Simulated MLE for Discrete Choices using Transformed Simulated Frequencies[1]

Donghoon Lee[2] and Kyungchul Song[3]

January 19, 2009

Abstract

This paper proposes a new method of simulated MLE for discrete choice models which is easy to implement and flexible enough to accommodate a variety of model specifications, and yet entirely free from the simulation bias for each finite number of simulations. By a deliberate transformation of the likelihood function, the estimation method is designed to generate consistent estimators even with a small number of simulations. The transformation is explicit, containing no unknowns that demand an additional step of estimation. The estimator achieves the efficiency of MLE as the simulation number increases fast enough. In order to emphasize the flexibility of the framework, we performed a Monte Carlo study using a dynamic model of schooling choice in which heterogeneity is introduced for discount factors and ability.

Key words: Simulated MLE, Discrete Choice Models, Simulated Frequency, Cube-Root Asymptotics

JEL Classifications: C12, C14, C52.

1 Introduction

Discrete choice models have long been very popular in applied researches across a wide range of empirical fields of economics. While a discrete choice model typically specifies the data

[1] A prevoius version was titled, "A Consistent SMLE When the Simulation Number was Finite." We are indebted to Sean D. Campbell for his valuable comments and inputs. We are also grateful to the seminar participants at Columbia University, Econometric Society Meeting, Greater New York Econometrics Colloquium, Lehigh University and New York University for useful comments. The views expressed are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

[2] Federal Reserve Bank of New York, 33 Liberty Street, New York, NY, 10045.

[3] Department of Economics, University of Pennsylvania, 528 McNeil Building, 3718 Locust Walk, Philadelphia, Pennsylvania 19104-6297.

generating process up to a parametric family of distributions, maximum likelihood estimation is infeasible in practice except for extremely simple models because the explicit evaluation of the likelihood is infeasible. The approach of simulation-based inference has been instrumental for overcoming this difficulty, providing the researcher with a much wider spectrum of flexibility in modeling. However, this simulation-based approach entails simulation error in the inference, and the problem of analyzing and controling this simulation bias has drawn attention in the literature. (See Hajivassiliou and Ruud (1994), Stern (1997) and Gouriéroux and Monfort (1997) for a review of the literature and references therein.)

There are mainly two popular methods of simulation-based inference: a method-of-moments approach and a likelihood-based approach. The method of simulated moment (MSM) approach developed by McFadden (1989) has been very popular and widely used in empirical researches. See Lee (1992), McFadden and Ruud (1994) and Keane (1994) for some surveys and improvements. First, the approach allows for flexible modeling of discrete responses as the latent process is permitted to be non-normal and nonlinear in errors. Second, the estimator from the simulated method of moment approach is known to be $\sqrt{n}$-consistent even with a finite number of simulations. However, MSM does not achieve the efficiency of MLE even with a large simulation number because the estimating equation does not utilize fully the likelihood information. Furthermore, MSM has an unattractive feature of having to choose moment conditions in practice. When the number of choice and the number of covariates are large, the choice of moment conditions can be a nontrival problem to an empirical researcher. Most importantly, the computation of an efficient weighting matrix for MSM is complicated as one has to simulate the first order derivative of the moment condition. The use of an efficient weighting matrix also requires estimating the same parameter twice, and hence in terms of computational cost, it is not clear to compare the performance of MSM using an optimal weighting matrix with that of MLE in finite samples.

On the other hand, the likelihood-based approach has the merit of achieving efficiency as the simulation number increases to infinity along with the sample size (Lerman and Manski (1981)). However, the existing methods of simulated MLE (SMLE) incur simulation bias (with the sample size $n \to \infty$ and the simulation number R fixed) that does not disappear unless one increases the simulation number to infinity. Most existing literatures on simulated MLE have focused on simulating the choice probabilities. Among the popular examples are the simulated frequency method of Lerman and Manski (1981), smoothed simulated MLE, the simulation method of Stern (1992), or a simulated MLE using the GHK simulator (Geweke (1989), Hajivassiliou (1990) and Keane (1993).) However, these methods induce simulation bias that disappears only when R increases to infinity. This is true even when simulated choice probabilities are unbiased because the expected logarithm is not equal to

the logarithm of expected values. (See Lee (1995) for an extensive analysis of asymptotic bias in simulated MLE.) There have been approaches in the literature to reduce the magnitude of this simulation bias by using bias correction. (See e.g. Lee (1995) and Gouriéroux and Monfort (1997)).

Hajivassiliou (1990) and McFadden and Hajivassiliou (1998) proposed a different method of SMLE that uses simulated scores to construct simulated moment conditions and proved efficiency of the estimators. (See Börsch-Supan and Hajivassiliou (1993) for a review and a simulation study.) The estimators do not suffer from simulation bias under a finite simulation number, but this is achieved only when the latent process follows a normal linear structure.

This paper suggests an entirely different approach in which the simulation bias is eliminated completely for each finite R, while maintaining the flexibility of modeling. In particular, we do not assume that the latent process is normal linear. Furthermore, the estimator achieves asymptotic efficiency as R increases to infinity. These desirable features are achieved by identifying a sequence of transforms of simulated probabilities that lead to identification of the parameter for each finite R. The sequence of transforms converge to a logarithmic function as R increases to infinity, so that the estimator achieves asymptotic efficiency when R increases fast enough. We summarize the characteristics of this new approach in the following.

(1) While our approach generates an estimator free from simulation bias under a finite simulation number, the estimator achieves the efficiency of MLE when the simulation number increases to infinity.

(2) While our method allows for flexible modeling as in Lerman and Manski (1981), especially admitting latent processes that are not normal linear, it does not suffer from a zero-probability problem.

(3) The method is very easy to implement, causing virtually no additional complication beyond that entailed by Lerman and Manski (1981)'s procedure.

To the best of our knowledge, our method is unique among the existing simulation based approaches in that it satisifes both Properties of (1) and (2). The existing methods of simulation-based inference either (a) fail to achieve the efficiency or (b) suffer from the simulation bias when the simulation number is small or (c) heavily rely on the assumption that the latent process is normal linear. In particular, Property (2) is significant for many empirical researches. For example, the latent process is allowed to be nonlinear in unobserved stochastic terms whose distributions are not normal. This feature is not shared by the approaches involving GHK simulators. While the original idea of Lerman and Manski (1981)

does not assume a normal linear latent process, it is well-known that their method suffers from a zero-probability problem. Some literatures have suggested proposals to deal with this problem (e.g. Stern (1992), Geweke (1989), Hajivassiliou (1990) and Keane (1993).) However, such proposals either assume normal linear latent processes or requires smoothing. In either case, the suggested properties fail to satisfy Property (1). Our approach does not suffer from this problem while maintaining the merit of flexible modeling. We illustrate the usefulness of this property in our simulation study based on the modeling of heterogeneity in schooling choices. Property (3) also deserves attention. Our estimator is obtained through a one-step optimization of an objective function. The evaluation of the objective function is computationally as fast as many existing SMLEs.

In this paper, we formally present conditions for identification and derive the asymptotic theory for the estimator in both the cases of simulation numbers fixed and increasing with the sample size. Our exposition is made through easily verifiable, high-level conditions to emphasize the flexibility of our approach. The conditions require only weak regularity conditions for the stochastic link between the decision variables and the observed covariates that ensures the identification of the parameters. We also demonstrate how our framework can also be immediately adapted to the case where only the cohort-level aggregate data are available. This set-up is often relevant to empirical researches in Industrial Organizations.

Here is the summary of the asymptotic theory of our estimators. When the simulation number is finite, the estimator follows the cube-root asymptotic theory in Kim and Pollard (1990). Similarly as for the maximum score estimator (Manski (1975)), this slower rate of convergence stems from "the sharp edge effect" due to the presence of discontinuities in the objective function. The simplex-based optimization algorithm is well-known to be useful for optimizing a discontinuous objective function.

In the case of an increasing number of simulations, we establish that the estimator is $\sqrt{n}$-consistent and asymptotically normal as the simulation number increases to infinity at a rate slightly faster than $\sqrt{n}$. This latter condition is only slightly stronger than the existing condition for many SMLEs. (See e.g. Lerman and Manski (1981) and Gourieroux and Monfort (1997).) Under this same condition, the estimator achieves the asymptotic efficiency of MLE.

To illustrate the usefulness of our approach, we performed a Monte Carlo simulation study based on a schooling choice model which involves heterogeneity in discount factor and ability. More specifically, the discount factor is assumed to be correlated with other observed individual characteristic and also an unobserved characteristic. The simultation methods considered in this study are, Lerman-Manski's SMLE, smoothed SMLE, MSM without using optimal weighting matrix and MSM with using optimal weighting matrix.

The optimal weighting matrix is not that which ensures the efficiency of MLE, but that of the usual optimal weighting matrix in GMM. Recall that the MSS based on a normal linear process and inferences based on GHK simulators cannot be employed for this model. Here is the summary of the findings from the study.

First, our estimator mostly dominates Lerman and Manski's simulation method and smoothed SMLEs regardless of the simulation number. The domination is prominent especially when the simulation number is small and the sample size is large. It is worth noting that when discrete choice models are complicated in structural modeling, reflecting various heterogenous components, the Lerman-Manski simulation method often emerges as one of the most attractive simulation methods. The general principle of Lerman-Manski's method, in particular, does not require that observed components are related to unobserved components in a specific manner.

Second, our estimator overall dominates MSM that does not use the optimal weighting matrix. When the sample size is small and the simulation number is large, our estimator performs better than MSM that uses the optimal weighting matrix. For other cases, our estimator performs well, comparable to this MSM with optimal weighting matrix. However, this direct comparison may not be fair because in order to estimate this optimal MSM, one has to estimate the parameters in the first step. Hence the computation of the optimal MSM takes roughly twice long as MSM without the optimal weighting matrix.

