n^\infty(\theta) = \frac{1}{n}\sum_{i=1}^n t_n^\infty(z_i,\theta)\Bigg\{I_{1i}\frac{\partial \ln f(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha,\theta)}{\partial\theta}\frac{\partial \ln f(\epsilon_i(\theta)|I_{1i}=1,x_i\alpha,\theta)}{\partial\theta'} + \sum_{l=1}^L I_{li}\frac{\partial \ln P_l(x_i,\theta)}{\partial\theta}\frac{\partial \ln P_l(x_i,\theta)}{\partial\theta'}\Bigg\}.$$

To show that $I_n(\bar{\theta}_n)$ will converge in probability to $I(\theta_o)$, we proceed in three steps.

(1) $I_n(\theta) - I_n^*(\theta)$ converges to zero in probability, uniformly in θ.
Under the assumptions in the main text, by Lemma 1, (A.2) and (B.1),

$$t_n(z_i,\theta)\left(\frac{\partial \ln C_n(z_i,\theta)}{\partial\theta} - \frac{\partial \ln C(z_i,\theta)}{\partial\theta}\right)$$
$$= \{O_P[(na_n^{m+k}\delta_{2,n}^{-2}/\ln^2 n)^{-1/2}] + O_p(a_n^{s_2^*}\delta_{2,n}^{-2})\}\left[1 + \frac{\partial \ln C(z_i,\theta)}{\partial\theta}\right]$$

uniformly in z_i and θ. Similarly,

$$t_n(z_i,\theta)\left(\frac{\partial \ln A_n(z_i,\theta)}{\partial\theta} - \frac{\partial \ln A(z_i,\theta)}{\partial\theta}\right)$$

$$= \{O_P[(na_n^{m+k}\delta_{1,n}^{-2}/\ln^2 n)^{-1/2}] + O_p(a_n^{s_2^*}\delta_{1,n}^{-2})\}\left[1 + \frac{\partial \ln A(z_i,\theta)}{\partial\theta}\right]$$

uniformly in z_i and θ. Then it follows easily by (B.2) and the assumptions in the main text that $I_n(\theta) - I_n^*(\theta) = o_p(1)$ uniformly in θ.

(2) From (A.1) and (A.2) and an expansion similar to (B.3) it follows that $\sup_{(z_i,\theta)}|t_n(z_i,\theta) - t_n^\infty(z_i,\theta)| = o_p(1)$. Hence $I_n^*(\theta) - I_n^\infty(\theta) = o_p(1)$ uniformly in θ.

(3) By applying the uniform law of large numbers in Proposition 5, $I_n^\infty(\theta)$ converges in probability to $I(\theta)$ uniformly in θ.

Therefore, $I_n(\theta)$ converges in probability to $I(\theta)$ uniformly in θ. Because $\bar{\bar{\theta}}_n$ is a consistent estimate of θ_o and $I(\theta)$ is continuous in θ, $I(\bar{\bar{\theta}})$ converges in probability to $I(\theta_o)$. Under the identification condition that $I(\theta_o)$ is nonsingular, $I_n^{-1}(\theta_n) \xrightarrow{p} I^{-1}(\theta_o)$.

B.3: Show That $(\partial S_n(\bar{\bar{\theta}}))/\partial\theta' \xrightarrow{p} -I(\theta_o)$. From (2.10),

$$\frac{\partial S_n(\theta)}{\partial\theta'} = \frac{1}{n}\sum_{i=1}^n t_n(z_i,\theta)\frac{\partial^2 \ln l_n(I_i,z_i,\theta)}{\partial\theta\partial\theta'} + \frac{1}{n}\sum_{i=1}^n \frac{\partial \ln l_n(I_i,z_i,\theta)}{\partial\theta}\frac{\partial t_n(z_i,\theta)}{\partial\theta'}.$$

