SEMIPARAMETRIC MAXIMUM PROFILE LIKELIHOOD ESTIMATION OF POLYTOMOUS AND SEQUENTIAL CHIOCE MODELS

by

Lung-Fei Lee

Discussion Paper No. 253, December 1989

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, Minn 55455

Semiparametric Maximum Profile Likelihood Estimation of Polytomous and Sequential Choice Models

by

Lung-Fei Lee
Department of Economics
University of Minnesota
Minneapolis,Mn 55455

May, 1989
December, 1989 (revised)

Abstract

This article considers semiparametric estimation of discrete choice models. The estimation methods are some semiparametric maximum profile likelihood methods which generalize Klein and Spady [1987] to the estimation of polytomous choice and sequential choice models. Special emphases are on the correction of asymptotic bias and negative density estimates caused by high order kernel density estimation. The estimators are shown to be $\sqrt{n}$ consistent and asymptotically normal. They attain the asymptotic efficiency bound of semiparametric estimation with some infinite dimensional parameter spaces of index probability functions.

JEL classification number: 211
Keywords:
Polytomous choices, sequential choices, semiparametric model, semiparametric estimation, profile likelihood, index probability functions, efficient estimators

Correspondence Address
Lung-Fei Lee, Department of Economics, University of Minnesota, Minneapolis, MN 55455

Semiparametric Maximum Profile Likelihood Estimation of Polytomous and Sequential Choice Models

by Lung-Fei Lee*

1. Introduction

In this article, we will consider estimation of the discrete choices models with any finite number of alternatives within the semiparametric framework. Asymptotic efficiency bounds for $\sqrt{n}$ consistent estimators for the binary choice models have been derived in Chamberlain [1986] and Cosslett [1987]. Severini and Wong [1987] provide a general characterization for semiparametric estimators to be asymptotically efficient when they are derived from the maximization of profile likelihoods. The term 'profile likelihood' refers to likelihood function with its unknown density or probability functions replaced by some estimated random functions. In the econometric literature, various approaches have been proposed for the estimation of discrete choice models (Manski [1975,1987] ,Han [1987] ,Ichimura [1987], Klein and Spady [1987] and Thompson [1989] among others). However, except the approach by Klein and spady [1987], the other approaches do not attain the asymptotic efficiency bound. The approach of Klein and spady [1987] is a maximum profile likelihood method. In their article, Klein and Spady [1987] has considered the estimation of binary choice models. Generalization of the method to the estimation of polytomous choice models is not straightforward.

The polytomous choice models are, in general, models with multiple indices. As shown in Ichimura and Lee [1988],estimation methods involving multiple indices and kernel density estimation require the selection of kernel functions with zero moments up to certain high order. Such kernel functions are needed to correct biases caused by kernel density estimations so that the semiparametric estimator can be properly located at the true parameter vector of interest. These kernel functions can neither be proper density functions nor positive value functions and the kernel estimated choice probabilities can be negative. This will create difficulty to evaluate the log likelihood function and the computation of the estimates. These complications raise questions on whether the Klein and Spady's approach can be generalized to the polytomous cases.

In this article, with some modifications, we are able to provide such generalization. We will show that our estimator is consistent and asymptotically normal. It satisfies Severini and Wong's efficient criterion and hence is asymptotically efficient with some infinite dimensional parameter spaces of probability functions. Our asymptotic analysis generalizes the analysis in Ichimura [1987] and Ichimura and Lee [1988] for semiparametric least squares estimation to semiparametric maximum profile likelihood estimation (MPLE). The methods are also applicable to the estimation of ordered choice models. It can further be generalized to the estimation of sequential discrete choice models with correlated disturbances.

Section 2 of this article will present the polytomous choice models and the semiparametric estimation methods. The regularity conditions for the models and the main results are summarized in section3. Section 4 provides generalization to the estimation of sequential discrete choice models. Section 5 reports some Monte Carlo simulations on the small sample performance of the estimators. The detail proofs of the consistency and asymptotic normality of the estimators are collected in sections 6, 7 and 8. An appendix is also included for some of the proofs.

* I appreciate having financial support under NSF grant number SES-8809939 for my research.

2.Polytomous Choices and Maximum Profile Likelihood Estimation

Within an utility maximization framework (McFadden [1974]), suppose there are L different alternatives, let

$$U_l = x\delta_l - u_l \tag{2.1}$$

be the associated utility for alternative $l, l = 1, ..., L$. Let I_l be a dichotomous indicator that the alternative l will be chosen:

$$I_l = 1 \Longleftrightarrow x\delta_l - x\delta_j \geq u_l - u_j, \qquad j \neq l,\ j = 1, \ldots, L \tag{2.2}$$

and $I_l = 0$, otherwise. Let $x\alpha_l = x\delta_l - x\delta_L$, $l = 1, \ldots, L-1$. Under the assumption that the disturbances $u_l, l = 1, ..., L$ are statistically independent with x, the choice probabilities conditional on x are functions of the indices $x\alpha_1, \ldots, x\alpha_{L-1}$ evaluated at the true parameter vector.

As a generalization, let us assume the number of indices are m where m can be greater than, equal to or less than the number of choices L. This generalization will include models with ordered choices, systems with multivariate qualitative dependent variables and discrete choice models with heteroskedastic disturbances u_l in (2.2) of which the conditional distribution conditional on x depends on some indices. To capture constraints, let θ be the vector of deep parameters in Θ, $\Theta \subseteq R^k$ and α_l's be functions of θ. As shown in Ichimura and Lee [1988], the model can be identified only if the indices $x\alpha_1(\theta), \cdots, x\alpha_m(\theta)$ are all continuous variables for each θ and they are distinguishable from each other. Hence, in this model, we assume that for each index $x\alpha_l(\theta)$, there exists a distinct continuous variable s_l in $x\alpha_l(\theta)$ which does not appear in other indices. Thus, $x = (s_1, \cdots, s_m, w)$ and $x\alpha_l(\theta) = s_l + w\beta_l(\theta)$,where w are explanatory variables which can appear in any indices. The choice probabilities are

$$Prob(I_l = 1|x) = E(I_l|x\alpha(\theta_o)) \tag{2.3}$$

where θ_o denotes the true parmeter vector, and $x\alpha(\theta) = (x\alpha_1(\theta), \cdots, x\alpha_m(\theta))$.

For any $\theta \in \Theta$, let $P_l(x\alpha|\theta) = E(I_l|x\alpha(\theta))$ be the conditional expectation of I_l conditional on $x\alpha(\theta)$. Given a random sample of size n, the probability function $P_l(x_i\alpha|\theta)$ at x_i can be estimated by nonparametric kernel regression function (Ichimura[1987] and Klein and Spady[1987]):

$$P_{n,l}(x_i, \theta) = \frac{A_{n,l}(x_i, \theta)}{B_n(x_i, \theta)} \tag{2.4}$$

where

$$A_{n,l}(x_i, \theta) = \frac{1}{(n-1)a_n^m} \sum_{j \neq i}^{n} I_{lj} K\left(\frac{x_i\alpha(\theta) - x_j\alpha(\theta)}{a_n}\right), \tag{2.5}$$

$$B_n(x_i, \theta) = \frac{1}{(n-1)a_n^m} \sum_{j \neq i}^{n} K\left(\frac{x_i\alpha(\theta) - x_j\alpha(\theta)}{a_n}\right) \tag{2.6}$$

and $K(\cdot)$ is a kernel function on R^m and $a_n > 0$ is a bandwidth sequence(Rao[1983]). As a generalization of Klein and Spady[1987], one may consider the following semiparametric maximum profile likehood estimation:

$$\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} I_{X_n}(x_i) \ln L_n(\theta|I_i, x_i) \tag{2.7}$$

where

$$\ln L_n(\theta|I_i, x_i) = \sum_{l=1}^{L} I_{li} \ln P_{n,l}(x_i, \theta) \tag{2.8}$$

and X_n is an appropriately chosen subset of x which effectively trim the values of x. To quarantee that the derived estimate $\hat{\theta}$ from (2.7) is $\sqrt{n}$ consistent and asymptotically normal with center located at θ_o, kernel functions with zero moments up to certain order are needed so as to reduce the asymptotic biases caused by the kernel density estimates. This requirement, however, can cause complications for implimentation of the estimation in (2.7). As the kernel function will not be strictly positive everywhere, the estimated functions $P_{n,l}(x_i, \theta)$ in (2.4) are not necessarily positive . While the estimation method in (2.7) is asymptotically well defined and the estimate $\hat{\theta}$ exists with probability one as the sample size n goes to infinity, the function $\ln P_{n,l}(x_i, \theta)$ may not be numerically well defined everywhere for a given sample. If that happens, there is a need to have systematic procedures to modify the objective function for estimation.

A possible modification of the method is

$$\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} I(A_{n,l}(x_i, \theta) > \Delta_n, B_n(x_i, \theta) > \Delta_n) \ln P_{n,l}(x_i, \theta) \tag{2.9}$$

where $\Delta_n \geq 0$ is a specified sequence such that $\lim_{n \to \infty} \Delta_n = 0$, and $I(A_{n,l}(x_i, \theta) > \Delta_n, B_n(x_i, \theta) > \Delta_n)$ is an indicator of the event $A_{n,l}(x_i, \theta) > \Delta_n$ and $B_n(x_i, \theta) > \Delta_n$. Essentially, we ignore the troublesome terms in the evaluation of the objective function. This does create discontinuity in the ojective function. However, we will show that the estimate $\hat{\hat{\theta}}$ derived from (2.9) is still $\sqrt{n}$ consistent and asymptotically normal. Indead, $\hat{\hat{\theta}}$ is asymptotically equivalent to $\hat{\theta}$ and is asymptotically efficient. More modification by imposing penalty on the likelihood function when $A_{n,l}(x_i, \theta) \leq \Delta_n$ or $B_n(x_i, \theta) \leq \Delta_n$ is also possible.[1]

[1] In the Monte Carlo experiment below, we have adopted a modification which bounds the estimated probabilities between zero and one. The asymptotic distribution will remain the same as the MPLE from (2.7) or (2.9) if the penalty functions are bounded.

3. The Main Results: Consistency and Asymptotic Distribution

In this section, we summarize the main results. The detail proofs are in subsequent sections. To justify the consistency of our estimator, we assume that the following conditions hold for our model.

Assumption 1 :

(1) The samples $(I_i, x_i), i = 1, ..., n$ are i.i.d.

(2) The parameter space Θ is a compact subset of R^k and θ_o is in the interior of Θ. $\alpha(\theta)$ is continuous and satisfies a Lipschitz condition on Θ.[2]

(3) The support S_x of $x, x = (s_1, \cdots, s_m, w)$, is $[T_1, T_2] \times S_w$ where $[T_1, T_2]$ is a m dimensional compact rectangle and S_w is the support of w which is also a compact set. For each w in its support, the conditional density $f(s|w)$ of $s, s = (s_1, \cdots, s_m)$, conditional on w is continuous in s, strictly positive and uniformly bounded on its support S_x.

(4) The choice probalilities $E(I_l|x\alpha(\theta_o)), l = 1, \ldots, L$ are strictly positive and continuous on S_x.

Assumption 2:

(1) The kernel function $K(v)$ on R^m is a function with bounded support D such that $\int_D K(v)dv = 1$ and $\int_D |K(v)|\, dv < \infty$.

(2) The kernel function $K(v)$ satisfies a Lipschitz condition and it goes to zero at the boundary of D.

(3) The bandwidth sequence a_n is a positive sequence such that $\lim_{n \to \infty} a_n = 0$.

(4) $X_n = \{x \in S_x : T_{1,l} + \delta_n \leq s_l \leq T_{2,l} - \delta_n, l = 1, \ldots, m\}$ where $T_{1,l}$ and $T_{2,l}$ are the l components of T_1 and T_2 respectively and $\{\delta_n\}$ is a positive sequence such that $\lim_{n \to \infty} \frac{\delta_n}{a_n} = \infty$.

An immediate consequence of assumption 1 is that the density function $p(t|\theta)$ of the indices $x\alpha(\theta)$ and the conditional expectations $E(I_l|x\alpha(\theta))$, $l = 1, \ldots, L$ are continuous in (x, θ) and are strictly positive on $S_x \times \Theta$. This is so, since $t = s + wB$ where $t = x\alpha, s = (s_1, \cdots, s_m)$ and $B = (\beta_1, \cdots, \beta_m)$,

$$p(t|\theta) = \int f(t - wB|w)d\nu(w) \tag{3.1}$$

and

$$\begin{aligned} E(I_l|t) &= E[P_l(x\alpha(\theta_o)|\theta_o)|t] \\ &= \int P_l(t - wB + wB(\theta_o)|\theta_o)f(t - wB|w)d\nu(w)/p(t|\theta) \end{aligned} \tag{3.2}$$

where $f(s|w)$ is the conditional density function of s conditional on w and $\nu(w)$ is the distribution function of w.

The asymptotic properties of our semiparametric maximum frofile likelihood estimations depend on the properties of the nonparametric functions. At point $x_i\alpha(\theta)$ bounded away from the boundary of its support, the function $B_n(x_i, \theta)$ in (2.6) is a consistent estimate of the density function $p(x_i\alpha|\theta)$ and $A_{n,l}(x_i, \theta)$ in (2.5) is a consistent estimate of $E(I_l|x_i\alpha(\theta))p(x_i\alpha|\theta)$. At any boundary point, one side of its neighborhood is empty and the kernel estimates can only estimate some fraction of its density value. The trimming set X_n in assumption 2(4) is designed to bound alway from the boundary of the indices $x\alpha(\theta)$. On $X_n \times \Theta$, nonparametric functions can converge in probability uniformly to their desirable limits.

The estimate $\hat{\theta}$ from (2.7) will be consistent by showing that the sample objective function

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} \ln P_{n,l}(x_i, \theta) \tag{3.3}$$

---

[2] A function $h(x)$ is said to satisfy a Lipschitz condition if there exists a constant c such that $\| h(x_1) - h(x_2) \| \leq c \| x_1 - x_2 \|$ for all x_1 and x_2.

converges in probability uniformly to a continuous limit function which attains a unique maximum at θ_o. The limit function is

$$\begin{aligned} Q^*(\theta) &= E\{\sum_{l=1}^{L} I_l \ln P_l(x\alpha|\theta)\} \\ &= \sum_{l=1}^{L} P_l(x\alpha(\theta_o)|\theta_o) \ln P_l(x\alpha|\theta) \end{aligned} \tag{3.4}$$

which is continuous in θ. This limit function has a unique minimizer if for any $\theta \neq \theta_o$, $P_l(x\alpha|\theta) \neq P_l(x\alpha_o|\theta_o)$ with positive probability for some subset of x, for some l.[3] This is an identification condition. This identification problem had been considered in Ichimura and Lee [1988] (see also Spady and Klein [1987] and Thompson [1989]). Essentially, we need all the indices to be distinguishable from each other.

Theorem 1. *Under assumptions 1 and 2, and* $\lim_{n\to\infty} \frac{n}{\ln n} a_n^m = \infty$, *if, for any* $\theta \neq \theta_o$, $P_l(x\alpha|\theta) \neq P_l(x\alpha_o|\theta_o)$ *with positive probability for some* l, *then the semiparametric maximum profile likelihood estimator* $\hat{\theta}$ *from (2.7) is consistent.*

The consistency of the estimate $\hat{\hat{\theta}}$ from the modified estimation (2.9) can be established by showing that the difference of the the two sample objective functions in (2.7) and (2.9) converges to zero uniformly in probability in θ.

Corollary 1. *Under the conditions in Theorem 1, the modified semiparametric maximum profile likelihood estimator* $\hat{\hat{\theta}}$ *from (2.9) is consistent.*

To derive the asymptotic distributions for our estimates, we need additional regularity conditions. In addition to assumptions 1 and 2, we assume the following conditions hold.