The remainder of this paper is organized as follows. In section 2, we define the class of discrete choice models and discuss SMLE. In section 3, we introduce transformed simulated frequency (TSF) and present our main result of identification of parameters for each finite simulation number. Section 4 establishes the asymptotic properties of the estimator. Section 5 is devoted to two examples. The first example concerns with static random utility models and the second one discusses the case when only cohort-level aggregate data are available. In Section 6, we present and discuss results from a Monte Carlo simulation study. Section 7 concludes. All the technical proofs are relegated to the appendix.

2 A Discrete Choice Model and SMLE

We introduce a discrete choice model and notations. Suppose that a binary decision variable, D_{ij}, of an agent i choosing the j-th choice, is stochastically linked with an observed covariate vector X_i as follows:

$$D_{ij} = \delta_j(X_i, \eta_i; \theta_0),$$

where $X_i = (X_{i1}, \cdots, X_{iJ})'$ represents a vector of observed covariates, $\eta_i = (\eta_{i1}, \cdots, \eta_{iJ})'$ denotes a random vector representing unobserved shocks, and $\theta_0 \in \Theta \subset \mathbf{R}^d$ represents the parameters to be estimated. The number J represents the number of the choices the agent encounters and n represents the number of agents in the data set. For example, δ_j can be specified as follows,

$$\delta_j(X_i, \eta_i; \theta_0) = 1\left\{u_j(X_i, \eta_i; \theta_0) \geq \max_{1 \leq k \leq J} u_k(X_i, \eta_i; \theta_0)\right\}.$$

Here the function $u_j(X_i, \eta_i; \theta)$ often has a structural interpretation as a random utility (McFadden (1974)). While such a structural interpretation is a prime example, our framework does not strictly require that $\delta_j(X_i, \eta_i; \theta)$ have a random utility specification of the above form.

The choice probability of the agent choosing the j-th option is defined by

$$p_j(X_i, \theta) = \mathbf{E}\left[\delta_j(X_i, \eta_i; \theta)|X_i\right].$$

The choice probability is obtained by "integrating out" the unobserved variable η_i conditional on the observed covariate X_i. It is interpreted as the probability of the j-th choice being made by an agent i with a covariate X_i. Given the choice probabilities $p_j(X_i, \theta)$, it is natural to form the log-likelihood of a random sample $\{D_i, X_i\}$ as follows:

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}\log(p_j(X_i, \theta)).$$

In the case that $p_j(x, \theta)$ has a closed form representation, the maximum likelihood estimation is straightforward. (e.g. Amemiya (1985).) However, the choice probability is often hard to evaluate, in particular when the number of choices, J, is large and one wants to admit flexibility in specifying the joint distribution of η_i. One can bypass this difficulty by using a choice probability simulator p^*_{jR} and constructing a simulated log-likelihood,

$$l^*_{n,R}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}\log(p^*_{jR}(X_i, \theta)). \tag{1}$$

The number R represents the repitition number of simulated stochastic variables. Then, the simulated maximum likelihood estimator (SMLE) is defined as a maximizer of the above simulated likelihood function,

$$\hat{\theta}^*_{n,R} \equiv \arg\max_{\theta\in\Theta} l^*_{n,R}(\theta).$$

As mentioned in the introduction, many literatures have focused on improving the simulation of p^*_{jR}. The simulation bias in this case disappears only with R increasing. This paper suggests a different approach of SMLE that is entirely free from the simulation bias even with fixed R. We explain this approach in the following section in detail.

3 Transformed Simulated Frequency MLE

3.1 Transformed Simulated Frequency

Suppose that the R number of stochastic errors $\eta^*_{i,r},\ r = 1, \cdots, R$, are drawn from the known distribution F. Then $\delta_j(X_i, \eta^*_{i,r}; \theta),\ r = 1, \cdots, R$, denotes simulated choices for each value of θ. The simulated frequency of each choice is obtained by

$$m_{jR}(X_i, \eta^*_i; \theta) = \sum_{r=1}^{R} \delta_j(X_i, \eta^*_{i,r}; \theta)$$

where $\eta^*_i = (\eta^*_{i,1}, \cdots, \eta^*_{i,R})$ is a random sample from the distribution F of η_i. The quantity $m_{jR}(X_i, \eta^*_i; \theta)$ represents the number of incidences that the j-th choice is made by an agent i that has covariates X_i and simulated stochastic errors η^*_i. From now on, we write briefly

$$m_{ji}(\theta) = m_{jR}(X_i, \eta^*_i; \theta) \tag{2}$$

and $m_i(\theta) = (m_{1i}(\theta), \cdots, m_{Ji}(\theta))'$.

Our method uses an objective function that involves a transformed version of the simulated frequencies m_{jR}. More specifically, let $\mathbb{N}_R = \{0, 1, 2, \cdots, R\}$ and define

$$\mathbb{N}_{R,J} = \left\{(m_1, \cdots, m_J) : m_j \in \mathbb{N}_R,\ j = 1, \cdots, J,\ \text{and}\ \Sigma_{j=1}^{J} m_j = R\right\}.$$

The set $\mathbb{N}_{R,J}$ denotes the space of J-tuples of simulated frequencies. Suppose that we are given a set of maps $T^j_R : \mathbb{N}_{R,J} \to \mathbf{R},\ j = 1, \cdots, J$. Then, we can construct an objective function whose maximization yields an estimator of θ_0 as follows:

$$\hat{\theta}(\{T^j_R\}) = \arg\max_{\theta\in\Theta} l^*_{n,R}(\theta; \{T^j_R\})$$

where

$$l^*_{n,R}(\theta; \{T^j_R\}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T^j_R(m_i(\theta)). \tag{3}$$

The estimator $\hat{\theta}(\{T^j_R\})$ certainly depends on the maps $T^j_R : \mathbb{N}_{R,J} \to \mathbf{R}$. For example, when we choose the map T^j_R to be

$$T^j_R(m) = \log\left(\frac{m_j}{R}\right), \tag{4}$$

the estimator $\hat{\theta}(\{T^j_R\})$ is reduced to the SMLE of Lerman and Manski (1981). As is well known, this choice of T^j_R leads to an estimator $\hat{\theta}(\{T^j_R\})$ that suffers from a simulation bias. Our paper's goal is to identify transforms T^j_R such that the estimator $\hat{\theta}(\{T^j_R\})$ does not suffer from a simulation bias at all when $n \to \infty$ for each finite R.

Absence of simulation bias implies that we can identify the true parameter when the simulation number R is finite and the sample size n is sufficiently large. To describe identification under fixed R, we consider the following population objective function

$$l_R(\theta; \{T^j_R\}) \equiv \mathbf{E} l^*_{n,R}(\theta; \{T^j_R\}). \tag{5}$$

The function $l_R(\theta; \{T^j_R\})$ can be thought of as the probability limit of the simulated objective function divided by n as n goes to infinity. The identification of θ_0 is proceeded in two steps: the identification of the choice probability function $p_j(x; \theta_0)$ from the data set and the identification of the parameter θ_0 from the choice probability function $p_j(x; \theta_0)$. The second step is standard in the literature of discrete choice models and we will delineate sufficient conditions later. For now, it suffices to note that transforms $\{T^j_R\}$ affects the identification of θ_0 only through affecting the identification of the choice probability. Therefore, the main question is what choice of transforms $\{T^j_R\}$ guarantees the identification of the choice probabilities for each finite R.

There are two main challenges in the search of transforms $\{T^j_R\}$ that leads to identification of θ_0 under fixed R. First, it is not even known *a priori* whether such transforms exist. Second, even when such transforms exist, the transforms may depend on unknown aspects of the data generating process so that the transforms eventually have to be estimated from the data. Such transforms are not attractive, because the procedure introduce additional noise into the inference procedure and may lead to inefficient estimator even when R is large. The major contribution of this paper is the discovery of the set of transforms $\{T^j_R\}$ that have an algebraically simple and explicit form and are independent of the underlying data generating process.

Intuitively, the search for such transforms that are independent of unknown aspects of

P should utilize only information that is fully known. Define a simplex $S_J = \{p \in [0,1]^J : \Sigma_j^J p_j = 1\}$. Given a transform $\{T_R^j\}$, we introduce a function $\Lambda_R(p, p_0; \{T_R^j\}) : S_J \times S_J \to \mathbf{R}$ as follows:

$$\Lambda_R(p, p_0; \{T_R^j\}) = \sum_{j=1}^{J} p_{j0} \int T_R^j(m) dF_R(m; p),$$

where $F_R(\cdot; p)$ is the multinomial distribution function on $\mathbb{N}_{R,J}$ with parameter (R, p), $p = (p_1, \cdots, p_J)' \in S_J$, $p_0 = (p_{10}, \cdots, p_{J0})' \in S_J$. The function Λ_R is uniquely determined once R and the transform $\{T_R^j\}$ are chosen, and it does not depend on any other specifics of the data generating process.

This function Λ_R "links" between the transform $\{T_R^j\}$ and the population objective function $l_R(\theta; \{T_R^j\})$ in the following way. First note that

$$\Lambda_R(p(x,\theta), p(x,\theta_0); \{T_R^j\}) = \mathbf{E}\left[l_{n,R}^*(\theta; \{T_R^j\}) | X_i = x\right],$$

where $p(x,\theta) = (p_1(x,\theta), \cdots, p_J(x,\theta))'$. From this, it follows that we can write

$$l_R(\theta; \{T_R^j\}) = \mathbf{E}\left[\Lambda_R(p(X_i,\theta), p(X_i,\theta_0); \{T_R^j\})\right].$$

Therefore, the identification of the choice probability $p(x,\theta_0)$ hinges on the way $\{T_R^j\}$ is related to Λ_R. In other words, the condition for $\{T_R^j\}$ that ensures the identification of $p(x,\theta_0)$ can be characterized as that for Λ_R. In fact the condition for Λ_R can be characterized from the first order necessary condition for the identification of $p(x,\theta_0)$.