As

$$\frac{\partial \ln l_n(I_i,z_i,\theta)}{\partial\theta\,\partial\theta'}$$

$$= I_{1i}\left[\frac{1}{C_n(z_i,\theta)}\frac{\partial^2 C_n(z_i,\theta)}{\partial\theta\,\partial\theta'} - \frac{1}{C_n^2(z_i,\theta)}\frac{\partial C_n(z_i,\theta)}{\partial\theta}\frac{\partial C_n(z_i,\theta)}{\partial\theta'}\right.$$

$$\left. - \frac{1}{A_n(I_1|x_i,\theta)}\frac{\partial^2 A_n(I_1|x_i,\theta)}{\partial\theta\,\partial\theta'} + \frac{1}{A_n^2(I_1|x_i,\theta)}\frac{\partial A_n(I_1|x_i,\theta)}{\partial\theta}\frac{\partial A_n(I_1|x_i,\theta)}{\partial\theta'}\right]$$

$$+ \sum_{l=1}^L I_{li}\left[\frac{1}{A_n(I_l|x_i,\theta)}\frac{\partial^2 A_n(I_l|x_i,\theta)}{\partial\theta\,\partial\theta'} - \frac{1}{A_n^2(I_l|x_i,\theta)}\frac{\partial A_n(I_l|x_i,\theta)}{\partial\theta}\frac{\partial A_n(I_l|x_i,\theta)}{\partial\theta'}\right.$$

$$\left. - \frac{1}{A_n(1|x_i,\theta)}\frac{\partial^2 A_n(1|x_i,\theta)}{\partial\theta\,\partial\theta'} + \frac{1}{A_n^2(1|x_i,\theta)}\frac{\partial A_n(1|x_i,\theta)}{\partial\theta}\frac{\partial A_n(1|x_i,\theta)}{\partial\theta'}\right].$$

By following similar approaches to those in Section B.2,

$$\frac{1}{n}\sum_{i=1}^n t_n(z_i,\theta)\frac{\partial^2 \ln l_n(I_i,z_i,\theta)}{\partial\theta\,\partial\theta'} \xrightarrow{p} E\left(\frac{\partial^2 \ln l(I,z,\theta)}{\partial\theta\,\partial\theta'}\right)$$

uniformly in θ. As $\bar{\bar{\theta}}_n$ is a consistent estimate of θ_o, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\bar{\bar{\theta}}_n)\frac{\partial^2 \ln l_n(I_i,z_i,\bar{\bar{\theta}})}{\partial\theta\,\partial\theta'} \xrightarrow{p} E\left(I_1\frac{\partial^2 \ln f(\epsilon(\theta_o)|I_1=1,x\alpha_o,\theta_o)}{\partial\theta\,\partial\theta'}+\sum_{l=1}^{L} I_l\frac{\partial^2 \ln P_l(x,\theta_o)}{\partial\theta\,\partial\theta'}\right).$$

Because $f(\epsilon(\theta)|I_1=1,x\alpha,\theta)$ is a density function, under the regularity conditions that differentiation operators with respect to θ and the integration operator are exchangeable, the following "likelihood equation" is valid:

$$\int\left[\frac{\partial^2 \ln f(\epsilon(\theta)|I_1=1,x\alpha,\theta)}{\partial\theta\,\partial\theta'}+\frac{\partial \ln f(\epsilon(\theta)|I_1=1,x\alpha,\theta)}{\partial\theta}\frac{\partial \ln f(\epsilon(\theta)|I_1=1,x\alpha,\theta)}{\partial\theta'}\right] \times f(\epsilon(\theta)|I_1=1,x\alpha,\theta)\,dy=0.$$

At θ_o, it follows that

$$E\left(\frac{\partial \ln f(\epsilon(\theta_o)|I_1=1,x\alpha_o,\theta_o)}{\partial\theta}\frac{\partial \ln f(\epsilon(\theta_o)|I_1=1,x\alpha_o,\theta_o)}{\partial\theta'}\middle| I_1=1,x\right) = -E\left(\frac{\partial^2 \ln f(\epsilon(\theta_o)|I_1=1,x\alpha_o,\theta_o)}{\partial\theta\,\partial\theta'}\middle| I_1=1,x\right).$$