Assumption 3

(1) The functions $\alpha(\theta)$ are twice differentiable and their second order derivatives satisfy a Lipschitz condition.

(2) The kernel function $K(v)$ on R^m is twice differentiable and its second order derivatives satisfy a Lipschitz condition.

Assumption 4

(1) The kernel function $K(v)$ is a high order kernel function with zero moments up to the order s^*, $s^* = m+2$, i.e.,

$$\int_D v_1^{i_1} \cdots v_m^{i_m} K(v)dv = 0$$

for all $0 \leq i_l, l = 1, \cdots, m$ and $1 \leq i_1 + \cdots + i_m < s^*$.

(2) The bandwidth sequence $\{a_n\}$ is chosen such that $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{m+4} = \infty$, $\lim_{n\to\infty} n a_n^{2(m+1)} = \infty$ but $\lim_{n\to\infty} n a_n^{2s^*} = 0$.

Assumption 5

The s^*+1 order derivatives, $\frac{\partial^{s^*+1}}{\partial s^{s^*+1}} f(s|w)$, of $f(s|w)$ are continuous in s and uniformly bounded.

The semiparametric MPLE from (2.7) satisfies the first order condition:

$$\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} \frac{\partial}{\partial\theta} \ln P_{n,l}(x_i, \hat{\theta}) = 0$$

which is well-defined with probability 1 as sample size increases. By Taylor expansion at $\theta = \theta_o$,

[3] For notational simplicity, α_o denotes $\alpha(\theta_o)$.

$$\sqrt{n}(\hat{\theta} - \theta_o)$$

$$= -\{\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\bar{\theta})\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o) \quad (3.5)$$

where $\bar{\theta}$ lies between $\hat{\theta}$ and θ_o. The convergence of the terms on the right hand side of (3.5) will depend on the uniform convergence of the first and second order derivatives of the nonparametric functions. These properties imply that when $\lim_{n\to\infty}\frac{n}{\ln n}a_n^{m+4} = \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\bar{\theta}) \xrightarrow{p} -\sum_{l=1}^{L} E(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta'}). \quad (3.6)$$

The remaining term in (3.5) $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o)$ converges in distribution to $N(0,\Omega)$ where

$$\begin{aligned}\Omega &= \sum_{l=1}^{L} E(P_l(x\alpha_o|\theta_o)\frac{\partial}{\partial\theta}\ln P_l(x\alpha_o|\theta_o)\frac{\partial}{\partial\theta'}\ln P_l(x\alpha_o|\theta_o)) \\ &= \sum_{l=1}^{L} E(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial P_l(x\alpha_o|\theta_o))}{\partial\theta}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta'})\end{aligned} \quad (3.7)$$

Theorem 2

Under assumptions 1-5 and the identification condition in theorem 1,

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

As for the modified semiparametric MPLE $\hat{\hat{\theta}}$ from (2.9), we can show that $\sqrt{n}(\hat{\hat{\theta}} - \theta_o)$ is asymptotically equivalent to $\sqrt{n}(\hat{\theta} - \theta_o)$. This equivalence can be established with the argument in Amemiya [1982] by showing that

$$\sup_{\Theta}|n(Q_n(\theta) - \tilde{Q}_n(\theta))| \xrightarrow{p} 0$$

where

$$\tilde{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}I(A_{n,l}(x_i,\theta) > \Delta_n,\ B_n(x_i,\theta) > \Delta_n)\ln P_{n,l}(x_i,\theta). \quad (3.8)$$

Corollary 2.

Under the conditions in theorem 2,

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

For the general polytomous choice models, our proposed maximum profile likelihood estimators satisfy the characterization of efficient semiparametric estimators in Severini and Wong [1987] in that the maximum profile likelihood estimator has the same limiting distribution of the likelihood estimator derived by using the probability functions $P_l(x\alpha|\theta)$ in the likelihood function. The intuition behind the Severini and Wong's efficient criterion is based on Stone's idea [1956] that any parametrized likelihood estimates will be at least as efficient as semiparametric estimates. The probability functions $P_l(x\alpha|\theta)$ provide a parametrized likelihood for our model. The asymptotic variance of the corresponding likelihood estimate provides an asymptotic lower bound. The asymptotic variances of our semiparametric estimates attain such a bound. Our estimates are therefore asymptotic efficient for the estimation of discrete choice models where the choice probabilities are functions of the indices $x\alpha$. For the binary choice model, the matrix Ω attains also the asymptotic efficiency bound for semiparametric estimation derived in Cosslett[1987] and Chamberlain [1986] as pointed out by Klein and Spady [1987]. The binary choice model considered by Cosslett and Chamberlain assumes that the disturbances in the model are i.i.d. and are independent with the regressors x. This assumption implies that the choice probabilities are monotone functions of the indices $x\alpha$. This model differs from the models in this article in that the choice probabilities in our models need not be monotonic in the indices. The maximum profile likelihood estimate does not impose such restrictions. Therefore, eventhought the binary choice model in Cosslett and Chamberlain has more structure, the maximum profile likelihood estimate for the index probability model is also asymptotic efficient for such model. For general discrete choice models, whether such property will still hold is an unkown question. The recent article by Thompson [1989] points out some basic structural differences between polytomous index probability models and probability models generated by i.i.d. disturbances. However, the asymptotic efficient bound for estimation of such polytomous choice models has not yet been derived.

For these estimators, the covariance matrix Ω can be consistently estimated by

$$-\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'} ln P_{n,l}(x_i,\hat{\hat{\theta}})$$

or

$$\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L}\frac{1}{P_{n,l}(x_i,\hat{\hat{\theta}})}\frac{\partial P_{n,l}(x_i,\hat{\hat{\theta}})}{\partial\theta}\frac{\partial P_{n,l}(x_i,\hat{\hat{\theta}})}{\theta'}.$$

4. Estimation of Sequential Choice Models

The above estimation methods can be extended to the estimation of sequential choice models. Consider a two stage decision problem. The consumer decides in the first stage the choice of the alternatives $1, \cdots, L$ with the choice probabilities (2.3). Conditional on the first stage outcome, a second decision will be made. Without loss of generality, consider the case that, only when $I_1 = 1$, there will be a second stage choice on L_1 alternatives. Since the first and the second stage decision can be correlated, the conditional probabilities are in general functions of indices $x\alpha_1(\theta_o), \cdots, x\alpha_m(\theta_o), x\alpha_{m+1}(\theta_o), \cdots, x\alpha_{\bar{m}}(\theta_o)$:

$$Prob(J_l = 1|I_1 = 1, x) = E(J_l|I_1 = 1, x\alpha_1(\theta_o), \cdots, x\alpha_{\bar{m}}(\theta_o)) \tag{4.1}$$

where J_l is a dichotomous indicator of the alternative $l, l = 1, \cdots, L_1$ in the second stage decision. The additional indices $x\alpha_{m+1}(\theta), \cdots, x\alpha_{\bar{m}}(\theta)$ where $x\alpha_{m+k}(\theta) = s_{m+k} + w\beta_{m+k}(\theta), k = 1, \cdots, \bar{m} - m$ capture the attributes of the alternatives in the second stage decision. To simplify notations, let $x\delta(\theta) = (x\alpha_1(\theta), \cdots, x\alpha_{\bar{m}}(\theta))$ and $\bar{P}_l(x_i\delta|\theta) = E(J_l|I_1 = 1, x\delta(\theta))$ be the conditional choice probability of alternative l in the second stage choice conditional on $I_1 = 1$ and the indices $x\delta(\theta)$.

Similar to equation(2.4), for each $\theta \in \Theta$ the probability function $\bar{P}_l(x_i\delta|\theta)$ can be estimated by

$$\bar{P}_{n,l}(x_i, \theta) = \frac{\bar{A}_{n,l}(x_i, \theta)}{\bar{B}_n(x_i, \theta)} \tag{4.2}$$

where

$$\bar{A}_{n,l}(x_i, \theta) = \frac{1}{(n-1)b_n^{\bar{m}}} \sum_{j \neq i}^{n} I_{1j} I_{lj} \bar{K}\left(\frac{x_i\delta(\theta) - x_j\delta(\theta)}{b_n}\right) \tag{4.3}$$

and

$$\bar{B}_n(x_i, \theta) = \frac{1}{(n-1)b_n^{\bar{m}}} \sum_{j \neq i}^{n} I_{1j} \bar{K}\left(\frac{x_i\delta(\theta) - x_j\delta(\theta)}{b_n}\right) \tag{4.4}$$

where $\bar{K}(\cdot)$ is a kernel function on $R^{\bar{m}}$ and $b_n > 0$ is a bandwidth sequence. The semiparametric MPLE which is analogous to (2.7) will be

$$\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} [I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} \ln P_{n,l}(x_i, \theta) + I_{\bar{X}_n}(x_i) I_{1i} \sum_{l=1}^{L_1} J_{li} \ln \bar{P}_{n,l}(x_i, \theta)]. \tag{4.5}$$

A modified method which is analogous to (2.9) will be

$$\begin{aligned}\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} [I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} I(A_{n,l}(x_i, \theta) > \Delta_n, B_n(x_i, \theta) > \Delta_n) \ln P_{n,l}(x_i, \theta) \\ + I_{\bar{X}_n}(x_i) I_{1i} \sum_{l=1}^{L_1} J_{li} I(\bar{A}_{n,l}(x_i, \theta) > \Delta_n, \bar{B}_n(x_i, \theta) > \Delta_n) \ln \bar{P}_{n,l}(x_i, \theta)].\end{aligned} \tag{4.6}$$

To justify these estimation methods, in addition to the previous assumptions, the following conditions are assumed:

Assumption 6.

(1) $\delta(\theta)$ is continuous and satisfies a Lipschitz condition on Θ.

(2) The support $\bar{S}_x$ of $x, x = (s_1, \cdots, s_{\bar{m}}, w)$, is $[\bar{T}_1, \bar{T}_2] \times S_w$ where $[\bar{T}_1, \bar{T}_2]$ is a $\bar{m}$ dimensional compact rectangle. For each w in S_w, the conditional density $\bar{f}(\bar{s}|w)$ where $\bar{s} = (s_1, \cdots, s_{\bar{m}})$ conditional on w is continuous in $\bar{s}$, strictly positive and uniformly bounded on $\bar{S}_x$.

(3) The probabilities $E(J_l|I_1 = 1, x\delta(\theta_o)), l = 1, \cdots, L_1$ are strictly positive and continuous on $\bar{S}_x$.

Assumption 7.

(1) The kernel function $\bar{K}(\cdot)$ on $R^{\bar{m}}$ is a function with bounded support $\bar{D}$ such that $\int_{\bar{D}} \bar{K}(v)dv = 1$ and $\int_{\bar{D}} |\bar{K}(v)|dv < \infty$.

(2) The kernel function $\bar{K}(\cdot)$ satisfies a Lipschitz condition and it goes to zero at the boundary of $\bar{D}$.
(3) The bandwidth sequence b_n is a positive sequence such that $\lim_{n\to\infty} b_n = 0$.
(4) $\bar{X}_n = \{x \in \bar{S}_x : \bar{T}_{1,l} + \bar{\delta}_n \leq s_l \leq \bar{T}_{2,l} - \bar{\delta}_n, l = 1, \cdots, \bar{m}\}$ where $\{\bar{\delta}_n\}$ is a positive sequence such that $\lim_{n\to\infty} \frac{\bar{\delta}_n}{b_n} = \infty$.

Assumption 8.
(1) The functions $\delta(\theta)$ are twice differentiable and their second order derivatives satisfy a Lipschitz condition.
(2) The kernel function function $\bar{K}(\cdot)$ on $R^{\bar{m}}$ is twice differentiable and its second order derivatives satisfy a Lipschitz condition.

Assumption 9.
(1) The kernel $\bar{K}(\cdot)$ is a high order kernel function with zero moments up to the order $\bar{s}, \bar{s} = \bar{m} + 2$.
(2) The bandwidth sequence $\{b_n\}$ is chosen with a rate such that $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{\bar{m}+4} = \infty, \lim_{n\to\infty} n b_n^{2(\bar{m}+1)} = \infty$ but $\lim_{n\to\infty} n b_n^{2\bar{s}} = 0$.

Assumption 10.
The $\bar{s}+1$ order derivatives, $\frac{\partial^{\bar{s}+1}}{\partial s^{\bar{s}+1}} \bar{f}(\bar{s}|w)$, are continuous in $\bar{s}$ and uniformly bounded.

These assumptions are parallel to the the assumptions 1)-5). They are needed owning to the additional indices and choice probabilities in the second stage decision. The asymptotic analysis of the properties of these MPLE for the sequential choice model will be similar to those for the polytomous choice models.

Theorem 3.
Under assumptions 1,2,6 and 7 and the conditions that $\lim_{n\to\infty} \frac{n}{\ln n} a_n^m = \infty$ *and* $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{\bar{m}} = \infty$, *if for any* $\theta \neq \theta_o, P_l(x\alpha|\theta) \neq P_l(x\alpha_o|\theta_o)$ *or* $\bar{P}_j(x\delta|\theta) \neq \bar{P}_j(x\delta_o|\theta_o)$ *with positive probability for some* l *or* j, *then both the semiparametric MPLE from (5.5) and (5.6) are consistent.*

Theorem 4.
Under assumptions 1-10 and the identification condition in theorem 3,

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

where $\hat{\theta}$ *is the MPLE from (4.5) or (4.6), and*

$$\begin{aligned} &\Sigma \\ &= \sum_{l=1}^{L} E\Big(\frac{1}{P_l(x\alpha_o|\theta_o)} \frac{\partial}{\partial\theta} P_l(x\alpha_o|\theta_o) \cdot \frac{\partial}{\partial\theta'} P_l(x\alpha_o|\theta_o)\Big) + \sum_{l=1}^{L_1} E\Big(\frac{P_1(x\alpha_o|\theta_o)}{\bar{P}_l(x\delta_o|\theta_o)} \frac{\partial}{\partial\theta} \bar{P}_l(x\delta_o|\theta_o) \cdot \frac{\partial}{\partial\theta'} \bar{P}_l(x\delta_o|\theta_o)\Big). \end{aligned} \tag{4.7}$$

5. Some Monte Carlo Simulations

To investigate the finite sample properties of the MPLE for polytomous choice models, we perform some small scale Monte Carlo simulations. Simulated data will be generated for two discrete choice models. Both are models with three choice alternatives but their disturbances are generated from different distributions. The utility associated with alternative l is

$$U_l = x_l + z_l\beta + w_l\gamma + u_l, \quad l = 1,2,3. \tag{5.1}$$

where x_1, x_2 and x_3 are generated by independently distributed truncated normal $N(0,1)$ random variables with support on [-1.8, 1.8]; z_1, z_2 and z_3 are generated by independent uniform variables with support on [-1, 1], and w_1, w_2 and w_3 are generated by independent Poisson variables with mean 2 but are truncated with support on [0, 5]. Such explanatory variables imply that the density functions of the corresponding indices are bounded alway from zero on their supports. The true parameters are $\beta_o = 1$ and $\gamma_o = -1$. In the first model, the disturbances u_l, $l = 1,2,3$ are generated by independent Gumbel type I extreme value distributions (EV D.) which imply the logit choice probabilities. In the second model, the disturbances u_l, $l = 1,2,3$ are generated by independent $N(0,1)$ variables. These two models will be estimated by our maximum profile likelihood procedure and by a logit likelihood subroutine (logit MLE). For the first model, the logit MLE is a parametrically asymptotically efficient estimator but, for the second model, it will be inconsistent. These designs allow us to investigate the efficient loss for the MPLE as compared with the logit MLE in the first model and the robust gain in the second model.