To see this, we assume the interchangeability of the derivative and expectation. Then, the first order condition (FOC) for θ_0 for maximizing l_R is written as,

$$\frac{\partial l_R(\theta; \{T_R^j\})}{\partial \theta}|_{\theta=\theta_0} = \sum_{j=1}^{J} \mathbf{E}\left[\lambda_{jR}(p(X_i,\theta_0), p(X_i,\theta_0); \{T_R^j\}) \frac{\partial p_j(X_i,\theta_0)}{\partial \theta}\right] = 0, \tag{6}$$

where

$$\lambda_{jR}(p, p_0; \{T_R^j\}) = \frac{\partial \Lambda_R(p, p_0; \{T_R^j\})}{\partial p_j}. \tag{7}$$

Since the choice probabilities sum up to one for all x and θ, differentiability of $p_j(x,\theta)$ at each θ implies

$$\sum_{j=1}^{J} \frac{\partial p_j(x,\theta)}{\partial \theta} = 0. \tag{8}$$

Therefore, the first order condition immediately follows if the following condition is satisfied:

for each $p_0 \in S_J$,

$$\lambda_{jR}(p_0, p_0; \{T_R^j\}) = \lambda_{kR}(p_0, p_0; \{T_R^j\}), \text{ for all } j, k = 1, 2, \cdots, J. \tag{9}$$

The condition in (9) has an important merit of not depending on any aspect of the data generating process. This means that if there are $\{T_R^j\}$ such that (9) is satisfied, we may be able to select a transform that does not depend on the data generating process and hence is fully known. The main idea of this paper is that we utilize the restriction in (9) to find an appropriate set of transforms $\{T_R^j\}$.

While (9) is stronger than the original first order condition in (6) for many data generating processes, our focusing on the restriction in (9) is almost necessary because the condition in (9) becomes equivalent to the original first order condition for some data generating processes. For example, consider an extreme case where $J = 2$ and X_i has a degenerate distribution with a single point mass at x and $\partial p_1(x, \theta_0)/\partial\theta = 1/2$ and $\partial p_2(x, \theta_0)/\partial\theta = -1/2$. Then the first order condition is equal to

$$\begin{aligned} 0 &= \sum_{j=1}^{J} \lambda_{jR}(p(x, \theta_0), x; \{T_R^j\}) \frac{\partial p_j(x, \theta_0)}{\partial\theta} \\ &= \frac{1}{2}\lambda_{1R}(p(x, \theta_0), x; \{T_R^j\}) - \frac{1}{2}\lambda_{2R}(p(x, \theta_0), x; \{T_R^j\}). \end{aligned}$$

In this case, the condition (9) implies that λ_{jR} is invariant to j. Since we do not want to impose *a priori* additional restrictions upon the data generating process, our search for $\{T_R^j\}$ should accommodate this kind of situation, i.e., a situation where the conditions (6) and (9) are equivalent. Therefore, it appears natural to focus on the condition (9) and let it guide the search for an appropriate $\{T_R^j\}$. It remains to identify $\{T_R^j\}$ that satisfies (9). The solution is given in the following algebraic result.

Lemma 1: *Transforms* $\{T_R^j\}$ *satisfy* (9) *if for each* $j = 1, \cdots, J$,

$$T_R^j(m) = -\sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\nu(m_{-j})}{R}, \tag{10}$$

where $\nu(m_{-j}) = \sum_{k=1, k\neq j}^{J} 1\{m_k > 0\}$. *Here we take the summation* $\sum_{s=0}^{-1}$ *to be zero.*

Lemma 1 is a pure algebraic result that does not involve any unknown specifics of the data generating process in the model. The proof of Lemma 1 is very simple, involving only algebraic computations. To find the form T_R^j, we first extract sufficient conditions for (9) and see what conditions are needed for T_R^j to satisfy these conditions. Then, these conditions

for T_R^j lead to an affine transform of the form in (10). We offer a proof of Lemma 1 that illuminates this discovery process.

It is not immediately clear how the choice of (10) is related to the MLE obtained by using (4) with sufficiently large R. To see this closely, note first that the simulated frequencies $m_{ji}(\theta)/R \to_P p_{ij}(\theta) \in (0,1]$ with $R \to \infty$, by the law of large numbers, and that

$$\frac{\nu(m_{-ji}(\theta))}{R} \leq \frac{J-1}{R} \to 0.$$

Finally, observe that

$$-\sum_{s=0}^{R-m_{ji}(\theta)-1} \frac{1}{R-s} \to \log(p_{ij}(\theta)).$$

This latter convergence is immediate as the sum is a Rieman lower sum of $-\int_{m_{ji}(\theta)/R}^{1}(1/x)dx$. Therefore, the objective function $l_{n,R}^*(\theta; \{T_R^j\})$ becomes closer to that of MLE as $R \to \infty$.

Let $m_{ji}(\theta)$ be as defined in (2). Then, the estimator that this paper propses is the following:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \left\{ -\sum_{s=0}^{R-m_{ji}(\theta)-1} \frac{1}{R-s} + \frac{\nu(m_{-ji}(\theta))}{R} \right\}. \tag{11}$$

The estimator is computed from a single-step maximization of the above objective function. Note that the objective function does not require any preliminary step of estimation. The form of the objective function is explicit and can be immediately evaluated. The specification of the discrete choice model affects the objective function only through the simulated frequencies $m_j(\theta)$. Hence this approach of transformed simulated frequency can be applied to any discrete choice models in which we can obtain the simulated frequencies $m_j(\theta)$. We will call $T_R^j(m)$ *transformed simulated frequency* (TSF) and the estimator $\hat{\theta}$ defined in (11) as TSF-MLE.

Lemma 1 in itself is not immediately seen to lead to the identification of θ_0 when one estimates θ_0 as in (11), because the form of T_R^j is derived from information contained only in the FOC of the (population) M-estimation problem. In Theorem 1 below, we establish that the choice of T_R^j in Lemma 1 indeed yields the identification of θ_0 for finite R. We introduce the following regularity conditions.

Assumption 1 : (i) Θ is compact with an interior containing θ_0 and for all $\theta \in \Theta$ and x in the support of X_i, the choice probability $p_j(x;\theta)$ belongs to S_J.
(ii) For each x in the support of X_i and for each $j \in \{1, \cdots, J\}$, $p_j(x;\theta)$ is twice-continuously

differentiable at $\theta = \theta_0$ and for some $\delta > 0$,

$$\mathbf{E}\left\| \sup_{\theta \in B(\theta_0,\delta)} \frac{\partial p(X_i;\theta)}{\partial \theta} \right\|^2 < \infty \text{ and } \mathbf{E}\left\| \sup_{\theta \in B(\theta_0,\delta)} \frac{\partial^2 p(X_i;\theta)}{\partial \theta \partial \theta'} \right\|^2 < \infty.$$

(iii) For all $\theta \notin \theta_0$, there exists $j \in \{1, \cdots, J\}$ such that $\mathbf{P}\{p_j(X_i;\theta_0) \neq p_j(X_i;\theta)\} > 0$.

Conditions in Assumption 1 are standard in the MLE of discrete choice models. Condition (i) requires that the choice probability function $p(x,\theta)$ is well-defined for all $\theta \in \Theta$. Condition (ii) is stronger than needed for identification. The twice differentiability is assumed envisaging the derivation of asymptotic normality in the next section. Condition (iii) is a minimal necessary condition used to idenfity θ_0 from the choice probabilities.

Theorem 1 (Identification) : *Suppose that* Assumption 1 *holds. Then for each* $\delta > 0$,

$$l_R(\theta_0; \{T_R^j\}) > \max_{\theta \in \Theta \setminus B(\theta_0,\delta)} l_R(\theta; \{T_R^j\}),$$

where $B(\theta_0,\delta) = \{\theta \in \Theta : ||\theta - \theta_0|| < \delta\}$.

The identification result in Theorem 1 is obtained by showing that the population objective function $\Lambda_R(p, p_0; \{T_R^j\})$ is globally strictly concave in $p \in S_J$ for each $p_0 \in S_J$. Therefore, the population objective function $l_R(\theta; \{T_R^j\})$ is uniquely maximized at $\theta = \theta_0$.

4 Asymptotic Properties

In this section we investigate the asymptotic properties of the estimator $\hat{\theta}$ defined in (11). The asymptotic properties of $\hat{\theta}$ are developed for two separate cases: when R is fixed and when R tends to infinity jointly with the sample size n. Let $\mathcal{X}$ be the support of X_i. We introduce the following assumptions.

Assumption 2 : (i) $\{(D_i, X_i, \eta_i)\}_{i=1}^n$ is i.i.d. from a common distribution.
(ii) For each $\tilde{\theta} \in \Theta$, $\sup_{x \in \mathcal{X}} \mathbf{E}[\sup_{\theta \in B(\tilde{\theta},\varepsilon)} |\delta_j(X_i, \eta_i; \tilde{\theta}) - \delta_j(X_i, \eta_i; \theta)|^2 | X_i = x] \leq C\varepsilon$, for some $C > 0$.
(iii) For some $\delta > 0$, $\inf_{\theta \in B(\theta_0,\delta)} \inf_{x \in \mathcal{X}} p_j(x;\theta_0) > \varepsilon_p > 0$, $j = 1, \cdots, J$, for some $\varepsilon_p > 0$.
(iv) $(X_i, \eta_i)_{i=1}^n$ and $(X_i, \eta_i^*)_{i=1}^n$ are distributionally identical.