Similarly, as $\sum_{l=1}^{L} P_l(x,\theta)=1$, it follows that

$$\sum_{l=1}^{L}\left\{\frac{\partial^2 \ln P_l(x,\theta)}{\partial\theta\,\partial\theta'}+\frac{\partial \ln P_l(x,\theta)}{\partial\theta}\frac{\partial \ln P_l(x,\theta)}{\partial\theta'}\right\}P_l(x,\theta)=0.$$

At θ_o,

$$E\left(\sum_{l=1}^{L} I_l\frac{\partial \ln P_l(x,\theta_o)}{\partial\theta}\frac{\partial \ln P_l(x,\theta_o)}{\partial\theta'}\middle| x\right) = -E\left(\sum_{l=1}^{L} I_l\frac{\partial^2 \ln P_l(x,\theta_o)}{\partial\theta\,\partial\theta'}\middle| x\right).$$

With these identities, it follows that

$$\frac{1}{n}\sum_{i=1}^{n} t_n(z_i,\bar{\bar{\theta}}_n)\frac{\partial^2 \ln l_n(I_i,z_i,\bar{\bar{\theta}})}{\partial\theta\,\partial\theta'} \xrightarrow{p} I(\theta_o).$$

It remains to show that $1/n\sum_{i=1}^{n}(\partial \ln l_n(I_i,z_i,\bar{\bar{\theta}})/\partial\theta)((\partial t_n(z_i,\bar{\bar{\theta}}))/\partial\theta')$ converges to zero in probability. By following the approach in Section B.2, we can show that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l_n(I_i,z_i,\theta)}{\partial\theta}\frac{\partial t_n(z_i,\theta)}{\partial\theta'}-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta}\frac{\partial t_n^{\infty}(z_i,\theta)}{\partial\theta'}\xrightarrow{p} 0$$

uniformly in θ, where

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta}\frac{\partial t_n^{\infty}(z_i,\theta)}{\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n} k_{c,n}^{\infty}(z_i,\theta)\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta}\frac{\partial \ln C(z_i,\theta)}{\partial\theta'} + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta}\sum_{j=0}^{L} k_{j,n}^{\infty}(z_i,\theta)\frac{\partial \ln A(I_j|x_i,\theta)}{\partial\theta'}$$

with $k_{c,n}^{\infty}(z_i,\theta)=\prod_{l=0}^{L} t^*(A(I_l|x_i,\theta)/\delta_{1,n})\times\nabla t^*(C(z_i,\theta)/\delta_{2,n})(C(z_i,\theta)/\delta_{2,n})$ and

$$k_{j,n}^{\infty}(z_i,\theta)=t^*(C(z_i,\theta)/\delta_{2,n})\prod_{l=0,l\neq j}^{L} t^*(A(I_l|x_i,\theta)/\delta_{1,n}) \times\nabla t^*(A(I_j|x_i,\theta)/\delta_{1,n})\frac{A(I_j|x_i,\theta)}{\delta_{1,n}}$$

for $j = 0,\ldots,L$. Because $k_{j,n}^{\infty}$ and $k_{c,n}^{\infty}$ are bounded and they do not vanish at z_i where $\frac{1}{2} < C(z_i,\theta)/\delta_{2,n} < 1$ and $\frac{1}{2} < A(I_l|x_i,\theta)/\delta_{1,n} < 1$, for all $l = 0,\ldots,L$, it follows from the uniform law of large numbers in Proposition 5 that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l(I_i,z_i,\theta)}{\partial\theta}\frac{\partial t_n^{\infty}(z_i,\theta)}{\partial\theta'}\xrightarrow{p}0$$

uniformly in θ. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln l_n(I_i,z_i,\bar{\bar{\theta}})}{\partial\theta}\frac{\partial t_n(z_i,\bar{\bar{\theta}})}{\partial\theta'}\xrightarrow{p}0.$$