For the discrete choice models with three alternatives, the number of indices is $m = 2$. The kernel function on R^2 chosen for the estimation of such models is the product of two univariate kernel functions of which each has the form:

$$K_4(t) = 2K(t) - \frac{1}{\sqrt{2}}K(\frac{t}{\sqrt{2}}) \tag{5.2}$$

where

$$K(t) = \begin{cases} \frac{35}{32}(1-t^2)^3, & \text{if } |t| < 1 \\ 0 & \text{, otherwise.} \end{cases}$$

The kernel function $K(t)$ is a proper density function. This density function is computationally simple and is twice continuously differentiable on the real line with bounded third order derivative. The kernel function $K_4(t)$ is a high order kernel with its first three moments being zero. The way of constructing such high order kernel functions from density functions is suggested in Bierens[1987]. The bandwith sequence $\{a_n\}$ is chosen as

$$a_n = \frac{c_1}{n^{1/7}} \tag{5.3}$$

where c_1 is a constant proportional factor independent with sample size n. In the Monte Carlo study, various values for c_1 will be used to check the sensitivity of the MPLE with the bandwidth parameter. The sequence $\{\delta_n\}$ is chosen as

$$\delta_n = c_2 a_n^{9/10} \tag{5.4}$$

where c_2 is another constant factor. The regressors are trimmed whenver $|x_1 - x_3| > 3.6 - \delta_n$ or $|x_2 - x_3| > 3.6 - \delta_n$. The constant factor c_2 in the δ is set to 0.3 so that the regressors are not excessively trimmed.[4] The chosen kernel and the bandwidth sequences satisfy the regularity conditions in the assumptions 2,3 and 4.

To avoid the potential empty space phenomenon, i.e., empty neighborhoods at some points in the kernel density estimation, and difficulty with negative probability values for the log likelihood function, the estimated probabilities $P_{n,l}(x_i,\theta)$ are modified by the following rules:

1. If $B_n(x_i,\theta) = 0$ and $A_{n,l}(x_i,\theta) = 0$, set $P_{n,l}(x_i,\theta) = 1/L$;

[4] Experiments with larger δ do not seem to change much the estimates.

2. if $B_n(x_i, \theta) = 0$, but $A_{n,l}(x_i, \theta) \neq 0$, set $B_n(x_i, \theta) = 10^{-10}$;
3. if $P_{n,l}(x_i, \theta) \leq 10^{-10}$, set $P_{n,l}(x_i, \theta) = 10^{-10}$; and
4. if $P_{n,l}(x_i, \theta) \geq 1 - 10^{-10}$, set $P_{n,l}(x_i, \theta) = 1 - 10^{-10}$.

These modifications guarantee the computed probabilities are properly bounded between zero and one. As we have argued in the theoretical sections, such modification will not change the asymptotic distribution of the MPLE. Such modifications are intuitively appealing and are adopted throughout the simulations. The log profile likelihood functions are maximized with the subroutine of downhill simplex method in Press et al [1986] with the logit estimates as one of the starting values.[5]

The results of the simulations are reported in four tables. In Table 1, we report results for the MPLE and the logit MLE estimates derived with various sample sizes (N) and a couple of bandwidth factors c_1.[6]Samples with sizes 100, 300 and 500 are generated. Throughtout the study, 200 simulations are performed for each case. The summary statistics, namely, mean, standard deviation (SD) and root mean squared error (RMSE), are reported. The bandwidth factors $c_1 = 8$ and $c_1 = 10$ provide compatible results. The biases of the MPLE with $c_1 = 10$ are on average smaller then the biases with $c_1 = 8$ for the model with EV distrurbances. But the reverse is observed for the model with normal disturbances. The biases of the MPLE with these factors are not large. The variances of the MPLE decrease as sample size increases with one exception. However the rate of decreasing is slower than the rate of the logit MLE. These sample sizes are probably not large enough for the asymptotic results to hold. Comparing the standard deviations of the MPLE with the ones of the logit MLE when the logit model is correctly specified, the loss of efficiency is about 20 to 40 percents for sample size 100. The loss of efficiency increases to about 50 percent for sample size 300 and about 63 percent for sample size 500. The RMSE ratio is on average about 1.5 to 1 for sample size 100. They are about 2 to 1 for sample size 300 and 2.9 to 1 for sample size 500.

For the second model with normal disturbances, the logit likelihood function is misspecified and the logit MLE is inconsistent. The biases of the logit MLE are much larger than the biases of the MPLE for both factors 8 and 10 but their variances are smaller. Overall, the MPLE have smaller RMSE than the logit MLE.

To investigate more the sensitivity of the MPLE with the bandwidth parameter, estimates with bandwidth factors 1, 5 and 15 in addition to the factors 8 and 10 are also computed. The summary statistics of the results, namely, mean, SD, RMSE, median, upper quartile (LQ) and lower quartile (LQ), are reported in Table 2. The sample size in these simulations are fixed at 300. The estimates are sensitive to the choice of the bandwidth factors. The mean, median, LQ and UQ of the estimates tend to increase as the bandwidth becomes wider. Among them, the estimates with the factors 8 and 10 have the smallest biases. The variances and the RMSE are the largest for estimates with the largest or the smallest bandwidth factors. The estimates with factors 8 and 10 have the smallest RMSE. The sensitivity of MPLE with the bandwidth factors in these models raises issues on how to choose the bandwidth factors in practice. In the statistical literature for the estimation of density or regression functions, there are suggestions such as the use of cross validation (Silverman [1986]) procedure to choose the bandwidth parameter. However, such procedure is designded for density estimation and whether it is useful for the estimation of coefficients in semiparametric models is questionable. For kernel density estimation, small bandwidth will introduce large variance but small bias and large bandwidth will have smaller variance but larger bias. For the estimation of the regression coefficients in the choice model, the evidence from Table 2 reveals different patterns. A simple procedure that seems to be useful for our model is based on the goodness-of-fit criterion. For each estimation, the average maximized log likelihood values are computed by deriving the maximized log likelhood function by the number of untrimmed observations. The column with the label AKL in Table 2 reports the mean of the average maximized log likelihood funtions for each factor. It is interesting to see that the models with the factors 8 and 10 provide the better fitted models. In the last row, we report the summary statistics for the estimator which corresponds to the estimate with the largest average log likelihood value among the five different bandwidth factors. This estimtor will have the same asymptotic properties as the MPLE with a fixed bandwidth seqence. The empirical evidence seems encouraging. The biases are small. The variances

---

[5] The sum of the vector (1, 0) and the logit estimator, and the sum of the vector (0, 1) and the logit estimator provide the other two starting points.

[6] The logit ML subroutine estimates the coefficient of the regressor x in addition to the coefficients of z and w. For comparison, we report only the coefficients of z and w in the tables.

and the RMSE are only slightly larger than the best ones with the fixed bandwidths. Comparing them with the logit MLE in Table 1 with the EV distribution, the RMSE ratios are on averge 2.7 to 1. For the misspecified model, the RMSE for this semiparametric estimator is slightly smaller than the RMSE of the logit MLE.

To compare the performance of the MPLE for choice models with different choice alternatives, we have also estimated some dichotomous choice models. The dichotomous models are generated by the same random utility stucture as in (5.1) except that the third choice is deleted. The disturbances in the utility functions are generated by EV distributions. For the dichtomous choice model, the number of index is $m = 1$. The bandwidth sequence is chosen as

$$a_n = \frac{c_1}{n^{1/5.5}}.$$

The kernel function is $K_4(t)$ in (5.2) and the δ_n sequence is the same in (5.4). The results are reported in Table 3 and Table 4. A striking difference of these results with the results in Table 1 and Table 2 is that the MPLE for the dichotomous choice model are less sensitive with the bandwidth parameter. The estimates for the models with bandwidth factor 10 have the smallest RMSE. However, the biases are still small even with bandwidth factor 2.5. The other features of these results are similar to those of the polytomous choice case. The mean,median, LQ and UQ tend to increase with the bandwidth factors. The larger variances and RMSE of the estimates are associated with the large factors as well as the small factors. The minimum average log likelihood criterion provides good semiparametric estimates.

TABLE 1.

Polytomous Choice Model with 3 alternatives

N		100			300			500		
EV D.		Mean	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE
factor	β_1	.991	.413	.411	.919	.375	.383	.907	.362	.373
8	β_2	-.863	.235	.271	-.856	.217	.260	-.850	.195	.246
factor	β_1	1.167	.547	.569	1.077	.385	.392	1.105	.373	.387
10	β_2	-1.038	.367	.367	-1.018	.208	.208	-1.016	.203	.203
logit	β_1	1.045	.340	.341	1.027	.182	.184	1.022	.136	.138
MLE	β_2	-1.042	.207	.210	-1.023	.111	.113	-1.010	.074	.075
Normal D.		Mean	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE
factor	β_1	1.053	.477	.478	1.061	.379	.383	1.036	.321	.323
8	β_2	-.963	.222	.224	-.980	.201	.202	-.995	.190	.190
factor	β_1	1.277	.490	.561	1.170	.338	.378	1.227	.347	.414
10	β_2	-1.192	.298	.353	-1.132	.207	.245	-1.143	.186	.234
logit	β_1	1.382	.387	.542	1.326	.215	.390	1.319	.161	.357
MLE	β_2	-1.358	.218	.419	-1.326	.118	.347	-1.321	.089	.333

TABLE 2.
Results of Various Bandwidth Factors
Polytomous choice with 3 alternatives
Design: sample size 300

				EV D.				
factor		Mean	SD	RMSE	Median	LQ	UQ	ALK
1	β_1	.611	.460	.602	.564	.311	.822	-3.433
	β_2	-.464	.289	.609	-.416	-.602	-.279	
5	β_1	.693	.360	.473	.649	.442	.907	-.819
	β_2	-.609	.224	.450	-.589	-.772	-.452	
8	β_1	.919	.375	.383	.868	.690	1.110	-.749
	β_2	-.856	.217	.260	-.843	-.971	-.726	
10	β_1	1.077	.385	.392	1.010	.832	1.320	-.742
	β_2	-1.018	.208	.208	-1.020	-1.140	-.897	
15	β_1	1.508	.461	.685	1.470	1.230	1.775	-.753
	β_2	-1.510	.282	.583	-1.490	-1.670	-1.325	
min	β_1	1.055	.423	.426	.962	.764	1.265	-.728
	β_2	-1.042	.343	.345	-.979	-1.195	-.805	

				Normal D.				
factor		Mean	SD	RMSE	Median	LQ	UQ	ALK
1	β_1	.714	.501	.576	.648	.411	.910	-2.955
	β_2	-.608	.353	.527	-.516	-.732	-.411	
5	β_1	.773	.341	.409	.775	.565	.969	-.691
	β_2	-.705	.199	.356	-.699	-.827	-.551	
8	β_1	1.061	.379	.383	.999	.817	1.240	-.635
	β_2	-.980	.201	.202	-.961	-1.085	-.854	
10	β_1	1.170	.338	.378	1.140	.953	1.370	-.635
	β_2	-1.132	.207	.245	-1.130	-1.235	-1.025	
15	β_1	1.697	.372	.790	1.700	1.465	1.910	-.657
	β_2	-1.684	.269	.735	-1.680	-1.790	-1.530	
min	β_1	1.086	.364	.373	1.030	.822	1.310	-.619
	β_2	-1.073	.311	.319	-1.015	-1.205	-.891	

TABLE 3.

Dichotomous Choice Models

N		100			300			500		
EV distribution		Mean	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE
factor	β_1	1.093	.720	.722	1.062	.398	.402	1.076	.364	.371
5	β_2	-1.037	.642	.640	-1.039	.278	.280	-1.022	.321	.321
factor	β_1	1.151	.709	.721	1.101	.349	.363	1.128	.326	.326
10	β_2	-1.148	.565	.581	-1.091	.245	.261	-1.085	.241	.255

TABLE 4.

Results of Various Bandwidth Factors
dichotomous choice models

Design: sample size 300

		EV distribution						
factor		Mean	SD	RMSE	Median	LQ	UQ	ALK
1	β_1	1.034	.495	.495	1.020	.666	1.365	-.505
	β_2	-.759	.297	.382	-.771	-.937	-.546	
2.5	β_1	1.020	.397	.397	.985	.739	1.270	-.449
	β_2	-.990	.292	.292	-.943	-1.170	-.789	
5	β_1	1.062	.398	.402	1.010	.784	1.305	-.440
	β_2	-1.039	.278	.280	-.993	-1.150	-.865	
10	β_1	1.101	.349	.363	1.070	.861	1.285	-.444
	β_2	-1.091	.245	.261	-1.070	-1.215	-.913	
15	β_1	1.322	.352	.477	1.320	1.090	1.525	-.471
	β_2	-1.312	.288	.424	-1.310	-1.460	-1.130	
20	β_1	1.723	.473	.864	1.720	1.450	1.995	-.524
	β_2	-1.683	.380	.781	-1.685	-1.895	-1.470	
min	β_1	1.011	.354	.354	.993	.765	1.210	-.434
	β_2	-1.024	.260	.261	-.985	-1.150	-.854	

6. **Proof of Consistency**

The asymptotic properties of our semiparametric maximum frofile likelihood estimations depend on the uniform convergence properties of the nonparametric functions. These follow from the following two propositions.

Proposition 1. (A Uniform Law of Large Numbers) *Let $\{y_i\}$ be a sequence of i.i.d. random vectors. Suppose that the measurable function $g(y,\beta,a_n)$ can be represented by the form:*

$$g(y,\beta,a_n) = \frac{1}{a_n^d} h(y,\beta,s(y,\beta)/a_n)$$

where $a_n = O(\frac{1}{n^p})$ with $p > 0, \beta \in B, s(y,\beta)$ is a m dimensional vector value function and $d \geq m$. Suppose that the following conditions are satisfied:
(i) B is a compact subset of a finite dimensional Euclidean space.
(ii) $h(y,\beta,s)$ is a bounded function which satisfies a Lipschitz condition with respect to β and s. The function $s(y,\beta)$ satisfies also a Lipschitz condition with β.
(iii) $E(h^2(y,\beta,s(y,\beta)/a_n) = O(a_n^m)$ uniformly in $\beta \in B$.
If $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{2d-m} = \infty$, then $\frac{1}{n}\sum_{i=1}^n g(y_i,\beta,a_n) - E(g(y,\beta,a_n)) \xrightarrow{p} 0$, uniformly in $\beta \in B$.
In particular, when $d = m$, for any $\epsilon > 0$, there exists constants c_0, c_1 and c_2 with $c_0 > 0, c_1 > 0$ such that

$$P(\sup_{\beta\in B} |\frac{1}{n}\sum_{i=1}^n g(y_i,\beta,a_n) - E(g(y,\beta,a_n))| \geq \epsilon) \leq c_0 \exp(-(c_1 n a_n^m + c_2 \ln n)).$$

Proof: This is a special case of a uniform law numbers in Ichimura[1987] (see also Ichimura and Lee [1988]).

Proposition 2.*Let $K(v)$ be a kernel function on R^m with a bounded support D. For each $\theta \in \Theta, T_2(\theta)$ and $T_1(\theta)$ denote the upper and lower limits of the values of the continuous random vector $t(z,\theta)$ in R^m. Suppose that the density function $g(t|\theta)$ of $t(z,\theta)$ and the conditional expectation $E(c(z,z_i,\theta)|t,z_i)$ are bounded and uniformly continuous in t, uniformly in (θ,z_i). Then*

$$\sup_{(z_i,\theta)\in Z_n} |E(c(z,z_i,\theta)\frac{1}{a_n^m}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i) - E(c(z,z_i,\theta)|t(z_i,\theta),z_i)g(t(z_i,\theta)|\theta)| \longrightarrow 0$$

where $Z_n = \{(z,\theta)|T_1(\theta)+\delta_n \leq t(z,\theta) \leq T_2(\theta)-\delta_n\}$ with $\delta_n > 0$ and $\lim_{n\to\infty}\frac{\delta_n}{a_n} = \infty$. Furthermore, if $K(v)$ is a kernel 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 $i_j \geq 0, j = 1,\ldots,m; i_1+\cdots+i_m < s^*$, and the functions $g(t|\theta)$ and $E(c(z,z_i,\theta)|t,z_i)$ are differentiable in t to the order s^* and these derivatives are uniformly bounded, then

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

Proof: This result is abstracted from Ichimura [1987] and Ichimura and Lee [1988] with special emphasis on the trimming sequence Z_n. For convenient reference, a proof is provided in the appendix.