Condition (ii) controls the manner the random decision rule $\delta_j(X_i, \eta_{i,r}; \theta)$ depends on θ and $(X_i, \eta_{i,r})$. The condition requires that the decision rule δ is locally uniformly L_2-continuous in θ (e.g. Chen, Linton, van Keilegom (2003)). This condition is a very useful

high-level condition that can be used to establish the stochastic equicontinuity of an empirical process involving a discontinuous function, and flexibly admits a wide class of specifications of δ. We present lower-level conditions in the case of random utility models in a later section (See Lemma 2 below.) Condition (iii) requires that the choice probabilities be bounded away from zero, and implies that $p_j(X_i;\theta_0) < 1 - \varepsilon_p$ for each $j = 1, \cdots, J$. Conditions in Assumption 2(ii) and (iii) can be weakened at the cost of introducing a more complicated procedure that involves a trimming sequence. Condition (iv) is certainly satisfied when X_i and η_i are independent and η_i^* are drawn i.i.d from F, as commonly assumed in the simulation literature.

Theorem 2 (The Rate of Convergence for Fixed R) : *Suppose that* Assumptions 1-2 *hold. Then for each fixed $R \geq 2$, we have*

$$n^{1/3}(\hat{\theta} - \theta_0) = O_P(1).$$

The rate of convergence follows the cube-root asymptotics of Kim and Pollard (1990). This rate of convergence is due to the fact that the objective function is discontinuous in the parameter. Not all the objective functions that are discontinuous in parameters yield an estimator with the cube-root asymptotics. For example, while the method of simulated moments (MSM) estimator of McFadden (1989) and the maximum rank correlation estimator of Sherman (1993) are obtained from maximizing discontinuous objective functions, they are asymptotically linear with bounded information and hence $\sqrt{n}$-consistent. However, the usual asymptotic linearity of an estimator breaks down in our case, and the rate of convergence becomes slower than the parametric rate. When R tends to infinity slightly faster than $\sqrt{n}$, not only is the $\sqrt{n}$-rate of convergence restored, but also the estimator achieves the efficiency of MLE.

Theorem 3 (Asymptotic Normality of the Estimator As $(n, R) \to \infty$ Jointly) : *Suppose that* Assumptions 1-2 *hold. As $n, R \to \infty$ jointly, with $\sqrt{n}\log(R)/R \to 0$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_p N(0, \Omega^{-1}),$$

where

$$\Omega = \mathbf{E}\left[\left(\sum_{j=1}^{J} D_{ij}\frac{\partial}{\partial\theta}\log p_j(X_i,\theta_0)\right)\left(\sum_{j=1}^{J} D_{ij}\frac{\partial}{\partial\theta'}\log p_j(X_i,\theta_0)\right)'\right].$$

The rate condition $\sqrt{n}\log(R)/R \to 0$ is satisfied when R increases slightly faster than $\sqrt{n}$. This condition is nearly close to the usual condition in the simulated MLE literature. In

particular, when $\sqrt{n}/R \to 0$, it is known that regardless of when one uses the simulation frequency of Lerman and Manski (1981) or the GHK simulator for $p^*_{jR}(X_i, \theta)$ in (1), the estimator is consistent and asymptotically normal. (See Gourieroux and Monfort (1991)).

5 Examples

5.1 Static Random Utility Models

We consider a static random utility model. Suppose that the utility of agent i with covariates X_i and stochastic errors η_i when she makes the j-th choice is given by $u_j(X_i, \eta_{ij}; \theta) = \mu_j(X_i, \theta) + \eta_{ij}$ for some function μ_j. Then she makes the j-th choice when

$$\Delta_j(X_i, \eta_i; \theta) = u_j(X_i, \eta_{ij}; \theta) - \max_{1 \le k \le J, k \ne j} u_k(X_i, \eta_{ik}; \theta)$$

is greater than zero. In this case, the decision rule δ_j is defined by

$$\delta_j(X_i, \eta_i; \theta) = 1\left\{\Delta_j(X_i, \eta_i; \theta) > 0\right\}.$$

Then the following lemma provides sufficient conditions that ensure the local uniform L_p-continuity of the random decision rule $\delta_j(X_i, \eta_i; \theta)$ in θ in Assumption 2(ii).

Lemma 2 : *Suppose that for each $\theta \in \Theta$, and for each x in the support of X,*

$$\sup_{1 \le j \le J} \left| \mu_j(x, \theta) - \mu_j(x, \tilde{\theta}) \right| \le C||\theta - \tilde{\theta}||.$$

Furthermore, assume that the conditional density of $\eta_{ij} - \eta_{ik}$ given $X_i = x$ is bounded uniformly over x in the support of X. Then the condition of Assumption 2(ii) *holds with $r = 1$.*

It is also easy to show the condition of Assumption 2(ii) for different specifications of random utilities. For example, consider the random utility specified as $u(X_i, \eta_{ij}; \theta) = A(\theta, \eta_{ij})' X_i$ where $A(\theta, \eta) = B(\theta) + \eta\Gamma(\theta)$ and X_i has a bounded support. In this case, the L_p-uniform continuity condition in Assumption 2(ii) is proved as follows. First consider

$$\begin{aligned} & u(X_i, \eta_{ij}; \theta) - u(X_i, \eta_{ij}; \tilde{\theta}) \\ = \; & (B(\theta) - B(\tilde{\theta})) \sum_{m=1}^{K} X_{im} + \sum_{m=1}^{K} (\Gamma_m(\theta) - \Gamma_m(\tilde{\theta})) \eta_{ij} X_{im}. \end{aligned}$$

Hence the L_p-continuity condition follows immediately when $B(\theta)$ and $\Gamma(\theta)$ are Lipschitz continuous in θ at θ_0. When X_i has an unbounded support, we may redefine $\tilde{u}(X_i, \eta_{ij}; \theta) = \Phi(A(\theta, \eta_{ij})' X_i)$ where Φ is a bounded strictly increasing function that is first order continuously differentiable with derivative ϕ such that $\sup_{x \in \mathbf{R}^{d_X}} \phi(x)||x|| < \infty$.

5.2 Simulated MLE with Cohort-Level Aggregate Data

In this section, we demonstrate that our results can be applied without difficulty to the case where we have only cohort-level aggregate data. The use of cohort-level aggregate data is common in the literature of empirical industrial organizations. (e.g. Pakes (1986) and Berry, Levinsohn, and Pakes (1995)). In such situations, modeling unobserved heterogeneity has drawn special attention in the literature, as the aggregate data do not contain direct information about the heterogeneity among agents. Our estimation method can be useful in this case because it allows for flexible modeling of unobserved heterogeneity.

Suppose that we have K number of cohorts and $n(k)$ number of agents in the k-th cohort. The individual decision variable $D_{ij}(k)$ corresponding to the agent i in cohort k choosing the j-th choice is defined as a binary variable such that

$$D_{ij}(k) = \delta_j(X(k), \eta_{ij}(k); \theta), \text{ when the } j\text{-th choice is made by the agent } i \text{ in cohort } k.$$

Note that the observed variable $X(k)$ is only a cohort-level aggregate covariate. The variables $D_{ij}(k)$ and $\eta_{ij}(k)$ represent the unobserved micro variables for each individual. Define

$$D_j(k) = \frac{1}{n(k)} \sum_{i=1}^{n(k)} D_{ij}(k)$$

and $D(k) = (D_1(k), \cdots, D_J(k))'$. The variable $D_j(k)$ indicates a proportion of agents in cohort k that have chosen the j-th choice. The econometrician observes only the cohort-level aggregate data $\{D(k), X(k)\}_{k=1}^K$. The (infeasible) log-likelihood of the micro data after normalizing by $n(k)$ is equal to

$$\sum_{k=1}^{K} \sum_{j=1}^{J} \frac{1}{n(k)} \sum_{i=1}^{n(k)} D_{ij}(k) \log \mathbf{P}\left\{D_{ij}(k) = 1 | X(k), \theta\right\}$$

When the conditional distribution of the stochastic error $\eta_{ij}(k)$ given $X(k)$ is identical for each individual i, the conditional probability $\mathbf{P}\left\{D_{ij}(k) = 1 | X(k), \theta\right\}$ is identical for all the individuals in the k-th cohort. This is the case when $\{\eta_{ij}(k) : i = 1, \cdots, n(k),\ k = 1, \cdots, K\}$

is i.i.d. and independent of $\{X(k) : k = 1, \cdots, K\}$. In this case, we can write the cohort-level likelihood as

$$\sum_{k=1}^{K} \sum_{j=1}^{J} D_j(k) \log \mathbf{P}\left\{D_{ij}(k) = 1 | X(k), \theta\right\}.$$

This is the log-likelihood using only the observable cohort characteristics and the proportion of agents in each cohort that made certain decisions. Let F be the fully known marginal distribution of $(\eta_{i1}(k), \cdots, \eta_{iJ}(k))$. Then, one draws R random sample from F to obtain $\{\eta_r^*(k)\}_{r=1}^R$ where $\eta_r^*(k) = (\eta_{r1}^*(k), \cdots, \eta_{rJ}^*(k))$. We define the simulated frequency

$$m_{jR}(k, \theta) = \frac{1}{R} \sum_{r=1}^{R} \delta_j(X(k), \eta_{r,j}^*(k); \theta).$$

Then using the transform that we propose here, we can construct an objective function as follows

$$l_{K,R}^*(\theta; \{T_R^j\}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} D_j(k) T_R^j(m_R(k, \theta)).$$

Note that

$$\mathbf{E}\left[D_j(k) | X(k)\right] = \mathbf{P}\left\{D_{ij}(k) = 1 | X(k)\right\}.$$

Hence one can check sufficient conditions with this choice probability. The results of Theorems 1-3 carry over to this case as long as the data $\{D(k), X(k)\}_{k=1}^K$ are cohort-wise i.i.d.

6 A Monte Carlo Study: A Model of Schooling Choice

6.1 The Model and the Simulation Design

In this section, we present and discuss results from a Monte Carlo simulation study. The model considered in the study is a model of schooling choice with observed ability and unobserved heterogeneous discount factor and preference. See Willis and Rosen (1982) and Keane and Wolpin (1997) for models of discrete choices under unobserved heterogeneity in the preferences.