APPENDIX C. SOME ALGEBRAIC EXPRESSIONS

C.1: Derivation of (B.13) and (B.14). Because $E(x|I_1=1,\epsilon(\theta_o),x\alpha_o) = E(x|x\alpha_o)$, it implies that $(\partial/\partial\epsilon)E(x|I_1=1,\epsilon,x\alpha_o)=0$. Therefore

$$\begin{aligned}\frac{\partial C(z,\theta_o)}{\partial\theta_k} &= \Bigg[-\mathrm{tr}\,\nabla E\left(x\,\frac{\partial\alpha_o}{\partial\theta_k}\Bigg|\,x\alpha_o\right)\times E(I_1|\epsilon(\theta_o),x\alpha_o)\\
&\quad + [E(\rho_{\theta_k}(x,\theta_o)|x\alpha_o)-\rho_{\theta_k}(x,\theta_o)]\nabla E(I_1|\epsilon(\theta_o),x\alpha_o)\Bigg]g(\epsilon(\theta_o),x\alpha|\theta_o)\\
&\quad + [E(\rho_{\theta_k}(x,\theta_o)|x\alpha_o)-\rho_{\theta_k}(x,\theta_o)]E(I_1|\epsilon(\theta_o),x\alpha_o)\nabla g(\epsilon(\theta_o),x\alpha_o|\theta_o)\\
&= -\mathrm{tr}\,\nabla E\left(x\,\frac{\partial\alpha_o}{\partial\theta_k}\Bigg|\,x\alpha_o\right)\times E(I_1|\epsilon(\theta_o),x\alpha_o)g(\epsilon(\theta_o),x\alpha_o|\theta_o)\\
&\quad + [E(\rho_{\theta_k}(x,\theta_o)|x\alpha_o)-\rho_{\theta_k}(x,\theta_o)]\times\nabla\{E(I_1|\epsilon(\theta_o),x\alpha_o)g(\epsilon(\theta_o),x\alpha_o|\theta_o)\}.\end{aligned}\tag{C.1}$$

Similarly,

$$\begin{aligned}\frac{\partial A(I_l|x,\theta_o)}{\partial\theta_k} &= -\mathrm{tr}\,\nabla E\left(x\,\frac{\partial\alpha_o}{\partial\theta_k}\Bigg|\,x\alpha_o\right)\times E(I_l|x\alpha_o)p(x\alpha_o|\theta_o)\\
&\quad + (x-E(x|x\alpha_o))\,\frac{\partial\alpha_o}{\partial\theta_k}\times\nabla\{E(I_1|x\alpha_o)p(x\alpha_o|\theta_o)\}.\end{aligned}\tag{C.2}$$

As $E(I_1|\epsilon(\theta),x\alpha)g(\epsilon(\theta),x\alpha|\theta) = g(\epsilon(\theta),x\alpha|I_1=1,\theta)E(I_1)$ and $E(I_1|x\alpha)p(x\alpha|\theta) = p(x\alpha|I_1=1,\theta)E(I_1)$, it follows that

$$\begin{aligned}&\nabla'\ln\{E(I_1|\epsilon(\theta),x\alpha)g(\epsilon(\theta),x\alpha|\theta)\} - (0,\nabla'\ln\{E(I_1|x\alpha)p(x\alpha|\theta)\})\\
&\quad = \nabla'\ln g(\epsilon(\theta),x\alpha|I_1=1,\theta) - (0,\nabla'\ln p(x\alpha|I_1=1,\theta))\\
&\quad = \bar{\nabla}'\ln f(\epsilon(\theta)|I_1=1,x\alpha,\theta),\end{aligned}$$

because $f(\epsilon(\theta)|I_1=1,x\alpha,\theta)=g(\epsilon(\theta),x\alpha|I_1=1,\theta)/p(x\alpha|I_1=1,\theta)$. Using these relations,

$$\frac{1}{C(z,\theta_o)}\frac{\partial C(z,\theta_o)}{\partial\theta_k}-\frac{1}{A(I_1|x,\theta_o)}\frac{\partial A(I_1|x,\theta_o)}{\partial\theta_k}$$
$$=[E(\rho_{\theta_k}(x,\theta_o)|x\alpha_o)-\rho_{\theta_k}(x,\theta_o)]\times\nabla\ln\{E(I_1|\epsilon(\theta_o),x\alpha_o)g(\epsilon(\theta_o),x\alpha_o|\theta_o)\}$$
$$-[\eta_{\theta_k}(x,\theta_o)-E(\eta_{\theta_k}(x,\theta_o)|x\alpha_o)]\times\nabla\ln\{E(I_1|x\alpha_o)p(x\alpha_o|\theta_o)\}$$
$$=[E(\rho_{\theta_k}(x,\theta_o)|x\alpha_o)-\rho_{\theta_k}(x,\theta_o)]\overline{\nabla}\ln f(\epsilon(\theta_o)|I_1=1,x\alpha_o,\theta_o),$$

which is (B.13). Equation (B.14) can be deduced in a similar way.