These two theorems are strong enough to imply the uniform convergence of $A_{n,l}(x_i,\theta)$ and $B_n(x_i,\theta)$ on $X_n \times \Theta$. The variances of $A_{n,l}(x_i,\theta)$ and $B_n(x_i,\theta)$ have the familar order $O(\frac{1}{na_n^m})$ uniformly on $S_x \times \Theta$, i.e.,

$$\sup_{S_x\times\Theta} var(A_{n,l}(x_i,\theta)|x_i) = O(\frac{1}{na_n^m}) \tag{6.1}$$

and

$$\sup_{S_x\times\Theta} var(B_n(x_i,\theta)|x_i) = O(\frac{1}{na_n^m}). \tag{6.2}$$

Proposition 1 implies that if $\lim_{n\to\infty} \frac{n}{\ln n} a_n^m = \infty$,

$$\sup_{S_x \times \Theta} |A_{n,l}(x_i, \theta) - E(A_{n,l}(x_i, \theta)|x_i)| \xrightarrow{p} 0 \tag{6.3}$$

and

$$\sup_{S_x \times \Theta} |B_n(x_i, \theta) - E(B_n(x_i, \theta)|x_i)| \xrightarrow{p} 0 \tag{6.4}$$

Proposition 2 guarantees that

$$\sup_{X_n \times \Theta} |E(A_{n,l}(x_i, \theta)|x_i) - E(I_l|x_i\alpha(\theta))p(x_i\alpha|\theta)| \rightarrow 0 \tag{6.5}$$

and

$$\sup_{X_n \times \Theta} |E(B_n(x_i, \theta)|x_i) - p(x_i\alpha|\theta)| \rightarrow 0. \tag{6.6}$$

Since $p(x_i\alpha|\theta)$ and $E(I_l|x_i\alpha(\theta))$ are bounded away from zero on $S_x \times \Theta$, uniform convergence implies that $\inf_{X_n \times \Theta} B_n(x_i, \theta)$ and $\inf_{X_n \times \Theta} A_{n,l}(x_i, \theta)$ are bounded away from zero in probability. The nonparametric probability function $P_{n,l}(x_i, \theta)$ as a ratio of $A_{n,l}(x_i, \theta)$ over $B_n(x_i, \theta)$ will therefore converge in probability to $E(I_l|x_i\alpha(\theta))$ uniformly in $(x_i, \theta) \in X_n \times \Theta$. These arguments are from Ichimura and Lee[1988]. In the previous article, the trimming sets are fixed and are not depended on sample size. For the approach here, the trimming sets are expanding as sample size increases, so special attention has been taken in its construction (assumption 2(4)) to overcome the troublesome caused at the boundary of the support of the indices. In summary, if $\lim_{n\to\infty} \frac{n}{\ln n} a_n^m = \infty$,

$$\sup_{X_n \times \Theta} |B_n(x_i, \theta) - p(x_i\alpha|\theta)| \xrightarrow{p} 0 \tag{6.7}$$

$$\sup_{X_n \times \Theta} |A_{n,l}(x_i, \theta) - E(I_l|x_i\alpha)p(x_i\alpha|\theta)| \xrightarrow{p} 0 \tag{6.8}$$

$$\sup_{X_n \times \Theta} |P_{n,l}(x_i, \theta) - E(I_l|x_i\alpha)| \xrightarrow{p} 0. \tag{6.9}$$

and there exists a positive constant $\delta > 0$ such that

$$\lim_{n\to\infty} P(\inf_{X_n \times \Theta} B_n(x_i, \theta) \geq \delta) = 1 \tag{6.10}$$

and

$$\lim_{n\to\infty} P(\inf_{X_n \times \Theta} A_{n,l}(x_i, \theta) \geq \delta) = 1. \tag{6.11}$$

These last two properties guarantee that the log likelihood function in (2.7) will be well defined with probability 1 as sample size increases.

Define the functions

$$\begin{aligned} Q^*(\theta) &= E\{\sum_{l=1}^{L} I_l \ln P_l(x\alpha|\theta)\} \\ &= \sum_{l=1}^{L} P_l(x\alpha(\theta_o)|\theta_o) \ln P_l(x\alpha|\theta) \end{aligned} \tag{6.12}$$

and

$$Q_n^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li} \ln P_l(x_i\alpha|\theta). \tag{6.13}$$

The classical uniform law of large number (e.g.,Amemiya[1985],theorem 4.2.1) implies that $Q_n^*(\theta) \xrightarrow{p} Q^*(\theta)$ uniformly in θ. On the other hand,

$$Q_n(\theta) - Q_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li}[\ln P_{n,l}(x_i,\theta) - \ln P_l(x_i\alpha|\theta)]$$

implies

$$\begin{aligned} &\sup_{\Theta} |Q_n(\theta) - Q_n^*(\theta)| \\ &\leq \sum_{l=1}^{L} \sup_{X_n\times\Theta} |\ln P_{n,l}(x_i,\theta) - \ln P_l(x_i\alpha|\theta)| \\ &\xrightarrow{p} 0 \end{aligned}$$

as $P_{n,l}(x_i,\theta)$ converges in probability to $P_l(x_i\alpha|\theta)$ uniformly on $X_n \times \Theta$, and $P_{n,l}(x_i,\theta)$ and $P_l(x_i\alpha|\theta)$ are bounded away from zero in probability. Jensen's inequality implies θ_o maximizes $Q^*(\theta)$. It will be the unique minimizer if for any $\theta \neq \theta_o$, $P_l(x\alpha|\theta) \neq P_l(x\alpha_o|\theta_o)$ with positive probability for some subset of x, for some l. This is an identification condition. Under this identification condition, the semiparametric maximum profile likelihood estimator $\hat{\theta}$ from (2.7) is consistent.

With (3.3) and (3.8),

$$\begin{aligned} &Q_n(\theta) - \bar{Q}_n(\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i) \sum_{l=1}^{L} I_{li}\{1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)\} \ln P_{n,l}(x_i,\theta). \end{aligned}$$

which implies

$$\begin{aligned} &\sup_{\Theta} |Q_n(\theta) - \bar{Q}_n(\theta)| \\ &\leq \sum_{l=1}^{L} \sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n,\ B_n(x_i,\theta) > \Delta_n)| \cdot \sup_{X_n\times\Theta} |\ln P_{n,l}(x_i,\theta)|\,. \end{aligned} \tag{6.14}$$

Since $P_{n,l}(x_i,\theta)$ is bounded alway from 0 on $X_n\times\Theta$ in probability, $\sup_{X_n\times\Theta} |\ln P_{n,l}(x_i,\theta)| = O_p(1)$. Uniform convergence of $Q_n(\theta) - \bar{Q}_n(\theta)$ to zero in probability will follow if

$$\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)| \xrightarrow{p} 0.$$

For any $\epsilon > 0$, by Markov inequality,

$$\begin{aligned} &P(\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n,\ B_n(x_i,\theta) > \Delta_n)| \geq \epsilon) \\ &\leq \frac{1}{\epsilon}E(\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n,\ B_n(x_i,\theta) > \Delta_n)|). \end{aligned}$$

Since $X_n \times \Theta$ is compact, $\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x,\theta) > \Delta_n,\ B_n(x,\theta) > \Delta_n)| = 1$ if and only if there exists $(\bar{x}_n,\ \bar{\theta}_n) \in X_n \times \Theta$ such that $A_{n,l}(\bar{x}_n,\ \bar{\theta}_n) \leq \Delta_n$ or $B_n(\bar{x}_n,\ \bar{\theta}_n) \leq \Delta_n$. Therefore,

$$\begin{aligned} &E(\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)|) \\ &= P(A_{n,l}(\bar{x}_n,\bar{\theta}_n) \leq \Delta_n, \text{or } B_n(\bar{x}_n,\bar{\theta}_n) \leq \Delta_n, \text{for some } (\bar{x}_n,\bar{\theta}_n) \in X_n \times \Theta) \\ &\leq P(\inf_{X_n\times\Theta} B_n(x,\theta) \leq \Delta_n) + P(\inf_{X_n\times\Theta} A_{n,l}(x,\theta) \leq \Delta_n) \end{aligned}$$

and

$$\begin{aligned} &P(\sup_{X_n\times\Theta} |1 - I(A_{n,l}(x,\theta) > \Delta_n,\ B_n(x,\theta) > \Delta_n)| \geq \epsilon) \\ &\leq \frac{1}{\epsilon}[P(\inf_{X_n\times\Theta} B_n(x,\theta) \leq \Delta_n) + P(\inf_{X_n\times\Theta} A_{n,l}(x,\theta) \leq \Delta_n)] \end{aligned} \tag{6.15}$$

which converge to zero as n goes to infinity, since Δ_n tends to zero and both $B_n(x,\theta)$ and $A_{n,l}(x,\theta)$ are bounded away form zero on $X_n \times \Theta$ in probability. The consistency of $\hat{\hat{\theta}}$ follows as $Q_n(\theta)$ converges to $Q^*(\theta)$ in probability uniformly on Θ.

7. Proof of Asymptotic Distribution

The convergence of the terms on the right hand side of (3.5) will depend on the convergence of the first and second order derivatives of the nonparametric functions. The following two propositions are useful.

Proposition 3

Let $K(v)$ be a kernel function on R^m with a bounded support D such that $K(v)$ goes to zero at the boundary of D and its gradient $\frac{\partial}{\partial v}K(v)$ is bounded. Suppose that the density function $g(t|\theta)$ of $t(z,\theta)$, its gradient $\frac{\partial}{\partial t}g(t|\theta)$, the conditional expectation $E(c(z,z_i,\theta)|t,z_i)$ and its derivative $\frac{\partial}{\partial t}E(c(z,z_i,\theta)|t,z_i)$ are all uniformly continuous in t, uniformly in (z_i,θ) and are uniformly bounded. Then

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

where Z_n is defined in proposition 2.

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

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

Proposition 4

Let $K(v)$ be a twice differentialbe kernel function on R^m with a bounded support D such that $K(v)$ and its gradient $\frac{\partial}{\partial v}K(v)$ go to zero at the boundary of D, and the gradient $\frac{\partial}{\partial v}K(v)$ and its hessian matrix $\frac{\partial^2}{\partial v\partial v'}K(v)$ are bounded. Suppose that the density function $g(t|\theta)$ of $t(z,\theta)$, its derivatives $\frac{\partial}{\partial t}E(c(z,z_i,\theta)|t)$ and $\frac{\partial^2}{\partial t\partial t'}E(c(z,z_i,\theta)|t)$ are uniformly continuous in t, uniformaly in (z_i,θ), and are uniformly bounded. Then

$$\lim_{n\to\infty}\sup_{(z_i,\theta)\in Z_n}|E(c(z,z_i,\theta)\frac{1}{a_n^{m+2}}\frac{\partial^2}{\partial v\partial v'}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}|z_i) - [g(t(z_i,\theta)|\theta)\frac{\partial^2}{\partial v\partial v'}E(c(z,z_i,\theta)|t(z_i,\theta))+\frac{\partial}{\partial v}g(t(z_i,\theta)|\theta)\frac{\partial}{\partial v'}E(c(z,z_i,\theta)|t(z_i,\theta)) + \frac{\partial}{\partial v}E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial}{\partial v'}g(t(z_i,\theta)|\theta)+E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial^2}{\partial v\partial v'}g(t(z_i,\theta)|\theta)]| = 0$$

These two propositions are again abstracted from the results in Ichimura[1987] and Ichimura and Lee[1988]. For reference the proofs are provided in the appendix.

The first order derivatives of $A_{n,l}(x_i,\theta)$ and $B_n(x_i,\theta)$ with respect to θ_k are

$$\frac{\partial A_{n,l}(x_i,\theta)}{\partial\theta_k} = \frac{1}{(n-1)a_n^{m+1}}\sum_{j\neq i}^{n} I_{lj}(x_i\frac{\partial\alpha}{\partial\theta_k}-x_j\frac{\partial\alpha}{\partial\theta_k})\frac{\partial}{\partial v}K(\frac{x_i\alpha-x_j\alpha}{a_n}) \tag{7.1}$$

and

$$\frac{\partial B_n(x_i,\theta)}{\partial\theta_k} = \frac{1}{(n-1)a_n^{m+1}}\sum_{j\neq i}^{n}(x_i\frac{\partial\alpha}{\partial\theta_k} - x_j\frac{\partial\alpha}{\partial\theta_k})\frac{\partial}{\partial v}K(\frac{x_i\alpha - x_j\alpha}{a_n}). \tag{7.2}$$

The variances of $\frac{\partial A_{n,l}(x_i,\theta)}{\partial\theta}$ and $\frac{\partial B_n(x_i)}{\partial\theta}$ have the familiar order $O(\frac{1}{na_n^{m+2}})$ uniformly in $(x_i,\theta) \in S_x \times \Theta$. It is apparent from (3.1) and assumption 5 that $p(t|\theta)$ will be continuously differentiable in t up to order s^*+1 and these derivatives are uniformly bounded. For any integrable function $h(x)$,

$$E(h(x)|t) = \int h(t-wB,w)f(t-wB|w)d\nu(w)/p(t|\theta). \tag{7.3}$$

So assumption 5 implies that if $h(x)$ is continuously differentiable to order s^*+1 and these derivatives are bounded, the expectation $E(h(x)|t)$ will also have the same differentiability and boundedness properties. These properties justify the conditions in propositions 5 and 6 applied to the first two derivatives of $A_{n,l}(x,\theta)$ and $B_n(x,\theta)$. Proposition 1 and proposition 3 imply that if $\lim_{n\to\infty}\frac{n}{\ln n}a_n^{m+2} = \infty$,

$$\sup_{X_n\times\Theta}\left|\frac{\partial B_n(x_i,\theta)}{\partial\theta} - \frac{\partial B(x_i,\theta)}{\partial\theta}\right| = o_p(1) \tag{7.4}$$

and

$$\sup_{X_n\times\Theta}\left|\frac{\partial A_{n,l}(x_i,\theta)}{\partial\theta} - \frac{\partial A_l(x_i,\theta)}{\partial\theta}\right| = o_p(1) \tag{7.5}$$

where

$$\frac{\partial B(x_i,\theta)}{\partial\theta_k} = -tr\frac{\partial}{\partial t}E(x|x_i\alpha)\frac{\partial\alpha}{\partial\theta_k}\cdot p(x_i\alpha|\theta) + (x_i - E(x|x_i\alpha))\frac{\partial\alpha}{\partial\theta_k}\frac{\partial}{\partial t}p(x_i\alpha|\theta) \tag{7.6}$$

and

$$\frac{\partial A_l(x_i,\theta)}{\partial\theta_k} = tr\frac{\partial}{\partial t}E(I_l(x_i-x)|x_i\alpha)\frac{\partial\alpha}{\partial\theta_k}\cdot p(x_i\alpha|\theta) + E(I_l(x_i-x)|x_i\alpha)\frac{\partial\alpha}{\partial\theta_k}\frac{\partial}{\partial t}p(x_i\alpha|\theta) \tag{7.7}$$

Similarly, proposition 1 and proposition 4 imply that if $\lim_{n\to\infty}\frac{n}{\ln n}a_n^{m+4} = \infty$,

$$\sup_{X_n\times\Theta}\left|\frac{\partial^2 B_n(x_i,\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 B(x_i,\theta)}{\partial\theta\partial\theta'}\right| = o_p(1) \tag{7.8}$$

and

$$\sup_{X_n\times\Theta}\left|\frac{\partial^2 A_{n,l}(x_i,\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 A_l(x_i,\theta)}{\partial\theta\partial\theta'}\right| = o_p(1). \tag{7.9}$$