Suppose that people make schooling decisions at the age 16 endowed with 10 years of education. They can choose among the 4 alternatives: 1) to drop out of high school and start working right away, 2) to graduate from high school attaining 12 years of education, 3) to graduate a 2-year college with 14 years of education, and 4) to graduate from college with 16 years of education. After finishing their respective schooling, they work until age 65 and there is no labor supply decision. Therefore, the number of periods in the model is 50

periods.

People are assumed to be heterogeneous in 1) two observed measures of ability (X_1 and X_2) which affect their labor market income, 2) unobserved discount factor and 3) unobserved random utility value of schooling. Labor market income is determined by individuals' ability and years of schooling and is assumed to follow the Mincer-type exponential distribution.

$$w_t = \exp\left(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \varepsilon_1\right),$$

where E is the years of education taking values of 10, 12, 14, and 16 and ε_1 is normal, i.i.d., across individuals and periods with standard deviation of σ_1. Once an individual enters labor market and starts working, going back to school is not permitted. In each period t, the utility is given by U_{1t} if the individual works, and U_{2t} if he attends school. Also, we assume that the individual observes the labor income shock only after he enters the labor market and, therefore, the expected value of the wage only enters the utility function. This set-up yields the following two utilities corresponding to entering the labor market and attending school:

$$\begin{aligned} U_{1t} &= \mathbf{E}\left(w\right) = \exp\left(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \frac{1}{2}\sigma_1^2\right) \\ U_{2t} &= \gamma_1 1\{\text{in high school}\} + \gamma_2 1\{\text{in two-year college}\} + \gamma_3 1\{\text{in four-year college}\} + \varepsilon_2, \end{aligned}$$

where E_t denotes the years of education received up to t, so that

$$E_{t+1} = E_t + 1\{\text{schooling is chosen at } t\}.$$

Here γ_1 is the average utility of attending high school (we assume that there's no tuition for attending high school), γ_2 the average utility of attending two year college including tuition cost, γ_3 the average utility of attending four year college including tuition cost, ε_2 is mean zero and normally distributed individual specific random effect on schooling utility which is independent across individuals, but is fixed over time for each individual. The standard deviation of ε_2 is denoted by σ_2.

We assume that people have different discount factor β, which is correlated with observed X_3 and specified as

$$\beta = \varepsilon_3 + \rho_0 + \rho_1 X_3$$

where ε_3 is normally distributed with mean 0 and standard deviation σ_3 and does not change

over time for each individual. The errors $\varepsilon_1, \varepsilon_2$, and ε_3 are independent.

Let $U(E = a)$ be the discounted utility from schooling choice $E = a$ at the beginning of life cycle. Given that working is an absorbing state, we can represent this multi period dynamic programming model in the following 4-choice static model:

$$\begin{aligned}
U(E &= 10) = \sum_{t=1}^{50} \beta^{t-1} e^{\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\alpha_3\times 10+\frac{1}{2}\sigma_1^2} \\
U\left(E=12\right) &= \sum_{t=1}^{2} \beta^{t-1}\left(\gamma_1+\varepsilon_2\right)+\sum_{t=3}^{50} \beta^{t-1} e^{\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\alpha_3\times 12+\frac{1}{2}\sigma_1^2} \\
U\left(E=14\right) &= \sum_{t=1}^{2} \beta^{t-1}\left(\gamma_1+\varepsilon_2\right)+\sum_{t=3}^{4} \beta^{t-1}\left(\gamma_2+\varepsilon_2\right) \\
&\quad +\sum_{t=5}^{50} \beta^{t-1} e^{\alpha_0+\alpha_1 X_1+\alpha_2 X_2+\alpha_3\times 14+\frac{1}{2}\sigma_1^2} \\
U\left(E=16\right) &= \sum_{t=1}^{2} \beta^{t-1}\left(\gamma_1+\varepsilon_2\right)+\sum_{t=3}^{6} \beta^{t-1}\left(\gamma_3+\varepsilon_2\right) \\
&\quad +\sum_{t=7}^{50} \beta^{t-1} e^{\alpha_0+\alpha_1 X+\alpha_2 X_2+\alpha_3\times 16+\frac{1}{2}\sigma_1^2}.
\end{aligned}$$

Given the model structure, we expect people with higher ability X_1 and X_2, higher discount factor β and higher utility value of schooling ε_2 to attain a higher level of schooling.

We assume that the econometrician observes the ability measures X_1 and X_2, the schooling outcome, and characteristics X_3 that affect discount factor. Discount factor β and the utility value of schooling ε_2 are not observed. Given that there is no selection into the labor market, the parameters in the wage equation are simply estimated by a linear regression of wage on observable characteristics. Given the purpose of this simulation excercise, we assume that the parameters in the wage equation are known and focus only on the parameters in the schooling utility and the parameters in the discount factor. Hence the parameters of interest in this exercise is as follows:

$$\begin{aligned}
\text{schooling utility parameters :} &\qquad \gamma_1, \gamma_2, \gamma_3, \sigma_2 \text{ and} \\
\text{discount factor parameters :} &\qquad \rho_0, \rho_1, \sigma_3.
\end{aligned}$$

For estimation, we consider our TSF-MLE, simulated MLE following Lerman and Manski (1981)'s proposal, McFadden (1989)'s smoothed SMLE, and the method of simulated moments of McFadden (1989). Our comparison is not exhaustive, but we believe that the simulated MLE of Lerman and Manski (1981) and the MSM of McFadden (1989) are the

most common approach that an empirical researcher adopts in this type of models. Note that we cannot apply simulation methods that involve GHK simulators because the unobserved heterogeneity in discount factor is nonlinear in the latent process. The moment conditions that we consider for MSM are as follows:

$$\begin{aligned}\sum_i (D_j - p_j(X_i;\theta)) \times 1 &= 0,\ j = 2,3,4 \\ \sum_i (D_j - p_j(X_i;\theta)) \times X_k &= 0,\ j = 2,3,4,\ k = 1,2,3.\end{aligned}$$

The sample size was chosen among $\{100, 200, 500, 1000\}$ and the simulation number from $\{10, 20, 50, 100\}$. When the simulation number was equal to or greater than 100, the comparison was not much informative as most estimators perform well in our data generating process. The Monte-Carlo simulation number was set to be 1000.

6.2 TSF-MLE, Lerman-Manski SMLE, and Smoothed SMLE

This section compares the performance of our estimator, the Lerman-Manski's procedure, and smoothed SMLE. The Lerman-Manski procedure uses simulated frequencies to compute simulated choice probabilities. To prevent the zero-probability problem, we substituted $0.5/R$ for simulated probabilities that turned out to be zero. The second kind is a smoothed SMLE which is computed by using the following smoothed simulated choice probability:

$$p_{j,R} = \frac{1}{R}\sum_{r=1}^{R} \frac{\exp(U_{j,r}/\lambda)}{\Sigma_{j=1}^{J}\exp(U_{j,r}/\lambda)}.$$

Here the parameter λ is a smoothing parameter, larger values indicating more smoothing, and $U_{j,r}$ denotes the simulated value function of choice j at the r-th simulation. The smoothing parameter chosen from $\{0.1, 0.01\}$ performed relatively better than other choices. The results are reported in Tables 1-4.

Table 1 compares the overall simulation errors in terms of the log-likelihood evaluation of the simulation-based estimator using the true log-likelihood $l_n(\theta)$. This number is bounded by $l_n(\hat{\theta}_{MLE})$ with $\hat{\theta}_{MLE}$ denoting the MLE of θ_0. As the number is higher, the simulation-based estimator suffers from a smaller overall simulation error. First, note that the performance of the Lerman-Manski is different from smoothed SMLEs. The simulation results show that the use of smoothing does not improve the performance, and sometimes, even worsen the quality of the estimator.

When the sample size is small, the performance of Lerman-Manski's procedure and

Table 1: TSF-MLE and Lerman-Manski SMLEs: Log Likelihood

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	-1,001.3	-992.3	-985.9	-984.6
$n = 100$	Lerman-Manski	-1,015.9	-994.6	-986.0	-984.6
	Smoothed SMLE ($\lambda = 0.1$)	-1,018.3	-996.9	-985.6	-983.9
	Smoothed SMLE ($\lambda = 0.01$)	-1,015.7	-996.4	-985.4	-983.9
	TSF-MLE	-1,008.0	-1,002.1	-998.0	-997.0
$n = 200$	Lerman-Manski	-1,018.2	-1,003.8	-998.2	-997.3
	Smoothed SMLE ($\lambda = 0.1$)	-1,035.6	-1,000.6	-998.9	-997.4
	Smoothed SMLE ($\lambda = 0.01$)	-1,031.1	-1,012.6	-999.1	-997.3
	TSF-MLE	-1,008.7	-1,005.3	-1,002.9	-1,002.4
$n = 500$	Lerman-Manski	-1,019.0	-1,006.6	-1,003.9	-1,002.5
	Smoothed SMLE ($\lambda = 0.1$)	-1,048.8	-1,028.1	-1,007.6	-1,004.0
	Smoothed SMLE ($\lambda = 0.01$)	-1,043.8	-1,027.7	-1,007.7	-1,004.1
	TSF-MLE	-1,007.8	-1,005.9	-1,004.5	-1,004.0
$n = 1000$	Lerman-Manski	-1,018.1	-1,006.8	-1,004.6	-1,004.2
	Smoothed SMLE ($\lambda = 0.1$)	-1,057.4	-1,040.3	-1,013.1	-1,006.5
	Smoothed SMLE ($\lambda = 0.01$)	-1,052.9	-1,037.7	-1,012.8	-1,006.5

Table 2: TSF-MLE and Lerman-Manski SMLEs: MAE of Estimated Standard Deviation of Discount Factor $(\times 100)$, $(\sigma_3 = 0.02.)$

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	0.53	0.47	0.42	0.40
$n = 100$	Lerman-Manski	0.57	0.51	0.43	0.40
	Smoothed SMLE ($\lambda = 0.1$)	0.66	0.54	0.42	0.40
	Smoothed SMLE ($\lambda = 0.01$)	0.68	0.55	0.43	0.40
	TSF-MLE	0.39	0.35	0.31	0.28
$n = 200$	Lerman-Manski	0.44	0.36	0.29	0.27
	Smoothed SMLE ($\lambda = 0.1$)	0.55	0.44	0.31	0.28
	Smoothed SMLE ($\lambda = 0.01$)	0.52	0.42	0.31	0.27
	TSF-MLE	0.26	0.22	0.18	0.17
$n = 500$	Lerman-Manski	0.37	0.25	0.19	0.18
	Smoothed SMLE ($\lambda = 0.1$)	0.55	0.47	0.28	0.20
	Smoothed SMLE ($\lambda = 0.01$)	0.47	0.47	0.27	0.20
	TSF-MLE	0.21	0.18	0.15	0.14
$n = 1000$	Lerman-Manski	0.36	0.22	0.16	0.14
	Smoothed SMLE ($\lambda = 0.1$)	0.52	0.54	0.32	0.19
	Smoothed SMLE ($\lambda = 0.01$)	0.49	0.53	0.31	0.19

Table 3: TSF-MLE and Lerman-Manski SMLEs: MAE of the estimator of $\rho_1 = 0.02$ $(\times 100)$.