C.2: Derivation of (B.5)–(B.8). Because $t_n(z_i,\theta_o)$ and $C(z_j,\theta_o)$ are functions of $(y_j-x_j\gamma_o,x_j\alpha_o)$ and are continuous everywhere, Proposition 2 implies that

$$\lim_{n\to\infty}E(T_{1,k}(z_j,I_j,z_i,a_n,\theta_o)|z_i)$$
$$=-g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)$$
$$\times\operatorname{tr}\nabla\Bigg\{[\rho_{\theta_k}(x_i,\theta_o)E(I_{1i}|\epsilon_i(\theta_o),x_i\alpha_o)-E(I_{1i}\rho_{\theta_k}(x_i,\theta_o)|\epsilon_i(\theta_o),x_i\alpha_0)]$$
$$\times\frac{1}{C(z_i,\theta_o)}\Bigg\}$$
$$-[\rho_{\theta_k}(x_i,\theta_o)E(I_{1i}|\epsilon_i(\theta_o),x_i\alpha_o)-E(I_{1i}\rho_{\theta_k}(x_i,\theta_o)|\epsilon_i(\theta_o),x_i\alpha_o)]\frac{1}{C(z_i,\theta_o)}$$
$$\times\nabla g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)$$
$$=-g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)\times\operatorname{tr}\nabla\Bigg\{[\rho_{\theta_k}(x_i,\theta_o)-E(\rho_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]$$
$$\times\frac{1}{g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)}\Bigg\}$$
$$-[\rho_{\theta_k}(x_i,\theta_o)-E(\rho_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\nabla\ln g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)$$
$$=-\operatorname{tr}\nabla[E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)-\eta_{\theta_k}(x_i,\theta_o)]$$
$$+\frac{1}{g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)}[\rho_{\theta_k}(x_i,\theta_o)-E(\rho_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\times\nabla g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)$$
$$-[\rho_{\theta_k}(x_i,\theta_o)-E(\rho_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\nabla\ln g(\epsilon_i(\theta_o),x_i\alpha_o|\theta_o)$$
$$=-\operatorname{tr}\nabla E\left[\frac{\partial x_i\alpha_o}{\partial\theta_k}\middle|x_i\alpha_o\right],$$

which is (B.5). Similarly,

$$\lim_{n\to\infty} E(T_{3,k}(x_j, I_j, x_i, a_n, \theta_n)|x_i)$$

$$= -p(x_i\alpha_o|\theta_o)$$

$$\times \operatorname{tr} \nabla\left\{[E(I_{1i}\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)E(I_{1i}|x_i\alpha_o)]\frac{1}{A_1(x_i,\theta_o)}\right\}$$

$$- [E(I_{1i}\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)E(I_{1i}|x_i\alpha_o)]\frac{1}{A_1(x_i,\theta_o)} \times \nabla p(x_i\alpha_o|\theta_o)$$

$$= -p(x_i\alpha_o|\theta_o) \times \operatorname{tr} \nabla\left\{[E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\frac{1}{p(x_i\alpha_o|\theta_o)}\right\}$$

$$- [E(\eta_{\theta_k}(x_i,\theta_o)|x,\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\nabla \ln p(x_i\alpha_o|\theta_o)$$

$$= -\operatorname{tr} \nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right) + [E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\nabla \ln p(x_i\alpha_o|\theta_o)$$

$$- [E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\nabla \ln p(x_i\alpha_o|\theta_o)$$

$$= -\operatorname{tr} \nabla E\left(\frac{\partial x_i\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right) \quad \text{a.e.,}$$