The explicit expressions of $\frac{\partial^2 B(x_i,\theta)}{\partial\theta\partial\theta'}$ and $\frac{\partial^2 A_l(x_i,\theta)}{\partial\theta\partial\theta'}$ are rather complicate but can be found in Ichimura and Lee[1988]. However, those expressions are not needed for our subsequent analysis. Since $P_{n,l}(x_i,\theta) = \frac{A_{n,l}(x_i,\theta)}{B_n(x_i,\theta)}$, it follows that when $\lim_{n\to\infty}\frac{n}{\ln n}a_n^{m+4} = \infty$,

$$\sup_{X_n\times\Theta}\left|\frac{\partial}{\partial\theta}P_{n,l}(x_i,\theta) - \frac{\partial}{\partial\theta}P_l(x_i\alpha|\theta)\right| = o_p(1) \tag{7.10}$$

and

$$\sup_{X_n\times\Theta}\left|\frac{\partial^2}{\partial\theta\partial\theta'}P_{n,l}(x_i,\theta) - \frac{\partial^2}{\partial\theta\partial\theta'}P_l(x_i\alpha|\theta)\right| = o_p(1) \tag{7.11}$$

where

$$\frac{\partial}{\partial \theta_k} P_l(x_i\alpha|\theta) = (x_i - E(x|x_i\alpha))\frac{\partial \alpha}{\partial \theta_k}\frac{\partial}{\partial t} P_l(x_i\alpha|\theta) \tag{7.12}$$

These properties imply that when $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{m+4} = \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\bar{\theta}) \xrightarrow{p} -\sum_{l=1}^{L} E(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta'}). \tag{7.13}$$

This can be seen as follows:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\theta)\\ &= T_{n,n}(\theta) - \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L}\frac{1}{P_{n,l}(x_i,\theta)}\frac{\partial P_{n,l}(x_i,\theta)}{\partial\theta}\frac{\partial P_{n,l}(x_i,\theta)}{\partial\theta'} + \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L}\frac{\partial^2 P_{n,l}(x_i,\theta)}{\partial\theta\partial\theta'}\end{aligned} \tag{7.14}$$

where

$$\begin{aligned}&T_{n,n}(\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L}(\frac{1}{P_{n,l}(x_i,\theta)}\frac{\partial^2 P_{n,l}(x_i,\theta)}{\partial\theta\partial\theta'} - \frac{1}{P_{n,l}^2(x_i,\theta)}\frac{\partial P_{n,l}(x_i,\theta)}{\partial\theta}\frac{\partial P_{n,l}(x_i,\theta)}{\partial\theta'})(I_{li} - P_{n,l}(x_i,\theta)).\end{aligned} \tag{7.15}$$

The last term is identically zero since $\sum_{l=1}^{L} P_{n,l}(x_i,\theta) = 1$. Uniform convergence of $P_{n,l}(x_i,\theta)$ in (6.9), its derivatives in (7.10) and $P_{n,l}(x_i,\theta)$ being bounded away from zero in probability on $X_n \times \Theta$ imply that the difference between $T_{n,n}(\theta)$ and $T_{n,\infty}(\theta)$, where

$$\begin{aligned}&T_{n,\infty}(\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L}(\frac{1}{P_l(x_i\alpha|\theta)}\frac{\partial^2 P_l(x_i\alpha|\theta)}{\partial\theta\partial\theta'} - \frac{1}{P_l^2(x_i\alpha|\theta)}\frac{\partial P_l(x_i\alpha|\theta)}{\partial\theta}\frac{\partial P_l(x_i\alpha|\theta)}{\partial\theta'})(I_{li} - P_l(x_i\alpha|\theta))\end{aligned} \tag{7.16}$$

goes to zero in probability uniformly in $\theta \in \Theta$. By the law of large number for i.i.d. variables, $T_{n,\infty}(\theta) \xrightarrow{p} T_\infty(\theta)$ where

$$\begin{aligned}&T_\infty(\theta)\\ &= \sum_{l=1}^{L} E[(\frac{1}{P_l(x\alpha|\theta)}\frac{\partial^2 P_l(x\alpha|\theta)}{\partial\theta\partial\theta'} - \frac{1}{P_l^2(x\alpha|\theta)}\frac{\partial P_l(x\alpha|\theta)}{\partial\theta}\frac{\partial P_l(x\alpha|\theta)}{\partial\theta'})(P_l(x_i\alpha_o|\theta_o) - P_l(x_i\alpha|\theta))]\end{aligned} \tag{7.17}$$

is continuous in θ. Since $\bar{\theta}$ is a consistent estimate of θ_o, it follows that $T_{n,n}(\bar{\theta}) \xrightarrow{p} T_\infty(\theta_o) = 0$. Hence the result in (7.13) follows.

It remains to investigate the remaining term in (3.5). As

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}(\frac{1}{A_{n,l}(x_i,\theta_o)}\frac{\partial A_{n,l}(x_i,\theta_o)}{\partial\theta} - \frac{1}{B_n(x_i,\theta_o)}\frac{\partial B_n(x_i,\theta_o)}{\partial\theta}),\end{aligned} \tag{7.18}$$

by a Taylor expansion up to the second order,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o) + L_n + R_n \tag{7.19}$$

where

$$\begin{aligned} L_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}[\frac{1}{A_l(i)}(\frac{\partial A_{n,l}(i)}{\partial\theta} - \frac{\partial A_l(i)}{\partial\theta}) - \frac{1}{A_l^2(i)}\frac{\partial A_l(i)}{\partial\theta}(A_{n,l}(i) - A_l(i)) \\ &\quad - \frac{1}{B(i)}(\frac{\partial B_n(i)}{\partial\theta} - \frac{\partial B(i)}{\partial\theta}) + \frac{1}{B^2(i)}\frac{\partial B(i)}{\partial\theta}(B_n(i) - B(i))] \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}[\frac{1}{A_l(i)}\frac{\partial A_{n,l}(i)}{\partial\theta} - \frac{1}{A_l^2(i)}\frac{\partial A_l(i)}{\partial\theta}A_{n,l}(i) - \frac{1}{B(i)}\frac{\partial B_n(i)}{\partial\theta} + \frac{1}{B^2(i)}\frac{\partial B(i)}{\partial\theta}B_n(i)] \end{aligned} \tag{7.20}$$

and

$$\begin{aligned} &R_n \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}[-\frac{1}{\tilde{A}_{n,l}^2(i)}(\frac{\partial A_{n,l}(i)}{\partial\theta} - \frac{\partial A_l(i)}{\partial\theta})(A_{n,l}(i) - A_l(i)) + \frac{1}{\tilde{A}_{n,l}^3(i)}\frac{\partial \tilde{A}_{n,l}(i)}{\partial\theta}(A_{n,l}(i) - A_l(i))^2 \\ &+ \frac{1}{\tilde{B}_n^2(i)}(\frac{\partial B_n(i)}{\partial\theta} - \frac{\partial B(i)}{\partial\theta})(B_n(i) - B(i)) - \frac{1}{\tilde{B}_n^3(i)}\frac{\partial \tilde{B}_n(i)}{\partial\theta}(B_n(i) - B(i))^2] \end{aligned} \tag{7.21}$$

where $A_{n,l}(i) = A_{n,l}(x_i,\theta_o)$, $B_n(i) = B_n(x_i,\theta_o)$, $A_l(i) = E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o)$, $B(i) = p(x_i\alpha_o|\theta_o)$, $\tilde{A}_{n,l}(i)$ lies between $A_{n,l}(i)$ and $A_l(i)$ and $\frac{\partial \tilde{A}_{n,l}(i)}{\partial\theta}$ lies between $\frac{\partial A_{n,l}(i)}{\partial\theta}$ and $\frac{\partial A_l(i)}{\partial\theta}$ etc.

The asymptotic distribution of (7.18) can be investigated by analyzing each term separately. The remainder term R_n can be shown to converge in probability to zero. The first two terms can be analyzed by U statistic theory and standard central limit theorem. The following proposition abstracts from the analysis in Ichimura and Lee[1988].

Proposition 5

Let $C_{j,n}(z_1,\cdots,z_{i-1},z_{i+1},\cdots,z_n;\ z_i)$, $j = 1,2$ be two sequences of measurable functions of an i.i.d. sample $\{z_i\}$. Suppose that, for each j,
(1) $\sup_{Z_n}|E(C_{j,n}(z_1,\cdots,z_n;\ z_i)|z_i) - C_j(z_i)| = O(a_n^{s_j})$, for some measurable functions $C_j(z_i)$,and
(2) $\sup_{Z_n} var(C_{j,n}(z_1,\cdots,z_n;\ z_i)|z_i) = O(\frac{1}{na_n^{r_j}})$, $j = 1,2$.

If $s_1 > r_2/2, s_2 > r_1/2$, $\lim_{n\to\infty} na_n^{r_1+r_2} = \infty$ and $\lim_{n\to\infty} na_n^{2(s_1+s_2)} = 0$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{Z_n}(z_i)\,|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|\cdot|C_{2n}(z_1,\cdots,z_n;z_i) - C_2(z_i)| \xrightarrow{p} 0.$$

Proof: See appendix.

Since $A_{n,l}(i)$ and $B_n(i)$ converge in probability uniformly on X_n to well defined limits which are bounded alway from zero in probability in (6.7),(6.8),(6.10) and (6.11), $\tilde{A}_{n,l}(i)$ and $\tilde{B}_n(i)$ are bounded away uniformly from zero in probability. Furthermore, since $\frac{\partial B_n(i)}{\partial\theta}$ and $\frac{\partial A_{n,l}(i)}{\partial\theta}$ converge in probability to $\frac{\partial p(x_i,\theta_o)}{\partial\theta}$ and $\frac{\partial A_l(x_i,\theta_o)}{\partial\theta}$ uniformly on X_n in (7.4) and (7.5) and these limits are bounded, $\frac{\partial \tilde{B}_n(i)}{\partial\theta}$ and $\frac{\partial \tilde{A}_{n,l}(i)}{\partial\theta}$ are bounded in probability. It follows that

$$\| R_n \| \leq \sum_{l=1}^{L}\{\sup_{X_n} \frac{1}{|\tilde{A}^2_{n,l}(i)|} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i) \| \frac{\partial A_{n,l}(i)}{\partial \theta} - \frac{\partial A_l(i)}{\partial \theta} \| \cdot |A_{n,l}(i) - A_l(i)|$$

$$+ \sup_{X_n} \frac{1}{|\tilde{A}^3_{n,l}(i)|} \| \frac{\partial \tilde{A}_{n,l}(i)}{\partial \theta} \| \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i)(A_{n,l}(i) - A_l(i))^2$$

$$+ \sup_{X_n} \frac{1}{|\tilde{B}^2_n(i)|} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i) \| \frac{\partial B_n(i)}{\partial \theta} - \frac{\partial B(i)}{\partial \theta} \| \cdot |B_n(i) - B(i)|$$

$$+ \sup_{X_n} \frac{1}{|\tilde{B}^3_n(i)|} \| \frac{\partial \tilde{B}_n(i)}{\partial \theta} \| \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i)(B_n(i) - B(i))^2\}$$

$$\leq O_p(1) \sum_{l=1}^{L}\{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i) \| \frac{\partial A_{n,l}(i)}{\partial \theta} - \frac{\partial A_l(i)}{\partial \theta} \| \cdot |A_{n,l}(i) - A_l(i)| + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i)(A_{n,l}(i) - A_l(i))^2$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i) \| \frac{\partial B_n(i)}{\partial \theta} - \frac{\partial B(i)}{\partial \theta} \| \cdot |B_n(i) - B(i)| + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{X_n}(x_i)(B_n(i) - B(i))^2\}.$$

As the kernel function $K(\cdot)$ has zero moments up to the order s^*, proposition 2 and proposition 3 imply that

$$\sup_{X_n \times \Theta} |E(B_n(x_i, \theta)|x_i) - B(x_i, \theta)| = O(a_n^{s^*}), \tag{7.22}$$

$$\sup_{X_n \times \Theta} |E(A_{n,l}(x_i, \theta)|x_i) - A_l(x_i, \theta)| = O(a_n^{s^*}), \tag{7.23}$$

$$\sup_{X_n \times \Theta} \| E(\frac{\partial B_n(x_i, \theta)}{\partial \theta}|x_i) - \frac{\partial B(x_i, \theta)}{\partial \theta} \| = O(a_n^{s^*}) \tag{7.24}$$

and

$$\sup_{X_n \times \Theta} \| E(\frac{\partial A_{n,l}(x_i, \theta)}{\partial \theta}|x_i) - \frac{\partial A(i)}{\partial \theta} \| = O(a_n^{s^*}). \tag{7.25}$$

As it has been pointed out before, the variances of $A_{n,l}(i)$ and $B_n(i)$ have order $O(\frac{1}{na_n^m})$ and the variances of $\frac{\partial A_{n,l}(i)}{\partial \theta}$ and $\frac{\partial B_n(i)}{\partial \theta}$ have order $O(\frac{1}{na_n^{m+2}})$ uniformly in x_i. Since $s^* = m + 2$, $\lim_{n\to\infty} na_n^{2(m+1)} = \infty$ and $\lim_{n\to\infty} na_n^{4s^*} = 0$ by assumption 4, proposition 5 implies that all the terms in (7.21) and R_n converge to zero in probability.

The term L_n in (7.19) will also converge to zero in probability. The following result on U statistic will be useful.

Proposition 6

Let $\{z_i\}$ be an i.i.d.sample and $\Phi_n(z_1, z_2, a_n)$ be a sequence of measurable functions with bandwidth a_n. Suppose that

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

$$\left| E(\Phi_n(z_1, z_2, a_n)|z_1) \right| \leq h_1(z_1); \quad \left| E(\Phi_n(z_2, z_1, a_n)|z_1) \right| \leq h_2(z_1),$$

(2) $E(\Phi_n(z_1, z_2, a_n)) = O(a_n^s)$ and $var(\Phi_n(z_1, z_2, a_n)) = O(\frac{1}{a_n^r})$ for all n, and

(3) $\lim_{n\to\infty} E(\Phi_n(z_1, z_2, a_n)|z_j) = 0, a.e., j = 1, 2$.

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

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

Proof: See appendix.