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	0.57	0.52	0.48	0.47
$n = 100$	Lerman-Manski	0.67	0.52	0.47	0.46
	Smoothed SMLE ($\lambda = 0.1$)	0.55	0.47	0.43	0.41
	Smoothed SMLE ($\lambda = 0.01$)	0.54	0.47	0.42	0.41
	TSF-MLE	0.43	0.37	0.34	0.32
$n = 200$	Lerman-Manski	0.52	0.40	0.34	0.31
	Smoothed SMLE ($\lambda = 0.1$)	0.58	0.36	0.28	0.27
	Smoothed SMLE ($\lambda = 0.01$)	0.54	0.38	0.28	0.26
	TSF-MLE	0.28	0.24	0.22	0.20
$n = 500$	Lerman-Manski	0.38	0.26	0.21	0.20
	Smoothed SMLE ($\lambda = 0.1$)	0.69	0.44	0.20	0.17
	Smoothed SMLE ($\lambda = 0.01$)	0.61	0.43	0.21	0.17
	TSF-MLE	0.21	0.18	0.16	0.15
$n = 1000$	Lerman-Manski	0.31	0.19	0.16	0.15
	Smoothed SMLE ($\lambda = 0.1$)	0.84	0.59	0.20	0.13
	Smoothed SMLE ($\lambda = 0.01$)	0.76	0.55	0.20	0.13

Table 4: TSF-MLE and Lerman-Manski SMLEs: Utility for Attending High School ($\gamma_1 = 0$).

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	968	856	740	667
$n = 100$	Lerman-Manski	989	875	747	655
	Smoothed SMLE ($\lambda = 0.1$)	1,196	897	746	659
	Smoothed SMLE ($\lambda = 0.01$)	1,173	877	728	641
	TSF-MLE	702	613	515	465
$n = 200$	Lerman-Manski	762	654	530	468
	Smoothed SMLE ($\lambda = 0.1$)	1,270	778	524	464
	Smoothed SMLE ($\lambda = 0.01$)	1,240	764	533	463
	TSF-MLE	492	408	348	315
$n = 500$	Lerman-Manski	541	462	362	316
	Smoothed SMLE ($\lambda = 0.1$)	1,565	812	405	306
	Smoothed SMLE ($\lambda = 0.01$)	1,515	831	412	324
	TSF-MLE	338	302	256	243
$n = 1000$	Lerman-Manski	406	351	269	252
	Smoothed SMLE ($\lambda = 0.1$)	1,800	1,031	394	239
	Smoothed SMLE ($\lambda = 0.01$)	1,782	1,008	397	243

smoothed SMLEs becomes comparable to our methods, although sometimes, the former does not appear to perform inferior to our method. However, when the sample size is large, the improved performance of our estimator becomes prominent over that of the competing procedures. This fact remains unchanged even with smoothing. This confirms our theoretical result that our estimator is consistent even when the simulation number is small, but the Lerman-Manski's procedures and the smoothed SMLEs do not possess this property.

Tables 2-4 report the difference between the estimated utilities from the true ones in terms of mean absolute errors (MAE). A similar comparison among the estimators is made for the performance in estimating the utility parameters. While not reported here, we observed a similar patten of performance for other parameters.

6.3 TSF-MLE and MSM

The next results compare the performance of MSM and TSF-MLE. We consider two MSMs in this case. The first MSM does not use optimal weighting matrix and the second MSM does. The optimal weighting matrix used here is not the weighting matrix that ensures the efficiency of the estimator as equivalent to MLE. Since the latter is much more complicated to compute, we choose to use rather the usual optimal weighting matrix from the GMM. Both estimators are known to be $\sqrt{n}$-consistent for each finite simulation number. The results are reported in Tables 5-8.

Table 5 again compares the performance of MSM and TSF-MLE in terms of the log-likelihood evaluation. First, our estimator performs better than MSM that does not use optimal weighting matrix. This is true for most ranges of simulation numbers and sample sizes considered. Outperformance by our estimator becomes conspicuous in particular when the sample size is 100 and simulation number is 100. This appears to reflect the fact that our estimator becomes more like an MLE as the simulation number becomes large while MSM does not. When an optimal weighting matrix was used, the quality of MSM substantially improves, and sometimes outperforms our estimator, especially when the sample size is large and the simulation number is small. This may be due to the slower rate of convergence of our estimator than MSM estimators. However, it should be noted that the computation of MSM using the optimal weighting matrix involves the first step estimation of the parameters. Therefore, the direct comparison of these two estimators does not appear to be fair as one can increase the simulation number of TSF-MLE taking advantage of its fast computing time. In our simulations, the Lerman-Manski SMLE estimators and MSM that does not use an

Table 5: TSF-MLE and MSM: Log Likelihood

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 100$	TSF-MLE	-1,001.3	-992.3	-985.9	-984.6
	MSM w/o Optimal Weighting Matrix	-1,036.0	-1,033.1	-1,031.5	-1,027.9
	MSM w/ Optimal Weighting Matrix	-1,014.6	-1,006.6	-1,004.9	-1,003.3
$n = 200$	TSF-MLE	-1,008.0	-1,002.1	-998.0	-997.0
	MSM w/o Optimal Weighting Matrix	-1,029.2	-1,027.3	-1,026.6	-1,023.8
	MSM w/ Optimal Weighting Matrix	-1,001.8	-1,000.6	-999.7	-999.2
$n = 500$	TSF-MLE	-1,008.7	-1,005.3	-1,002.9	-1,002.4
	MSM w/o Optimal Weighting Matrix	-1,019.4	-1,018.0	-1,016.2	-1,014.9
	MSM w/ Optimal Weighting Matrix	-1,002.2	-1,001.8	-1,001.4	-1,001.3

Table 6: TSF-MLE and MSM: MAE of Estimated Standard Deviation of Discount Factor shock ($\times 100$), ($\sigma_3 = 0.02$.)

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 100$	TSF-MLE	0.53	0.47	0.42	0.40
	MSM w/o Optimal Weighting Matrix	0.88	0.87	0.87	0.85
	MSM w/ Optimal Weighting Matrix	0.70	0.68	0.66	0.63
$n = 200$	TSF-MLE	0.39	0.35	0.31	0.28
	MSM w/o Optimal Weighting Matrix	0.71	0.44	0.31	0.28
	MSM w/ Optimal Weighting Matrix	0.45	0.42	0.31	0.27
$n = 500$	TSF-MLE	0.26	0.22	0.18	0.17
	MSM w/o Optimal Weighting Matrix	0.55	0.47	0.28	0.49
	MSM w/ Optimal Weighting Matrix	0.47	0.47	0.27	0.23

Table 7: TSF-MLE and MSM: MAE of the estimator of $\rho_1 = 0.02$ $(\times 100)$.

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	0.57	0.52	0.48	0.47
$n = 100$	MSM w/o Optimal Weighting Matrix	0.61	0.60	0.59	0.59
	MSM w/ Optimal Weighting Matrix	0.56	0.55	0.54	0.53
	TSF-MLE	0.43	0.37	0.34	0.32
$n = 200$	MSM w/o Optimal Weighting Matrix	0.44	0.43	0.43	0.43
	MSM w/ Optimal Weighting Matrix	0.35	0.34	0.33	0.33
	TSF-MLE	0.28	0.24	0.22	0.20
$n = 500$	MSM w/o Optimal Weighting Matrix	0.28	0.28	0.27	0.27
	MSM w/ Optimal Weighting Matrix	0.21	0.20	0.20	0.20

Table 8: TSF-MLE and MSM: Utility for Attending High School $(\gamma_1 = 0)$.

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	968	856	740	667
$n = 100$	MSM w/o Optimal Weighting Matrix	836	807	790	785
	MSM w/ Optimal Weighting Matrix	946	913	894	879
	TSF-MLE	702	613	515	465
$n = 200$	MSM w/o Optimal Weighting Matrix	549	532	531	526
	MSM w/ Optimal Weighting Matrix	570	549	545	544
	TSF-MLE	492	408	348	315
$n = 500$	MSM w/o Optimal Weighting Matrix	325	321	322	318
	MSM w/ Optimal Weighting Matrix	346	336	330	330

optimal weighting matrix took approximately the same computation time as our estimator. The MSM estimator that uses an optimal weighting matrix took about twice as long.