which is (B.7). On the other hand, Proposition 1 implies that

$$\lim_{n\to\infty} E(T_{2,k}(z_j, I_j, z_i, a_n, \theta_o)|z_i) = E\left[I_{1i}\frac{\partial C(z_i,\theta_o)}{\partial\theta_k}\middle|\epsilon_i(\theta_o), x_i\alpha_o\right]$$

$$\times\left(\frac{1}{C(z_i,\theta_o)}\right)^2 g(\epsilon_i(\theta_o), x_i\alpha_o|\theta_o)$$

$$= E\left[\frac{\partial C(z_i,\theta_o)}{\partial\theta_k}\middle| I_{1i}=1,\epsilon_i(\theta_o), x_i\alpha_o\right]\frac{1}{C(z_i,\theta_o)} \quad \text{a.e.}$$

With relation (C.1),

$$E\left[\frac{\partial C(z_i,\theta_o)}{\partial\theta_k}\middle| I_{1i}=1,\epsilon_i(\theta_o), x_i\alpha_o\right]$$

$$= -\operatorname{tr} \nabla E\left(x_i\frac{\partial\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right) \times E(I_{1i}|\epsilon_i(\theta_o), x_i\alpha_o)g(\epsilon_i(\theta_o), x_i\alpha_o|\theta_o)$$

$$+ [E(\rho_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - E(\rho_{\theta_k}(x_i,\theta_o)|I_{1i}=1,\epsilon_i(\theta_o), x_i\alpha_o)]$$

$$\times \nabla\{E(I_{1i}|\epsilon_i(\theta_o), x_i\alpha_o)g(\epsilon_i(\theta_o), x_i\alpha_o|\theta_o)\}$$

$$= -\operatorname{tr} \nabla E\left(x\frac{\partial\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right) \times E(I_{1i}|\epsilon_i(\theta_o), x_i\alpha_o)g(\epsilon_i(\theta_o), x_i\alpha_o|\theta_o) \quad \text{a.e.}$$

Hence $\lim_{n\to\infty} E(T_{2,k}(z_j,I_j,z_i,a_n,\theta_n)|z_i) = -\text{tr}\,\nabla E(x_i|x_i\alpha_o)(\partial\alpha(\theta_o)/\partial\theta_k)$ a.e., which is (B.6). Relation (C.2) implies that

$$E\left(\frac{\partial A_1(x_i,\theta_o)}{\partial\theta_k}\middle| I_{1i}=1, x_i\alpha_o\right) = -\text{tr}\,\nabla E\left(x_i\frac{\partial\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right)\times E(I_{1i}|x_i\alpha_o)p(x_i\alpha_o|\theta_o)$$
$$+ [E(\eta_{\theta_k}(x_i,\theta_o)|I_{1i}=1,x_i\alpha_o)$$
$$- E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\times\nabla\{E(I_{1i}|x_i\alpha_o)p(x_i\alpha_o|\theta_o)\}$$
$$= -\text{tr}\,\nabla E\left(x_i\frac{\partial\alpha_o}{\partial\theta_k}\middle| x_i\alpha_o\right)\times E(I_{1i}|x_i\alpha_o)p(x_i\alpha_o|\theta_o).$$

Hence it follows that

$$\lim_{n\to\infty} E(T_{4,k}(x_j,I_j,x_i,a_n,\theta_o)|x_i) = E\left(I_{1i}\frac{\partial A_1(x_i,\theta_o)}{\partial\theta}\middle| x_i\alpha_o\right)\left(\frac{1}{A_1(x_i,\theta_o)}\right)^2 p(x_i\alpha_o|\theta_o)$$
$$= E\left(\frac{\partial A_1(x_i,\theta_o)}{\partial\theta}\middle| I_{1i}=1,x_i\alpha_o\right)\frac{1}{A_1(x_i,\theta_o)}$$
$$= -\text{tr}\,\nabla E(x_i|x_i\alpha_o)\frac{\partial\alpha_o}{\partial\theta_k}\quad\text{a.e.,}$$

which is (B.8).