Let

$$\begin{aligned}
&\Phi_n(x_i, I_i, x_j, I_j, a_n) \\
&= I_{X_n}(x_i) \sum_{l=1}^{L} I_{li}\{\frac{1}{A_l(i)}[I_{lj}(\frac{\partial x_i\alpha(\theta_o)}{\partial\theta} - \frac{\partial x_j\alpha(\theta_o)}{\partial\theta})\frac{1}{a_n^{m+1}}\frac{\partial}{\partial w}K(\frac{x_i\alpha_o - x_j\alpha_o}{a_n}) - \frac{\partial A_l(i)}{\partial\theta}] \\
&- \frac{1}{A_l^2(i)}\frac{\partial A_l(i)}{\partial\theta}[I_{lj}\frac{1}{a_n^m}K(\frac{x_i\alpha_o - x_j\alpha_o}{a_n}) - A_l(i)] \\
&- \frac{1}{B(i)}[(\frac{\partial x_i\alpha(\theta_o)}{\partial\theta} - \frac{\partial x_j\alpha(\theta_o)}{\partial\theta})\frac{1}{a_n^{m+1}}\frac{\partial}{\partial w}K(\frac{x_i\alpha_o - x_j\alpha_o}{a_n}) - \frac{\partial B(i)}{\partial\theta}] \\
&+ \frac{1}{B^2(i)}\frac{\partial B(i)}{\partial\theta}[\frac{1}{a_n^m}K(\frac{x_i\alpha_o - x_j\alpha_o}{a_n}) - B(i)]\}.
\end{aligned} \tag{7.26}$$

We note that from (7.20), $L_n = \frac{1}{\sqrt{n(n-1)}}\sum_{i=1}^n \sum_{j\neq i}^n \Phi_n(x_i, I_i, x_j, I_j, a_n)$. It is apparent from (7.26) that $E(\Phi_n^2(x_i, I_i, x_j, I_j, a_n)|x_i, I_i) = O(\frac{1}{a_n^{m+2}})$ uniformly in (x_i, I_i). Therefore $E(\Phi_n^2(x_i, I_i, x_j, I_j, a_n)) = O(\frac{1}{a_n^{m+2}})$. It is also clear from the biases in (7.22)-(7.25) that

$$E(\Phi_n(x_i, I_i, x_j, I_j, a_n)|x_i, I_i) = O(a_n^{s^*})$$

uniformly in (x_i, I_i). Hence $E(\Phi_n(x_i, I_i, x_j, I_j, a_n)) = O(a_n^{s^*})$. It remains to analyze the following term:

$$\begin{aligned}
&E(\Phi_n(x_j, I_j, x_i, I_i, a_n)|x_i, I_i) \\
&= \sum_{l=1}^{L}\{I_{li}E[W_{1,n}(x_j, I_{lj}, x_i)|x_i] - I_{li}E[W_{2,n}(x_j, I_{lj}, x_i)|x_i] \\
&- E[W_{3,n}(x_j, I_{lj}, x_i)|x_i] + E[W_{4,n}(x_j, I_{lj}, x_i)|x_i]\}
\end{aligned} \tag{7.27}$$

where

$$W_{1,n}(x_j, I_{lj}, x_i) = E(I_{X_n}(x_j)I_{lj}(\frac{\partial x_j\alpha_o}{\partial\theta} - \frac{\partial x_i\alpha_o}{\partial\theta})|x_j\alpha_o)\frac{1}{a_n^{m+1}}\frac{\partial}{\partial v}K(\frac{x_j\alpha_o - x_i\alpha_o}{a_n})/A_l(j), \tag{7.28}$$

$$W_2,n(x_j, I_{lj}, x_i) = E(I_{X_n}(x_j)I_{lj}\frac{\partial A_l(j)}{\partial\theta}|x_j\alpha_o)\frac{1}{a_n^m}K(\frac{x_j\alpha_o - x_i\alpha_o}{a_n})/A_l^2(j), \tag{7.29}$$

$$W_{3,n}(x_j, I_{lj}, x_i) = E(I_{X_n}(x_j)I_{lj}(\frac{\partial x_j\alpha_o}{\partial\theta} - \frac{\partial x_i\alpha_o}{\partial\theta})|x_j\alpha_o)\frac{1}{a_n^{m+1}}\frac{\partial}{\partial v}K(\frac{x_j\alpha_o - x_i\alpha_o}{a_n})/B(j) \tag{7.30}$$

and

$$W_{4,n}(x_j, I_{lj}, x_i) = E(I_{X_n}(x_j)I_{lj}\frac{\partial B(j)}{\partial\theta}|x_j\alpha_o)\frac{1}{a_n^m}K(\frac{x_j\alpha_o - x_i\alpha_o}{a_n})/B^2(j). \tag{7.31}$$

Under assumption 1(3), all the conditional expectations in (7.28)-(7.31) are continuous almost everywhere in $x_j\alpha_o$. At the continuity points, similar arguments in the proofs of the propositions 2 and 3 imply . for each component θ_k,[7]

[7] We recall that in the following equations the partial derivatives with respect to the argument t refer to the partial derivatives with respect to the indices $x\alpha$.

$$
\begin{aligned}
&\lim_{n\to\infty} E(W_{1,n}(x_j, I_{lj}, x_i)_k|x_i) \\
&= -p(x_i\alpha_o|\theta_o)\cdot tr\frac{\partial}{\partial t}[E(I_l(x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)\frac{1}{A_l(i)}] \\
&\quad - E(I_l(x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)\frac{1}{A_l(i)}\frac{\partial}{\partial t}p(x_i\alpha_o|\theta_o)/p(x_i\alpha_o|\theta_o) \\
&= -p(x_i\alpha_o|\theta_o)\cdot tr\frac{\partial}{\partial t}[E((x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)/p(x_i\alpha_o|\theta_o)] \\
&\quad - (E(x|x_i\alpha_o)-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}\frac{\partial}{\partial t}p(x_i\alpha_o|\theta_o)/p(x_i\alpha_o|\theta_o) \\
&= -tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}
\end{aligned}
\tag{7.32}
$$

and

$$
\begin{aligned}
&\lim_{n\to\infty} E(W_{3,n}(x_j, I_{lj}, x_i)_k|x_i) \\
&= -p(x_i\alpha_o|\theta_o)\cdot tr\frac{\partial}{\partial t}[E(I_l(x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)\frac{1}{B(i)}] - E(I_l(x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)\frac{1}{B(i)}\frac{\partial}{\partial t}p(x_i\alpha_o|\theta_o) \\
&= -p(x_i\alpha_o|\theta_o)tr\frac{\partial}{\partial t}[E((x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)E(I_l|x_i\alpha_o)/p(x_i\alpha_o|\theta_o)] \\
&\quad - E((x-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o)E(I_l|x_i\alpha_o)\frac{\partial}{\partial t}p(x_i\alpha_o|\theta_o)/p(x_i\alpha_o|\theta_o) \\
&= -E(I_l|x_i\alpha_o)\cdot tr\frac{\partial}{\partial t}E(x\frac{\partial\alpha(\theta_o)}{\partial\theta_k}|x_i\alpha_o) - (E(x|x_i\alpha_o)-x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}\frac{\partial}{\partial t}E(I_l|x_i\alpha_o)
\end{aligned}
\tag{7.33}
$$

by using the porperty $E(x|I_{li}=1, x_i\alpha_o) = E(x|x_i\alpha_o)$. At θ_o, equation(7.7) is

$$
\begin{aligned}
&\frac{\partial A_l(x_i,\theta_o)}{\partial\theta_k} \\
&= tr\frac{\partial}{\partial t}[E(x_i - x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}E(I_l|x_i\alpha_o)]p(x_i\alpha_o|\theta_o) + (x_i - E(x|x_i\alpha_o))\frac{\partial\alpha(\theta_o)}{\partial\theta_k}E(I_l|x_i\alpha_o)\frac{\partial p(x_i\alpha_o|\theta_o)}{\partial t} \\
&= -tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o) + (x_i - E(x|x_i\alpha_o))\frac{\partial\alpha(\theta_o)}{\partial\theta_k}\frac{\partial}{\partial t}[E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o)].
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\lim_{n\to\infty} E(W_{2n}(x_j, I_{lj}, x_i)|x_i) \\
&= [-tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o)]E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o)/A_l^2(i) \\
&= -tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k} \quad a.e.
\end{aligned}
\tag{7.34}
$$

Similarly, from equation (7.6),

$$
\begin{aligned}
&\lim_{n\to\infty} E(W_{4n}(x_j, I_{lj}, x_i)|x_i) \\
&= [-tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}p(x_i\alpha_o|\theta_o)]E(I_l|x_i\alpha_o)p(x_i\alpha_o|\theta_o)/B^2(i) \\
&= -tr\frac{\partial}{\partial t}E(x|x_i\alpha_o)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}E(I_l|x_i\alpha_o) \qquad a.e.
\end{aligned}
\tag{7.35}
$$

Therefore, it follows from(7.27),

$$\begin{aligned}
&\lim_{n\to\infty} E(\Phi_n(x_j, I_j, x_i, I_i, a_n)|x_i, I_i)\\
&= \sum_{l=1}^{L}\{I_{li}\lim_{n\to\infty} E[W_{1,n}(x_j, I_{lj}, x_i)|x_i] - I_{li}\lim_{n\to\infty} E[W_{2,n}(x_j, I_{lj}, x_i)|x_i]\\
&\quad - \lim_{n\to\infty} E[W_{3,n}(x_j, I_{lj}, x_i)|x_i] + \lim_{n\to\infty} E]W_{4,n}(x_j, I_{lj}, x_i)|x_i]\}\\
&= \sum_{l=1}^{L}(E(x|x_i\alpha_o) - x_i)\frac{\partial\alpha(\theta_o)}{\partial\theta_k}\frac{\partial}{\partial t}E(I_l|x_i\alpha_o) \;\; a.e.\\
&= 0 \;\; a.e.
\end{aligned}$$

since $\sum_{l=1}^{L} I_l = 1$ implies $\sum_{l=1}^{L}\frac{\partial}{\partial t}E(I_l|x_i\alpha_o) = \frac{\partial}{\partial t}E(\sum_{l=1}^{L} I_l|x_i\alpha_o) = 0$. Assumption 4(2) implies that $\sqrt{n}a_n^{s^*} \longrightarrow 0$ and $na_n^{m+2} \longrightarrow \infty$. The conditions in proposition 6 are satified and we conclude $L_n \xrightarrow{P} 0$.

As L_n and R_n converge to zero in probability, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i, \theta_o)$ is asymptotically equivalent to $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o))$ from equation (7.19). The mean of the later quantity is

$$\begin{aligned}
&\sqrt{n}E(I_{X_n}(x)\sum_{l=1}^{L} I_l\frac{\partial}{\partial\theta}\ln P_l(x\alpha_o|\theta_o))\\
&= \sqrt{n}E(I_{X_n}(x)\sum_{l=1}^{L}\frac{\partial}{\partial\theta}P_l(x\alpha_o|\theta_o))\\
&= 0
\end{aligned}$$

because $\sum_{l=1}^{L} P_l(x\alpha|\theta) = 1$ for all $\theta \in \Theta$ and $\frac{\partial}{\partial\theta}\sum_{l=1}^{L} P_l(x\alpha|\theta) = 0$. Central limit theorem for double array implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o) \xrightarrow{D} N(0,\Omega) \tag{7.36}$$

where

$$\begin{aligned}
\Omega &= \sum_{l=1}^{L} E(P_l(x\alpha_o|\theta_o)\frac{\partial}{\partial\theta}\ln P_l(x\alpha_o|\theta_o)\frac{\partial}{\partial\theta'}\ln P_l(x\alpha_o|\theta_o))\\
&= \sum_{l=1}^{L} E(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial P_l(x\alpha_o|\theta_o))}{\partial\theta}\frac{\partial P_l(x\alpha_o|\theta_o)}{\partial\theta'})
\end{aligned} \tag{7.37}$$

The asymptotic distribution of the semiparametric MLE $\hat{\theta}$ from (2.7) follows.

From inequality (6.14),

$$\sup_{\Theta}|n(Q_n(\theta) - \bar{Q}_n(\theta))| \leq \sum_{l=1}^{L}\sup_{X_n\times\Theta} n|1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(X_i,\theta) > \Delta_n)| \cdot O_p(1).$$

which will converge in probability to zero if

$$\sup_{X_n\times\Theta} n|1 - I(A_{n,l}(x_i,\theta) > \delta_n, B_n(x_i,\theta) > \Delta_n)| \xrightarrow{P} 0$$

for each l. In fact, $I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)$ can converge in probability to unity uniformly on $X_n \times \Theta$ at a nearly exponential rate. Specifically, we can show that for any finite $s \geq 0$,

$$n^s \sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)| \xrightarrow{p} 0. \tag{7.38}$$

For any given $\epsilon > 0$, we have from (6.15) that

$$\begin{aligned} &P(n^s \sup_{X_n\times\Theta} |1 - I(A_{n,l}(x_i,\theta) > \Delta_n, B_n(x_i,\theta) > \Delta_n)| \geq \epsilon) \\ &\leq \frac{n^s}{\epsilon}[P(\inf_{X_n\times\Theta} B_n(x_i,\theta) \leq \Delta_n) + P(\inf_{X_n\times\Theta} A_{n,l}(x_i,\theta) \leq \Delta_n)]. \end{aligned} \tag{7.39}$$

Since $p(x\alpha|\theta)$ is strictly bounded alway from zero on $S_x \times \Theta$, there exists a positive constant δ such that $p(x\alpha|\theta) \geq \delta$ for all $(x,\theta) \in S_x \times \Theta$. When n is sufficiently large, $\Delta_n < \frac{\delta}{2}$ and

$$P(\inf_{X_n\times\Theta} B_n(x_i,\theta) \leq \Delta_n) \leq P(\inf_{X_n\times\Theta} B_n(x_i,\theta) < \frac{\delta}{2}).$$

As $E(B_n(x_i,\theta)|x_i)$ converges to $p(x_i\alpha|\theta)$ uniformly on $X_n\times\Theta$ from (6.6), $E(B_n(x_i,\theta)|x_i) \geq p(x_i\alpha|\theta) - \frac{\delta}{4} \geq \frac{3}{4}\delta$ on $X_n \times \Theta$ for sufficiently large n. Therefore the event $\inf_{X_n\times\Theta} B_n(x_i,\theta) < \frac{\delta}{2}$ implies $E(B_n(x_i,\theta)|x_i) - B_n(x_i,\theta) \geq \frac{3}{4}\delta - \frac{1}{2}\delta = \frac{\delta}{4}$ for some $(x_i,\theta) \in X_n \times \Theta$ for large n. Hence

$$\begin{aligned} &P(\inf_{X_n\times\Theta} B_n(x_i,\theta) < \frac{\delta}{2}) \\ &\leq P(\sup_{X_n\times\Theta} |B_n(x_i,\theta) - E(B_n(x_i,\theta)|x_i)| \geq \frac{\delta}{4}) \\ &\leq c_0 exp(-(c_1 n a_n^m + c_2 \ln n)) \end{aligned}$$

for some constants c_0, c_1 and c_2 where $c_0 > 0$ and $c_1 > 0$ by the uniform law of large number in proposition 1. It follows that

$$n^s P(\inf_{X_n\times\Theta} B_n(x_i,\theta) \leq \Delta_n) \leq c_0 exp(-(c_1 n a_n^m + (c_2 - s)\ln n))$$

which will converge to zero as $\frac{n a_n^m}{\ln n} \to \infty$. Similarly, since $E(I_l|x\alpha(\theta))$ is also bounded alway from zero on $S_x \times \Theta$, $n^s P(\inf_{X_n\times\Theta} A_{n,l}(x_i,\theta) \leq \Delta_n) \to 0$ as $\frac{n a_n^m}{\ln n} \to \infty$. The result (7.38) follows from (7.39). Therefore,

$$\sup_{\Theta} |n(Q_n(\theta) - \bar{Q}_n(\theta))| \xrightarrow{p} 0.$$

which implies that $\sqrt{n}(\hat{\hat{\theta}} - \theta_o)$ is asymptotically equivalent to $\sqrt{n}(\hat{\theta} - \theta_o)$ (Amemiya [1982]).

8. Proof of The Asymptotic Properties for The Sequential Choice Models

Under the assumptions 6-10, corresponding to equations(6.7),(6.8) and (6.9), we have

$$\sup_{\bar{X}_n \times \Theta} |\bar{B}_n(x_i,\theta) - E(I_1|x_i\delta(\theta))\bar{p}(x_i\delta|\theta)| \xrightarrow{p} 0, \tag{8.1}$$

$$\sup_{\bar{X}_n \times \Theta} |\bar{A}_{n,l}(x_i,\theta) - E(J_l|I_{1i} = 1, x_i\delta(\theta))E(I_1|x_i\delta(\theta))\bar{p}(x_i\delta|\theta)| \xrightarrow{p} 0 \tag{8.2}$$

and hence

$$\sup_{\bar{X}_n \times \Theta} |\bar{P}_{n,l}(x_i,\theta) - \bar{P}_l(x_i\delta|\theta)| \xrightarrow{p} 0. \tag{8.3}$$

Similar to the proofs of Theorem 1 and Corollary 1, we can establish consistency of our estimators for the sequential choice model.