7 Conclusion

In this paper we propose an alternative to the conventional simulated maximum likelihood estimator for discrete choice models that is consistent when the number of simulations is finite. This alternative approach involves a simple transform of simulated frequency and hence incurs no computational burden beyond that of the conventional simulated MLE. We have derived the estimator's convergence rate when the number of simulations is fixed and we have established a rate of the increase in simulation numbers that ensures the estimator's asymptotic equivalence with MLE. Monte Carlo simulation studies show that the performance of our estimator is satisfactory, dominating Lerman and Manski (1981)'s SMLE using simulated frequencies, and SMLE using smoothed simulated choice probabilities. MSM estimators are shown to perform better than SMLEs when the simulation number is small. However, the performance of MSM becomes inferior to SMLEs when the simulation number becomes large.

While the simulation bias is completely eliminated in our estimation method, the rate of convergence for finite R does not achieve the $\sqrt{n}$-rate. It would be interesting to extend the method of this paper so that the estimator may achieve the parametric rate. A research in this direction is in progress by the authors.

8 Appendix: Proofs of the Results

Throughout the proofs, the notation C denotes a constant that can take different values in different places.

Proof of Lemma 1 : The proof is algebraically straightforward. Indeed, we can write out $\lambda_k(p, p_0; T_R)$ as in (13) and check the condition (9). However, we offer a different proof which reveals the discovery process of the transforms $\{T_R^j\}$. First, we let $\{T_R^j\}$ take the following form:

$$T_R^j(m) = T_R(m_j, m_{-j}) \tag{12}$$

for some map T_R. Observe that

$$\Lambda_R(p, p_0; \{T_R^j\}) = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,J}} T_R^j(m) \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} {p_1}^{m_1} \cdots {p_J}^{m_J}.$$

Note that the derivative of Λ_R with respect to p_k at $p = p_0$ is

$$\begin{aligned}\lambda_k(p_0, p_0; \{T_R^j\}) &= \frac{\partial}{\partial p_k}\Lambda_R(p, p_0; \{T_R^j\})|_{p=p_0} \\ &= \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,J}} T_R^j(m) \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} m_k p_{10}^{m_1} p_{20}^{m_2} \cdots p_{k0}^{m_k - 1} \cdots p_{J0}^{m_J}.\end{aligned} \tag{13}$$

Let $c_k(m_1, \cdots, m_J)$ be the coefficient of $p_1^{m_1} ... p_J^{m_J}$ in the expansion of $\lambda_k(p, p_0; \{T_R^j\})$ as above. For brevity, put $p = p_0$ so that we write $\lambda_1(p) \equiv \lambda_1(p, p; \{T_R^j\})$ as

$$\lambda_1(p) = \sum_{j=1}^{J} \sum_{m \in \mathbb{N}_{R,J}} T_R^j(m) \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} m_1 p_1^{m_1 - 1} p_2^{m_2} ... p_{j-1}^{m_{j-1}} p_j^{m_j + 1} p_{j+1}^{m_{j+1}} \cdots p_J^{m_J}.$$

Let us compute $c_1(m_1, \cdots, m_J)$. Then it suffices to show that $c_j(m_1, \cdots, m_J)$ is the same for all $j = 1, \cdots, J$, or, without loss of generality, that

$$c_1(m_1, \cdots, m_J) = c_2(m_1, \cdots, m_J).$$

First observe that $c_2(m_1, m_2, \cdots, m_J) = c_1(m_2, m_1, \cdots, m_J)$ by the form of $\{T_R^j\}$ in (12) and Λ_R. Hence it suffices to show that

$$c_1(m_1, m_2, \cdots, m_J) = c_1(m_2, m_1, \cdots, m_J).$$

To show this, first note that

$$c_1(m_1, \cdots, m_J) = T_R^1(m) \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} m_1 + U_R \tag{14}$$

where

$$\begin{aligned}U_R &= \sum_{j=2}^{J} T_R(m_j - 1; m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \\ &\quad \times \begin{pmatrix} R \\ m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J \end{pmatrix} (m_1 + 1).\end{aligned}$$

The relation in (14) holds for any $(m_1, \cdots, m_J) \in \mathbb{N}_{R,J}$ and we can simply extend the domain of T_R to negative numbers by taking $T_R(m_j; m_{-j}) = 0$ if $m_j < 0$. By noting

$$\begin{pmatrix} R \\ m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J \end{pmatrix} (m_1 + 1) = \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} m_j,$$

we write the coeffient $c_1(m_1, \cdots, m_J)$ in (14) as

$$\begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} \left[m_1 T_R(m_1; m_{-1}) + \sum_{j=2}^{J} m_j T_R(m_j - 1; m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \right].$$

Since the factor in front of the above bracket does not depend on "1", it suffices to show that

$$\begin{aligned} & c_1\left(m_1, m_2, \cdots, m_J\right) \\ = \; & m_1 T_R\left(m_1; m_2, \cdots, m_J\right) + \sum_{j \neq 1} m_j T_R\left(m_j - 1; m_1 + 1, m_2, \cdots, m_J\right) \\ = \; & m_2 T_R\left(m_2; m_1, \cdots, m_J\right) + \sum_{j \neq 2} m_j T_R\left(m_j - 1; m_2 + 1, m_1, \cdots, m_J\right) \\ = \; & c_1(m_2, m_1, m_3 \cdots, m_J). \end{aligned}$$

By rearranging terms on both sides of the second equality, we obtain

$$\begin{aligned} & m_1\left[T_R\left(m_1; m_2, \cdots, m_J\right) - T_R\left(m_1 - 1; m_2 + 1, m_3, \cdots, m_J\right)\right] \\ & + \sum_{j=3}^{J} m_j\left[T_R\left(m_j - 1; m_1 + 1, m_2, \cdots, m_J\right) - T_R\left(m_j - 1; m_2 + 1, m_1, \cdots, m_J\right)\right] \\ = \; & m_2\left[T_R\left(m_2; m_1, m_3, \cdots, m_J\right) - T_R\left(m_2 - 1; m_1 + 1, m_3, \cdots, m_J\right)\right]. \end{aligned} \tag{15}$$

Therefore, the proof is complete once we show that the above equality is satisfied by our choice of (24). One can check this equality immediately by considering each case: $m_1 = m_2 = 0$ and $m_1, m_2 > 0$ and $m_1 > 0,\ m_2 = 0$ and finally $m_1 = 0,\ m_2 > 0$. However, here we take a different route, showing how the form of (24) was discovered. In the proof we generate sufficient conditions for the equality in (15). Then these sufficient conditions lead to the solution of (24).

Without loss of generality, we assume $m_1 \geq m_2$ and $m_3 \geq m_4 \geq \cdots \geq m_J$. If $m_1 = m_2 = 0$, the equality in (15) is trivially satisfied.

Case 1) $m_1, m_2 > 0$. Then, the condition (15) is satisfied if

$$m_1\left[T_R\left(m_1; m_2, \cdots, m_J\right) - T_R\left(m_1 - 1; m_2 + 1, m_3, \cdots, m_J\right)\right] = 1, \tag{16}$$

and

$$T_R\left(m_j - 1; m_1 + 1, m_2, \cdots, m_J\right) - T_R\left(m_j - 1; m_2 + 1, m_1, \cdots, m_J\right) = 0. \tag{17}$$

Restriction (17) implies that $T_R\left(m_1; m_2, ..., m_J\right)$ depends on $(m_2, \cdots, m_J)$ only through $\nu(m_2, \cdots, m_J)$, the number of non-zero elements from the non-choices $\{m_2, ..., m_J\}$. To see this, choose $(m_2', \cdots, m_J')$ such that $\nu(m_2', \cdots, m_J') = \nu(m_2, \cdots, m_J)$. Then, we can show that

$$T_R\left(m_1; m_2, ..., m_J\right) = T_R\left(m_1; m_2', ..., m_J'\right),$$

by repeating the process in (17) with adding and subtracting by 1 between two non-zero members from $\{m_2, ..., m_J\}$.

Therefore we write

$$T_R\left(m_1, m_2, ..., m_J\right) = T_R(m_1, \nu\left(m_2, ..., m_J\right)),$$

where ν denotes the number of non-zero elements in the non-choice set. Using the observation in (17), (16) can be re-written as

$$m_1\left[T_R\left(m_1; \nu\left(m_2, \cdots, m_J\right)\right) - T_R\left(m_1 - 1; \nu\left(m_2 + 1, m_3, \cdots, m_J\right)\right)\right] = 1, \tag{18}$$

and note that $\nu(m_2,\cdots,m_J)=\nu(m_2+1,m_3,\cdots,m_J)$. Hence we extract one condition for T_R that leads to (18):

$$T_R(m,\nu)-T_R(m-1,\nu)=\frac{1}{m}\text{ for all possible } m. \tag{19}$$

Case 2) $m_1>0$ and $m_2=0$. If further, $m_3=0$ then m_1 is simply R. In this case,

$$\begin{aligned} & m_1\left[T_R(m_1;m_2,\cdots,m_J)-T_R(m_1-1;m_2+1,m_3,\cdots,m_J)\right] \\ = \;& R\left[T_R(R;m_2=0,\cdots,m_J=0)-T_R(R-1;m_2+1,m_3=0,\cdots,m_J=0)\right]=0 \end{aligned} \tag{20}$$

or

$$T_R(R,0)=T_R(R-1,1). \tag{21}$$

If on the other hand $m_3>0$, we have from (9)

$$\begin{aligned} & m_1\left[T_R(m_1;m_2,\cdots,m_J)-T_R(m_1-1;m_2+1,m_3,\cdots,m_J)\right] \\ & +\sum_{j=3}^{J}m_j\left[T_R(m_j-1;m_1+1,m_2,\cdots,m_J)-T_R(m_j-1;m_2+1,m_1,\cdots,m_J)\right] \\ = \;& 0. \end{aligned}$$