C.3: Derivation of (B.9)–(B.12). The expressions of W_1 and W_2 are similar to the expressions T_3 and T_4, respectively. Following similar analysis to that previously described for T_3 and T_4, (B.9) and (B.10) hold. Proposition 2 implies that

$$\lim_{n\to\infty} E(W_{3,k}(x_j,I_{lj},x_i,a_n,\theta_n)|x_i)$$
$$= -p(x_i\alpha_o|\theta_o)$$
$$\times\,\text{tr}\,\nabla\left\{[E(I_{li}\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)E(I_{li}|x_i\alpha_o)]\frac{1}{p(x_i\alpha_o|\theta_o)}\right\}$$
$$- [E(I_{li}\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)E(I_{li}|x_i\alpha_o)]\nabla\ln p(x_i\alpha_o|\theta_o)$$
$$= -p(x_i\alpha_o|\theta_o)\Bigg\{\text{tr}\,\nabla[E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\frac{E(I_{li}|x_i\alpha_o)}{p(x_i\alpha_o|\theta_o)}$$
$$+ [E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]$$
$$\times\left(\frac{1}{p(x_i\alpha_o|\theta_o)}\nabla E(I_{li}|x_i\alpha_o) - \frac{E(I_{li}|x_i\alpha_o)}{p^2(x_i\alpha_o|\theta_o)}\nabla p(x_i\alpha_o|\theta_o)\right)\Bigg\}$$
$$- [E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\frac{E(I_{li}|x_i\alpha_o)}{p(x_i\alpha_o|\theta_o)}\nabla p(x_i\alpha_o|\theta_o)$$
$$= -\text{tr}\,\nabla E(x_i|x_i\alpha_o)\frac{\partial\alpha_o}{\partial\theta_k}E(I_{li}|x_i\alpha_o)$$
$$- [E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o) - \eta_{\theta_k}(x_i,\theta_o)]\nabla E(I_{li}|x_i\alpha_o)\quad\text{a.e.,}$$

which is (B.11). Finally, Proposition 1 and (C.2) imply that

$$\lim_{n\to\infty} E(W_{4,k}(x_j, I_{lj}, x_i, a_n, \theta_o)|x_i)$$

$$= E\left(I_{li} \left.\frac{\partial A(1|x_i,\theta_o)}{\partial\theta}\right| x_i\alpha_o\right)\left(\frac{1}{A(1|x_i,\theta_o)}\right)^2 p(x_i\alpha_o|\theta_o)$$

$$= E\left(\left.\frac{\partial A(1|x_i,\theta_o)}{\partial\theta}\right| I_{li}=1, x_i\alpha_o\right)\frac{E(I_{li}|x_i\alpha_o)}{p(x_i\alpha_o|\theta_o)}$$

$$= \left\{-\text{tr}\,\nabla E(x_i|x_i\alpha_o)\frac{\partial\alpha_o}{\partial\theta_k} p(x_i\alpha_o|\theta_o) + [E(\eta_{\theta_k}(x_i,\theta_o)|I_{li}=1, x_i\alpha_o)\right.$$

$$\left. - E(\eta_{\theta_k}(x_i,\theta_o)|x_i\alpha_o)]\nabla p(x_i\alpha_o|\theta_o)\right\} \times \frac{E(I_{li}|x_i\alpha_o)}{p(x_i\alpha_o|\theta_o)}$$

$$= -\text{tr}\,\nabla E(x_i|x_i\alpha_o)\frac{\partial\alpha_o}{\partial\theta_k} E(I_{li}|x_i\alpha_o) \quad \text{a.e.,}$$

which is (B.12).

APPENDIX D: A USEFUL LEMMA FOR DERIVATIVES OF CONDITIONAL EXPECTATIONS

In this section, we derive some formulas, which are useful here and interesting on their own, for probability limits for derivatives of some nonparametric estimates with respect to certain parameter vectors.

For a random vector $(y, x) \in R^{1+k}$ and a vector $\alpha \in R^k$, let $E(y|x\alpha = t)$ denote the conditional expectation of y given $x\alpha$ evaluated at t. It is obvious that this conditional expectation depends on the parameter vector α through $x\alpha$. In the following lemma, we derive a formula for the derivative of $E(y|x\alpha = t)p(x\alpha = t)$ with respect to the parameter α, where $p(x\alpha = t)$ denotes the density of $x\alpha$ evaluated at t.