The semiparametric MPLE $\hat{\theta}$ from (4.5) satisfies the first order condition:

$$\frac{1}{n}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\hat{\theta}) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_{n,l}(x_i,\hat{\theta})] = 0.$$

By Taylor expansion at $\theta = \theta_o$,

$$\begin{aligned}
&\sqrt{n}(\hat{\theta} - \theta_o) \\
&= -\{\frac{1}{n}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\bar{\theta}) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln \bar{P}_{n,l}(x_i,\hat{\theta})]\}^{-1} \\
&\quad \cdot \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_{n,l}(x_i,\theta_o)].
\end{aligned} \tag{8.4}$$

When both $\lim_{n\to\infty} \frac{n}{\ln n} a_n^{m+4} = \infty$ and $\lim_{n\to\infty} \frac{n}{\ln n} b_n^{\bar{m}+4} = \infty$, it is apparent that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln P_{n,l}(x_i,\bar{\theta}) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial^2}{\partial\theta\partial\theta'}\ln \bar{P}_{n,l}(x_i,\bar{\theta})] \\
&\xrightarrow{p} -[\sum_{l=1}^{L} E\Big(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial}{\partial\theta}P_l(x\alpha_o|\theta_o)\cdot\frac{\partial}{\partial\theta'}P_l(x\alpha_o|\theta_o)\Big) \\
&\quad + \sum_{l=1}^{L_1} E\Big(\frac{P_1(x\alpha_o|\theta_o)}{\bar{P}_l(x\delta_o|\theta_o)}\frac{\partial}{\partial\theta}\bar{P}_l(x\delta_o|\theta_o)\cdot\frac{\partial}{\partial\theta'}\bar{P}_l(x\delta_o|\theta_o)\Big)]
\end{aligned} \tag{8.5}$$

by the similar arguments for the proof of equation(7.13). Proposition 6 will also imply that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_{n,l}(x_i,\theta_o) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_{n,l}(x_i,\theta_o)]$$

is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_l(x_i\delta_o|\theta_o)]. \tag{8.6}$$

The mean of the first term in (8.6) has been shown to be zero. The mean of the second term is also zero because

$$E(I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_l(x_i\delta_o|\theta_o)|x_i)$$

$$= P_1(x_i\alpha_o|\theta_o)\sum_{l=1}^{L_1}\frac{\partial}{\partial\theta}\bar{P}_l(x_i\delta_o|\theta_o)$$

$$= 0$$

which follows from $\sum_{l=1}^{L_1}\bar{P}_l(x_i\delta|\theta) = 1$ for all $\theta \in \Theta$. Since

$$E\Big(\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o)\cdot\sum_{r=1}^{L_1} I_{1i}J_{ri}\frac{\partial}{\partial\theta'}\ln \bar{P}_r(x_i\delta_o|\theta_o)|x_i\Big)$$

$$= \sum_{r=1}^{L_1} E(I_{1i}J_{ri}|x_i)\frac{\partial}{\partial\theta}\ln P_1(x_i\alpha_o|\theta_o)\cdot\frac{\partial}{\partial\theta'}\ln \bar{P}_r(x_i\delta_o|\theta_o)$$

$$= \frac{\partial}{\partial\theta}P_1(x_i\alpha_o|\theta_o)\cdot\sum_{r=1}^{L_1}\frac{\partial}{\partial\theta'}\bar{P}_r(x_i\delta_o|\theta_o)$$

$$= 0,$$

the two terms in (8.6) are uncorrelated. Central limit theorem for double array will imply that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[I_{X_n}(x_i)\sum_{l=1}^{L} I_{li}\frac{\partial}{\partial\theta}\ln P_l(x_i\alpha_o|\theta_o) + I_{\bar{X}_n}(x_i)I_{1i}\sum_{l=1}^{L_1} J_{li}\frac{\partial}{\partial\theta}\ln \bar{P}_l(x_i\delta_o|\theta_o)]\xrightarrow{D} N(0,\Sigma) \qquad (8.7)$$

where

$$\Sigma$$

$$= \sum_{l=1}^{L} E\Big(\frac{1}{P_l(x\alpha_o|\theta_o)}\frac{\partial}{\partial\theta}P_l(x\alpha_o|\theta_o)\cdot\frac{\partial}{\partial\theta'}P_l(x\alpha_o|\theta_o)\Big) + \sum_{l=1}^{L_1} E\Big(\frac{P_1(x\alpha_o|\theta_o)}{\bar{P}_l(x\delta_o|\theta_o)}\frac{\partial}{\partial\theta}\bar{P}_l(x\delta_o|\theta_o)\cdot\frac{\partial}{\partial\theta'}\bar{P}_l(x\delta_o|\theta_o)\Big). \qquad (8.8)$$

Appendix

Proof of Proposition 2:

$$\begin{aligned}
&E(c(z,z_i,\theta)\frac{1}{a_n^m}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}|z_i) \\
&= \int_{T_1(\theta)}^{T_2(\theta)} E(c(z,z_i,\theta)|t,z_i)\frac{1}{a_n^m}K(\frac{t(z_i,\theta)-t}{a_n})g(t|\theta)dt \\
&= \int_{\frac{1}{a_n}(t(z_i,\theta)-T_2(\theta))}^{\frac{1}{a_n}(t(z_i,\theta)-T_1(\theta))} E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta)K(w)dw.
\end{aligned}$$

For any $(z_i,\theta)\in Z_n$, $\frac{1}{a_n}(t(z_i,\theta)-T_1(\theta))\geq\frac{\delta_n}{a_n}$ and $\frac{1}{a_n}(t(z_i,\theta)-T_2(\theta))\leq-\frac{\delta_n}{a_n}$. Since $\frac{\delta_n}{a_n}$ tends to infinity when n is large enough, D will be contained in the rectangle $[-\frac{\delta_n}{a_n},\ \frac{\delta_n}{a_n}]$. Hence, when n is large enough,

$$\begin{aligned}
&\int_{\frac{1}{a_n}(t(z_i,\theta)-T_2(\theta))}^{\frac{1}{a_n}(t(z_i,\theta)-T_1(\theta))} E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta)K(w)dw \\
&= \int_D E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta)K(w)dw
\end{aligned}$$

for all $(z_i,\theta)\in Z_n$. Therefore,

$$\begin{aligned}
&\sup_{(z_i,\theta)\in Z_n}|E(c(z,z_i,\theta)\frac{1}{a_n^m}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i)-E(c(z,z_i,\theta)|t(z_i,\theta),z_i)g(t(z_i,\theta)|\theta)| \\
&= \sup_{(z_i,\theta)\in Z_n}|\int_D[E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta) \\
&\quad -E(c(z,z_i,\theta)|t(z_i,\theta),z_i)g(t(z_i,\theta)|\theta)]K(w)dw| \\
&\leq \sup_{(t,z_i,\theta)}\sup_{w\in D}|E(c(z,z_i,\theta)|t-a_nw,z_i)g(t-a_nw|\theta)-E(c(z,z_i,\theta)|t,z_i)g(t|\theta)|\int_D|K(v)|dv
\end{aligned}$$

which converges to zero by uniform continuity and boundedness of the functions.

Let $R(z_i,\theta,a_nw)=E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta)$. Let $R_{i_1\cdots i_k}(z_i,\theta,t)$ denote the kth partial derivatives of $R(z_i,\theta,t)$ with respect to the components $t_{i_1},\cdots,t_{i_k}$ where $t=(t_1,\cdots,t_m)\in R^m$. By a Taylor expansion up to order s^*,

$$R(z,\theta,a_nw)=R(z,\theta,0)+\frac{1}{i!}\sum_{i=1}^{s^*-1}\frac{\partial^iR(z,\theta,0)}{\partial a_n^i}a_n^i+\frac{1}{s^*!}\frac{\partial^{s^*}R(z,\theta,\bar{a}_nw)}{\partial a_n^{s^*}}a_n^{s^*}.$$

We note that $\frac{\partial^i}{\partial a_n^i}R(z,\theta,a_nw)=\sum_{j_1=1}^m\cdots\sum_{j_i=1}^mR_{j_1\cdots j_i}(z,\theta,a_nw)w_{j_1}\cdots w_{j_i}$. As the first s^*-1 moments of $K(w)$ are zero, $\int_D\frac{\partial^iR(z,\theta,0)}{\partial a_n^i}dw=0,\ 1\leq i\leq s^*-1$. Hence

$$\int_D[R(z_i,\theta,a_nw)-R(z_i,\theta,0)]K(w)dw=a_n^{s^*}\frac{1}{s^*!}\int_D\frac{\partial^{s^*}}{\partial a_n^{s^*}}R(z_i,\theta,\bar{a}_nw)K(w)dw$$

and

$$\begin{aligned}
&\sup_{(z_i,\theta)\in Z_n}|\int_D[E(c(z,z_i,\theta)|t(z_i,\theta)-a_nw,z_i)g(t(z_i,\theta)-a_nw|\theta) \\
&\quad -E(c(z,z_i,\theta)|t(z_i,\theta),z_i)g(t(z_i,\theta)|\theta)]K(w)dw| \\
&\leq a_n^{s^*}\frac{1}{s^*!}\sum_{i_1=1}^m\cdots\sum_{i_{s^*}=1}^m\sup_{(z_i,\theta,t)}|R_{i_1\cdots i_{s^*}}(z_i,\theta,t)|\cdot\int_D\|w\|^{s^*}|K(w)|dw \\
&= O(a_n^{s^*})
\end{aligned}$$

by uniform boundedness of the derivatives.

Q.E.D.

Proof of Proposition 3

$$
\begin{aligned}
&E(c(z,z_i,\theta)\frac{1}{a_n^{m+1}}\frac{\partial}{\partial w}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i)\\
&=\frac{1}{a_n}\int_{\frac{1}{a_n}(t(z_i,\theta)-T_2(\theta))}^{\frac{1}{a_n}(t(z_i,\theta)-T_1(\theta))} E(c(z,z_i,\theta)|t(z_i,\theta)-a_n w,z_i)\cdot g(t(z_i,\theta)-a_n w|\theta)\frac{\partial}{\partial w}K(w)dw\\
&=\frac{1}{a_n}\int_D E(c(z,z_i,\theta)|t(z_i,\theta)-a_n w,z_i)\cdot g(t(z_i,\theta)-a_n w|\theta)\frac{\partial}{\partial w}K(w)dw
\end{aligned}
$$

for any $(z_i,\theta)\in Z_n$ when n is large enough. By Taylor expansion at $a_n=0$,

$$
\begin{aligned}
&E(c(z,z_i,\theta)|t(z_i,\theta)-a_n w,z_i)\cdot g(t(z_i,\theta)-a_n w|\theta)\\
&=E(c(z,z_i,\theta)|t(z_i,\theta),z_i)\cdot g(t(z_i,\theta)|\theta)\\
&-a_n[g(t(z_i,\theta)-\hat{a_n}w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)-\hat{a_n}w,z_i)\\
&+E(c(z,z_i,\theta)|t(z_i,\theta)-\hat{a_n}w,z_i)\cdot\frac{\partial}{\partial t'}g(t(z_i,\theta)-\hat{a_n}w|\theta)]w.
\end{aligned}
$$

Since $K(w)$ goes to zero at the boundary of D, $\int_D \frac{\partial}{\partial w}K(w)dw=0$. It follows that

$$
\begin{aligned}
&E(c(z,z_i,\theta)\frac{1}{a_n^{m+1}}\frac{\partial}{\partial w'}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i)\\
&=-\int_D[g(t(z_i,\theta)-\hat{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)-\hat{a}_n w,z_i)\\
&\quad+E(c(z,z_i,\theta)|t(z_i,\theta)-\hat{a}_n w,z_i)\frac{\partial}{\partial t'}g(t(z_i,\theta)-\hat{a}_n w|\theta)]w\frac{\partial}{\partial w'}K(w)dw
\end{aligned}
$$

By integration by parts, $\int_D w\frac{\partial}{\partial w}K(w)dw=-I_l$. Therefore,

$$
\begin{aligned}
&\sup_{(z_i,\theta)\in Z_n}|E(c(z,z_i,\theta)\frac{1}{a_n^{m+1}}\frac{\partial}{\partial w'}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n}|z_i)\\
&\quad-g(t(z_i,\theta)|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta),z_i)-E(c(z,z_i,\theta)|t(z_i,\theta),z_i)\frac{\partial}{\partial t'}g(t(z_i,\theta)|\theta)|\\
&=\sup_{(z_i,\theta)\in Z_n}|\int_D[g(t(z_i,\theta)|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta),z_i)\\
&\quad+E(c(z,z_i,\theta)|t(z_i,\theta),z_i)\frac{\partial}{\partial t'}g(t(z_i,\theta)|\theta)-g(t(z_i,\theta)-\bar{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w,z_i)\\
&\quad-E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w,z_i)\frac{\partial}{\partial t'}g(t(z_i,\theta)-\bar{a}_n w|\theta)]\cdot w\frac{\partial}{\partial w'}K(w)dw|\\
&\le(\sup_{(t,\theta)}\sup_{w\in D}\|g(t|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t,z_i)-g(t-\bar{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t-\bar{a}_n w,z_i)\|\\
&\quad+\sup_{(t,\theta)}\sup_{w\in D}\|E(c(z,z_i,\theta)|t,z_i)\frac{\partial}{\partial t'}g(t|\theta)-E(c(z,z_i,\theta)|t-\bar{a}_n w,z_i)\frac{\partial}{\partial t'}g(t-\bar{a}_n w|\theta)\|)\\
&\quad\cdot\int_D\|w\|\cdot\|\frac{\partial}{\partial w'}K(w)\|dw
\end{aligned}
$$

which converges to zero under our assumption.

Let

$$R(z_i,\theta,a_n w) = \frac{\partial}{\partial t}[g(t|\theta)E(c(z,z_i,\theta)|t,z_i)]|_{t=t(z_i,\theta)-a_n w}.$$

By Taylor expansion up to order s^*,

$$\begin{aligned}
&R(z,\theta,a_n w)\\
&= R(z,\theta,0) + \frac{1}{i!}\sum_{i=1}^{s^*-1}\frac{\partial^i R(z,\theta,0)}{\partial a_n^i}a_n^i + \frac{1}{s^*!}\frac{\partial^{s^*}R(z,\theta,\bar{a}_n w)}{\partial a_n^{s^*}}a_n^{s^*}\\
&= R(z,\theta,0) + \frac{1}{i!}\sum_{i=1}^{s^*-1}(\sum_{j_1=1}^{m}\cdots\sum_{j_i=1}^{m}R_{j_1\cdots j_i}(z,\theta,0)w_{j_1}\cdots w_{j_i})a_n^i\\
&\quad + \frac{1}{s^*!}\sum_{j_1=1}^{m}\cdots\sum_{j_{s^*}=1}^{m}R_{j_1\cdots j_{s^*}}(z,\theta,\bar{a}_n w)w_{j_1}\cdots w_{j_{s^*}}a_n^{s^*}
\end{aligned}$$

where $R_{j_1\cdots j_i}(z,\theta,t) = \frac{\partial^i R(z,\theta,t)}{\partial t_{j_1}\cdots\partial t_{j_i}}$. For any $j \notin \{j_1\cdots,j_i\}$,

$$\begin{aligned}
&\int_D w_{j_1}w_{j_2}\cdots w_{j_i}\frac{\partial}{\partial w_j}K(w)dw\\
&= \int_D w_{j_1}\cdots w_{j_i}\int\frac{\partial}{\partial w_j}K(w)dw_j dw_1\cdots dw_{j-1}dw_{j+1}\cdots dw_m\\
&= 0.
\end{aligned}$$

Also, for any $p \geq 1$, $j \notin \{j_1,\cdots,j_i\}$ and $1 \leq i+p-1 \leq s^*-1$,

$$\begin{aligned}
&\int_D w_{j_1}w_{j_2}\cdots w_{j_i}w_j^p\frac{\partial}{\partial w_j}K(w)dw\\
&= \int_D w_{j_1}\cdots w_{j_i}\int w_j^p\frac{\partial}{\partial w_j}K(w)dw_j dw_1\cdots dw_{j-1}dw_{j+1}\cdots dw_m\\
&= -p\int_D w_{j_1}\cdots w_{j_i}w_j^{p-1}K(w)dw\\
&= 0.
\end{aligned}$$

It follows that

$$\begin{aligned}
&|\int_D\{\frac{\partial}{\partial t}[g(t|\theta)E(c(z,z_i,\theta)|t,z_i)]|_{t=t(z_i,\theta)-a_n w} - \frac{\partial}{\partial t}[g(t|\theta)E(c(z,z_i,\theta)|t,z_i)]|_{t=t(z_i,\theta)}\}w\frac{\partial}{\partial w'}K(w)dw|\\
&\leq a_n^{s^*}\frac{1}{s^*!}\sup_{(z_i,\theta)}\sup_{w\in D}\sum_{j_1=1}^{l}\cdots\sum_{j_{s^*}=1}^{l}\| R_{j_1\cdots j_{s^*}}(z_i,\theta,\bar{a}_n w)\|\int_D\| w_{j_1}\cdots w_{j_{s^*}}w\|\cdot\|\frac{\partial}{\partial w'}K(w)\| dw\\
&= O(a_n^{s^*}).
\end{aligned}$$

Q.E.D.