LEMMA 2. *Assume that $E(y|x\alpha = t)p(x\alpha = t)$ and $E(yx|x\alpha = t)p(x\alpha = t)$ are continuously differentiable with respect to α and t and these two functions and their derivatives are uniformly bounded by a square integrable function. Then $\partial[E(y|x\alpha = t)p(x\alpha = t)]/\partial\alpha = -(E(yx|x\alpha = t)p(x\alpha = t))'$, where $'$ denotes the derivative with respect to t.*

Proof. Let $E(y|x\alpha = t) = h(\alpha, t)$. By the definition of the conditional expectation,

$$\int y\lambda(x\alpha)\,dP = \int E(y|x\alpha)\lambda(x\alpha)\,dP = \int p(\alpha,t)h(\alpha,t)\lambda(t)\,dt \qquad \textbf{(D.1)}$$

for any real function $\lambda \in C$, where P is the underlying probability measure, C denotes the set of real-valued functions on R that have compact support and continuous partial derivatives of all order, and $p(\alpha,t) = p(x\alpha = t)$, the density function of $x\alpha$ evaluated at t. Let $g(\alpha,t) = E(yx|x\alpha = t)$, the conditional expectation of xy given $x\alpha$ evaluated at t. Taking the partial derivative with respect to α on both sides, for the left-hand side we obtain

$$L_\alpha = \int xy\lambda'(x\alpha)\,dP = \int E(xy|x\alpha)\lambda'(x\alpha)\,dP = \int g(\alpha,t)p(\alpha,t)\lambda'(t)\,dt$$
$$= -\int [g(\alpha,t)p(\alpha,t)]'\lambda(t)\,dt. \tag{D.2}$$

The last equality follows from an integration by parts formula. On the other hand, the derivative of the right-hand side is equal to $R_\alpha = \int \lambda(t)([\partial(h(\alpha,t)p(\alpha,t))]/\partial\alpha)\,dt$. Therefore,

$$\int \lambda(t)\left[\frac{\partial(h(\alpha,t)p(\alpha,t))}{\partial\alpha} + (g(\alpha,t)p(\alpha,t))'\right]dt = 0 \tag{D.3}$$

holds for any function $\lambda \in C$. Let $f(\alpha,t) = [\partial(h(\alpha,t)p(\alpha,t))]/\partial\alpha + (g(\alpha,t)p(\alpha,t))'$. Then we have $\int \lambda(t)f(\alpha^*,t)\,dt = 0$ holds for any function $\lambda \in C$ and α^* in the domain of α. Let $f_l(\alpha,t)$ denote the lth component $f(\alpha,t)$. For each l, as $f_l(\alpha,t)$ is continuous and uniformly bounded by a square integrable function, Lemma A.2 in Chamberlain (1986) implies that for any $\epsilon > 0$ we can choose a $\lambda_l \in C$ such that $(\int (f_l(\alpha^*,t) - \lambda_l(t))^2\,dt)^{1/2} < \epsilon$. Therefore

$$\int f_l^2(\alpha^*,t)\,dt = \left\| \int f_l(\alpha^*,t)(\lambda_l(t) - f_l(\alpha^*,t))\,dt \right\|$$
$$\leq \left(\int f_l^2(\alpha^*,t)\,dt\right)^{1/2}\left(\int (f_l(\alpha^*,t) - \lambda_l(t))^2\,dt\right)^{1/2}$$
$$\leq \left(\int f_l^2(\alpha^*,t)\,dt\right)^{1/2}\epsilon.$$

Because it holds for any $\epsilon > 0$, thus $\int f_l^2(\alpha^*,t)\,dt = 0$. Hence $f_l^2(\alpha^*,t) = 0$ on the real line. Consequently, we have $[\partial(h(\alpha,t)p(\alpha,t))]/\partial\alpha = -(g(\alpha,t)p(\alpha,t))'$ as asserted. ■