Proof of Proposition 4

$$\begin{aligned}
&E(c(z,z_i,\theta)\frac{1}{a_n^{m+2}}\frac{\partial^2}{\partial t^2}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i)\\
&= \frac{1}{a_n{}^2}\int_{\frac{1}{a_n}(t(z_i,\theta)-T_2(\theta))}^{\frac{1}{a_n}(t(z_i,\theta)-T_1(\theta))}E(c(z,z_i,\theta)|t(z_i,\theta)-a_n w)g(t(z_i,\theta)-a_n w|\theta)\frac{\partial^2}{\partial w^2}K(w)dw\\
&= \frac{1}{a_n^2}\int_D E(c(z,z_i,\theta)|t(z_i,\theta)-a_n w)g(t(z_i,\theta)-a_n w|\theta)\frac{\partial^2}{\partial w^2}K(w)dw
\end{aligned}$$

for $(z_i,\theta) \in Z_n$ when n is sufficiently large. By a Taylor expansion at $a_n = 0$ up to the second order,

$$
\begin{aligned}
&E(c(z,z_i,\theta)|t-a_n w)g(t-a_n w|\theta) - E(c(z,z_i,\theta)|t)g(t) \\
&= -a_n[g(t|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t) + E(c(z,z_i,\theta)|t)\frac{\partial}{\partial t'}g(t|\theta)]w \\
&+ \frac{a_n^2}{2}w'[\frac{\partial}{\partial t}g(t-\bar{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t-\bar{a}_n w) + g(t-\bar{a}_n w|\theta)\frac{\partial^2}{\partial t\partial t'}E(c(z,z_i,\theta)|t-\bar{a}_n w) \\
&+ \frac{\partial}{\partial t}E(c(z,z_i,\theta)|t-\bar{a}_n w)\frac{\partial}{\partial t'}g(t-\bar{a}_n w|\theta) + E(c(z,z_i,\theta)|t-\bar{a}_n w)\frac{\partial^2}{\partial t\partial t'}g(t-\bar{a}_n w|\theta)]w.
\end{aligned}
$$

Since $K(w)$ and $\frac{\partial}{\partial w}K(w)$ go to zero at the boundary of D, it is easy to see with integration by parts that $\int_D \frac{\partial^2}{\partial w^2}K(w) = 0$, $\int_D w\frac{\partial^2}{\partial w_l\partial w_r}K(w)dw = 0$, $\int_D w_l^2\frac{\partial^2}{\partial w_l^2}K(w)dw = 2$, $\int_D w_l w_r\frac{\partial^2}{\partial w_l\partial w_r}K(w)dw = 1$ for $l \neq r$, and $\int_D w_l w_r \frac{\partial^2}{\partial w_p\partial w_q}K(w)dw = 0$ for all other cases. Hence

$$
\begin{aligned}
&\sup_{(z_i,\theta)\in Z_n} |E(c(z,z_i,\theta)\frac{1}{a_n^{m+2}}\frac{\partial^2}{\partial w_l\partial w_r}K(\frac{t(z_i,\theta)-t(z,\theta)}{a_n})|z_i) - g(t(z_i,\theta)|\theta)\cdot\frac{\partial^2}{\partial w_l\partial w_r}E(c(z,z_i,\theta)|t(z_i,\theta)) \\
&- \frac{\partial}{\partial t_l}g(t(z_i,\theta)|\theta)\frac{\partial}{\partial t_r}E(c(z,z_i,\theta)|t(z_i,\theta)) - \frac{\partial}{\partial t_l}E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial}{\partial t_r}g(t(z_i,\theta)|\theta) \\
&- E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial^2}{\partial t_l\partial t_r}g(t(x_i,\theta)|\theta)| \\
&= \frac{1}{2}\sup_{(z_i,\theta)\in Z_n}|\int_D w'[\frac{\partial}{\partial t}g(t(z_i,\theta)-\bar{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w) \\
&- \frac{\partial}{\partial t}g(t(z_i,\theta)|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)) + g(t(z_i,\theta)-\bar{a}_n w|\theta)\frac{\partial^2}{\partial t\partial t'}E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w) \\
&- g(t(z_i,\theta)|\theta)\frac{\partial^2}{\partial t^2}E(c(z,z_i,\theta)|t(z_i,\theta)) + \frac{\partial}{\partial t}E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w)\frac{\partial}{\partial t'}g(t(z_i,\theta)-\bar{a}_n w|\theta) \\
&- \frac{\partial}{\partial t}E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial}{\partial t'}g(t(z_i,\theta)|\theta) + E(c(z,z_i,\theta)|t(z_i,\theta)-\bar{a}_n w)\frac{\partial^2}{\partial t\partial t'}g(t(z_i,\theta)-\bar{a}_n w|\theta) \\
&- E(c(z,z_i,\theta)|t(z_i,\theta))\frac{\partial^2}{\partial t\partial t'}g(t(z_i,\theta)|\theta)]w\frac{\partial^2}{\partial w_l\partial w_r}K(w)dw|
\end{aligned}
$$

which is less than

$$
\begin{aligned}
&\frac{1}{2}(\sup_{(t,\theta)}\sup_{w\in D} \| \frac{\partial}{\partial t}g(t-\bar{a}_n w|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t-\bar{a}_n w) - \frac{\partial}{\partial t}g(t|\theta)\frac{\partial}{\partial t'}E(c(z,z_i,\theta)|t) \| \\
&+ \sup_{(t,\theta)}\sup_{w\in D} \| g(t-\bar{a}_n w|\theta)\frac{\partial^2}{\partial t\partial t'}E(c(z,z_i,\theta)|t-\bar{a}_n w) - g(t|\theta)\frac{\partial^2}{\partial t\partial t'}E(c(z,z_i,\theta)|t) \| \\
&+ \sup_{(t,\theta)}\sup_{w\in D} \| \frac{\partial}{\partial t}E(c(z,z_i,\theta)|t-\bar{a}_n w)\frac{\partial}{\partial t'}g(t-\bar{a}_n w|\theta) - \frac{\partial}{\partial t}E(c(z,z_i,\theta)|t)\frac{\partial}{\partial t'}g(t|\theta) \| \\
&+ \sup_{(t,\theta)}\sup_{w\in D} \| E(c(z,z_i,\theta)|t-\bar{a}_n w)\frac{\partial^2}{\partial t\partial t'}g(t-\bar{a}_n w|\theta) - E(c(z,z_i,\theta)|t)\frac{\partial^2}{\partial t\partial t'}g(t|\theta) \|) \\
&\cdot\int_D \| w' \| \cdot \| w \| \cdot |\frac{\partial^2}{\partial w_l\partial w_r}K(w)|dw
\end{aligned}
$$

and hence converges to zero under our assumptions.

Q.E.D.

Proof of Proposition 5

For any $\delta > 0$, by Markov inequality and Cauchy inequality,

$$P(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{Z_n}(z_i)|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|\cdot|C_{2,n}(z_1,\cdots,z_n;z_i) - C_2(z_i)| > \delta)$$
$$\leq \frac{1}{\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} E(I_{Z_n}(z_i)|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|\cdot|C_{2,n}(z_1,\cdots,z_n;z_i) - C_2(z_i)|)$$
$$\leq \frac{1}{\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{E(I_{Z_n}(z_i)|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|^2)$$
$$E(I_{Z_n}(z_i)|C_{2,n}(z_1,\cdots,z_n;z_i) - C_2(z_i)|^2)\}^{\frac{1}{2}}.$$

The moment conditions (1) and (2) imply that

$$E(I_{Z_n}(z_i)|C_{j,n}(z_1,\cdots,z_n;z_i) - C_j(z_i)|^2)$$
$$= E(I_{Z_n}(z_i)[var(C_{j,n}(z_1,\cdots,z_n;z_i)|z_i) + (E(C_{j,n}(z_1,\cdots,z_n;z_i)|z_i) - C_j(z_i))^2])$$
$$\leq O(\frac{1}{na_n^{r_j}}) + O(a_n^{2s_j}), \quad j = 1, 2.$$

Hence,

$$E(I_{Z_n}(z_i)|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|^2)\cdot E(I_{Z_n}(z_i)|C_{2,n}(z_1,\cdots,z_n;z_i) - C_2(z_i)|^2)$$
$$\leq O(\frac{1}{n^2a_n^{r_1+r_2}}) + O(\frac{1}{n}a_n^{2s_1-r_2}) + O(\frac{1}{n}a_n^{2s_2-r_1}) + O(a_n^{2(s_1+s_2)})$$

and

$$P(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_{Z_n}(z_i)|C_{1,n}(z_1,\cdots,z_n;z_i) - C_1(z_i)|\cdot|C_{2,n}(z_1,\cdots,z_n;z_i) - C_2(z_i)| > \delta)$$
$$\leq \frac{1}{\delta}[O(\frac{1}{na_n^{r_1+r_2}}) + O(a_n^{2s_1-r_2}) + O(a_n^{2s_2-r_1}) + O(na_n^{2(s_1+s_2)})]^{1/2}$$

which converges to zero.

Q.E.D.

Proof of Proposition 6:

Let $U_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Phi_n(z_i,z_j,a_n)$. Define the symmetric function,

$$\Psi_n(z_i,z_j.a_n) = \frac{1}{2}(\Phi_n(z_i,z_j,a_n) + \Phi_n(z_j,z_i,a_n)).$$

U_n can be rewritten in the standard U statistic form:

$$U_n = \frac{1}{C_2^n}\sum_{i=1}^{n}\sum_{j>i}\Psi_n(z_i,z_j,a_n)$$

where $C_2^n = n(n-1)/2$. As

$$E(nU_n^2) = n\cdot var(U_n) + n\cdot E^2(\Psi_n(z_1,z_2,a_n))$$
$$= n\cdot var(U_n) + O(na_n^{2s})$$

and $na_n^{2s} \to 0$, it remains to show that $n\cdot var(U_n)$ goes to zero. From Hoeffding[1948],

$$n\cdot var(U_n) = \frac{2}{(n-1)}[2(n-2)var(E[\Psi_n(z_1,z_2,a_n)|z_1]) + var(\Psi_n(z_1,z_2,a_n))]$$
$$= \frac{4(n-2)}{(n-1)}var(E[\Psi_n(z_1,z_2,a_n)|z_1]) + \frac{2}{n-1}var(\Psi_n(z_1,z_2,a_n))$$

As

$$\begin{aligned}
& var(E[\Psi_n(z_1, z_2, a_n)|z_1]) \\
&= E\{E^2[\Psi_n(z_1, z_2, a_n)|z_1]\} - E^2(\Psi_n(z_1, z_2, a_n)), \\
&= E\{E^2[\Psi_n(z_1, z_2, a_n)|z_1]\} - O(a_n^{2s})
\end{aligned}$$

it follows that

$$\begin{aligned}
& \lim_{n\to\infty} var(E[\Psi_n(z_1, z_2, a_n)|z_1]) \\
&= E\{\frac{1}{4}[\lim_{n\to\infty} E(\Phi_n(z_1, z_2, a_n)|z_1) + \lim_{n\to\infty} E(\Phi_n(z_2, z_1, a_n)|z_1)]^2\} \\
&= 0
\end{aligned}$$

by Lebesque dominated convergence theorem. The remaining term converges also to zero because

$$\begin{aligned}
& \frac{1}{n} var(\Psi_n(z_1, z_2, a_n)) \\
&\le \frac{1}{4n}\{[var(\Phi_n(z_1, z_2, a_n))]^{1/2} + [var(\Phi_n(z_1, z_2, a_n))]^{1/2}\}^2 \\
&= O(\frac{1}{na_n^r}).
\end{aligned}$$

Q.E.D.

References

1. Amemiya, T.[1982], "Two Stage Least Absolute Deviations Estimators", **Econometrica** 50; 689-711.
2. Bierens,H.J.[1985], "Kernel Estimators of Regression Functions", manuscript, University of Amsterdam.
3. Chamberlain, G.[1986], "Asymptotic Efficiency in Sem-parametric Models with Censoring", **Journal of Econometrics** 32; 189-218.
4. Cosslett, S.R.[1987], "Efficiency Bounds for Distribution-Free Estimators of the Censored Regression Models", **Econometrica** 55; 559-585.
5. Cosslett,S.R.[1983], "Distribution-free Maximum Likelihood Estimator of the Binary Choice Model", **Econometrica** 51; 765-782.
6. Han,A.K.[1987], "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator", **Journal of Econometrics** 35; 303-316.
7. Ichimura, H.[1987], "Estimation of Single Index Models", Ph.D. Thesis, Department of Economics, M.I.T.
8. Ichimura, H. and L.F.Lee [1988], "Semiparametric Estimation of Multiple Index Models: Single Equation Estimation", Manuscript, Department of Economics, University of Minnesota, forthcoming in **Non-parametric and Semiparametric Methods in Econometrics and Statistics**, edited by W.A. Barnett, J. Powell and G. Tauchen.
9. Klein, R.W. and R.H. Spady [1987], "An Efficient Semiparametric Estimator for Discrete Choice Models", Economics Research Group, Bell Communications Research, Morristown, N.J.
10. Manski,C.F.[1975], "Maximum Score Estimation of the Stochastuc Utility Model of Choice", **Journal of Econometrics** 3; 205-228.
11. Manski, C.F.[1985],"Semiparanetric Analysis of discrete Response: Asymptotic Properties of the Maximum Score Estimator", **Journal of Econometrics** 27; 313-333.
12. Matzkin, R.L.[1987], "Semiparametric Estimation of Monotonic and Concave Utility Functions: The Discrete Choice Case", Discussion Paper no.830, Cowles Foundation for Research in Economics, Yale University.
13. McFadden, D.[1987], "Conditional Logit Analysis of Qualitative Choice Bahavior", in P. Zarembka, ed., **Frontiers in Econometrics**, New York, Academic Press.
14. Press,W.H., B.P.Flannery, S.A.Teukosky and W.T.Vetterling [1986], **Numerical Recipes**, Cambridge University Press, Cambridge.
15. Severini,T.A. and W.H. Wong [1987], "Profile Likelihood and Semiparametric Models", unpublised manuscript, Department of Statistics, University of Chicago.
16. Silverman,B.[1986],**Density Estimation**, London: Chapman and Hall.
17. Stone, C.[1956], "Efficient Nonparametric Testing and Estimation." **Proc. Third Berkeley Symp. Math. Statist. Prob. 1** pp.187-196. University of california Press, Berkeley, California.
18. Thompson, T.S.[1989], "Least Squares Estimation of Semiparametric Discrete Choice Models", Discussion Paper No.249, Center for Economic Research,Department of Economics, University of minnesota, Minneapolis.