By subtracting and adding back $T_R(m_1-1;m_2,m_3+1,\cdots,m_J)$, we can write the above equation as

$$\begin{aligned} & m_1\left[T_R(m_1;m_2,m_3,\cdots,m_J)-T_R(m_1-1;m_2,m_3+1,\cdots,m_J)\right] \\ = \;& m_1\left[T_R(m_1-1;m_2+1,m_3,\cdots,m_J)-T_R(m_1-1;m_2,m_3+1,\cdots,m_J)\right] \\ & +\sum_{j=3}^{J}m_j\left[T_R(m_j-1;m_2+1,m_1,\cdots,m_J)-T_R(m_j-1;m_1+1,m_2,\cdots,m_J)\right] \end{aligned} \tag{22}$$

Note that the left hand side in (22) is 1 by (18) and the difference in the number of non-zero elements in T_R for each difference term on the right-hand side is exactly 1. For example, $\nu(m_2+1,m_3,\cdots,m_J)=\nu(m_2,m_3+1,\cdots,m_J)+1$ and $\nu(m_2+1,m_1,\cdots,m_J)=\nu(m_1+1,m_2,\cdots,m_J)+1$. Therefore, if

$$T_R(m,\nu)-T_R(m,\nu-1)=c$$

for some c independent of m and ν, (22) is satisfied. In this case, (22) becomes

$$1=\sum_{j=1}^{J}cm_j=cR \text{ or } c=\frac{1}{R}.$$

Therefore, we extract a condition for (22):

$$T_R(m,\nu)-T_R(m,\nu-1)=\frac{1}{R} \tag{23}$$

for all m and ν. To summarize, conditions (19), (21), and (23) are sufficient for (15).

Now, if we define

$$T_R(m_j,m_{-j})=-\sum_{s=0}^{R-m_j-1}\frac{1}{R-s}+\frac{\nu(m_{-j})}{R}, \tag{24}$$

this choice of T_R satisfies conditions (19), (21), and (23), and hence the equation (15) follows, completing the proof. On the other hand, it is also worth noting that the conditions (19), (21), and (23) for T_R also lead to the form of (24) up to an affine transform. This is the way the transform T_R is determined. ■

Proof of Theorem 1 : We first consider the case of $J = 3$. Recall

$$\Lambda_R\left(p, p_0; \{T_R^j\}\right) = \sum_j p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) T_{Rj}\left(m_1, m_2, m_3\right) p_1^{m_1} p_2^{m_2} p_3^{m_3}$$

where we define $T_{Rj}\left(m_1, m_2, m_3\right) = T_R(m_j; m_{-j})$. We show that $\Lambda_R(p, p_0; \{T_R^j\})$ is globally (strictly) concave in $p \in S_J$. Then $\Lambda_R(p, p_0; \{T_R^j\})$ is uniquely maximized at $p = p_0$ and by Assumption 1(iii), we obtain the identification result.

Recall that λ_j denotes the derivative of $\Lambda_R(p, p_0; \{T_R^j\})$ with respect to p_j, so that

$$\begin{aligned} \lambda_1 - \lambda_3 &= \sum_j p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) T_{Rj}\left(m_1, m_2, m_3\right) \\ &\quad \times \left\{ m_1 p_1^{m_1 - 1} p_2^{m_2} p_3^{m_3} 1\{m_1 > 0\} - m_3 p_1^{m_1} p_2^{m_2} p_3^{m_3 - 1} 1\{m_3 > 0\} \right\}. \end{aligned}$$

By relabeling the terms (m_3 as $m_3 + 1$ and m_1 as $m_1 - 1$),

$$\begin{aligned} \lambda_3 &= \sum_j p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ m_1 - 1, m_2, m_3 + 1 \end{array} \right) T_{Rj}\left(m_1 - 1, m_2, m_3 + 1\right) (m_3 + 1) p_1^{m_1 - 1} p_2^{m_2} p_3^{m_3} 1\{m_1 > 0\} \\ &= \sum_j p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) T_{Rj}\left(m_1 - 1, m_2, m_3 + 1\right) m_1 p_1^{m_1 - 1} p_2^{m_2} p_3^{m_3} 1\{m_1 > 0\}. \end{aligned}$$

Hence the difference $\lambda_1 - \lambda_3$ is equal to $\sum_{m \in \mathbb{N}_{R,3}} B_R(m) p_1^{m_1 - 1} p_2^{m_2} p_3^{m_3}$, where

$$B_R(m) = \sum_{j=1}^{3} p_{j0} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) \left\{ T_{Rj}\left(m_1, m_2, m_3\right) - T_{Rj}\left(m_1 - 1, m_2, m_3 + 1\right) \right\} m_1 1\{m_1 > 0\}$$

However, by the definition of T_{Rj}, we have

$$\begin{aligned} & m_1 \left[T_{Rj}\left(m_1, m_2, m_3\right) - T_{Rj}\left(m_1 - 1, m_2, m_3 + 1\right) \right] \\ = \; & m_1 \times 1\{j = 1\} \left[\frac{1}{m_1} - \frac{1\{m_3 = 0\}}{R} \right] + m_1 \times 1\{j = 2\} \left[\frac{1\{m_1 = 1\}}{R} - \frac{1\{m_3 = 0\}}{R} \right] \\ & + m_1 \times 1\{j = 3\} \left[\frac{-1}{m_3 + 1} + \frac{1\{m_1 = 1\}}{R} \right]. \end{aligned}$$

Plugging this back into $B_R(m)$ we obtain

$$\begin{aligned} B_R(m) &= p_{10} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) \left[1 - 1\{m_3 = 0\} \frac{m_1}{R} \right] + p_{20} \left(\begin{array}{c} R - 1 \\ m_1 - 1, m_2, m_3 \end{array} \right) \left[1\{m_1 = 1\} - 1\{m_3 = 0\} \right] \\ & + p_{30} \left(\begin{array}{c} R \\ m_1 - 1, m_2, m_3 + 1 \end{array} \right) \left[-1 + 1\{m_1 = 1\} \frac{m_3 + 1}{R} \right]. \end{aligned}$$

Now, write the summand in $\lambda_1 - \lambda_3$:

$$B_R(m)p_1^{m_1-1}p_2^{m_2}p_3^{m_3} = \left\{ \frac{p_{10}}{p_1} \left(\begin{array}{c} R \\ m_1, m_2, m_3 \end{array} \right) p_1^{m_1}p_2^{m_2}p_3^{m_3} - p_{10} \left(\begin{array}{c} R-1 \\ m_1 - 1, m_2, 0 \end{array} \right) p_1^{m_1-1}p_2^{m_2} \right\} I(m_1 > 0)$$
$$+p_{20} \left(\begin{array}{c} R-1 \\ 0, m_2, m_3 \end{array} \right) p_2^{m_2}p_3^{m_3} - p_{20} \left(\begin{array}{c} R-1 \\ m_1 - 1, m_2, 0 \end{array} \right) p_1^{m_1-1}p_2^{m_2} I(m_1 > 0)$$
$$-\frac{p_{30}}{p_3} \left(\begin{array}{c} R \\ m_1 - 1, m_2, m_3 + 1 \end{array} \right) p_1^{m_1-1}p_2^{m_2}p_3^{m_3+1} I(m_1 > 0) + p_{30} \left(\begin{array}{c} R-1 \\ 0, m_2, m_3 \end{array} \right) p_2^{m_2}p_3^{m_3}.$$

Summing the above over $m \in \mathbb{N}_{R,3}$ and rearranging the terms, we obtain that $\lambda_1 - \lambda_3$ is equal to

$$\frac{p_{10}}{p_1} \left[1 - \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ 0, m_2, m_3 \end{array} \right) p_2^{m_2}p_3^{m_3} \right] - p_{10} (p_1 + p_2)^{R-1} + p_{20} (p_2 + p_3)^{R-1} - p_{20} (p_1 + p_2)^{R-1}$$
$$-\frac{p_{30}}{p_3} \left[1 - \sum_{m \in \mathbb{N}_{R,3}} \left(\begin{array}{c} R \\ m_1 - 1, m_2, 0 \end{array} \right) p_1^{m_1-1}p_2^{m_2} \right] + p_{30} (p_2 + p_3)^{R-1}$$

or

$$p_{10} \left[\frac{1}{p_1} \left(1 - (p_2 + p_3)^R \right) - (p_1 + p_2)^{R-1} \right] + p_{20} \left[(p_2 + p_3)^{R-1} - (p_1 + p_2)^{R-1} \right]$$
$$-p_{30} \left[\frac{1}{p_3} \left(1 - (p_1 + p_2)^R \right) - (p_2 + p_3)^{R-1} \right].$$

Using the fact that $p_1 + p_2 + p_3 = 1$ and $p_{10} + p_{20} + p_{30} = 1$, we find that the above becomes,

$$\begin{aligned} & \frac{p_{10}}{p_1} \left[1 - (1 - p_1)^R \right] + (1 - p_{10}) (1 - p_1)^{R-1} - \frac{p_{30}}{p_3} \left[1 - (1 - p_3)^R \right] - (1 - p_{30}) (1 - p_3)^{R-1} \\ = & \frac{p_{10}}{p_1} + \left(1 - \frac{p_{10}}{p_1} \right) (1 - p_1)^{R-1} - \frac{p_{30}}{p_3} - \left(1 - \frac{p_{30}}{p_3} \right) (1 - p_3)^{R-1}. \end{aligned}$$

Therefore, $\partial (\lambda_1 - \lambda_3) / \partial p_1$ is equal to

$$\lambda_{11} - \lambda_{31} = -\frac{p_{10}}{p_1^2} + \frac{p_{10}}{p_1^2} (1 - p_1)^{R-1} - \left(1 - \frac{p_{10}}{p_1} \right) (R - 1) (1 - p_1)^{R-2}$$

and by symmetry, $\partial (\lambda_3 - \lambda_1) / \partial p_3$ is equal to

$$\lambda_{33} - \lambda_{31} = -\frac{p_{30}}{p_3^2} + \frac{p_{30}}{p_3^2} (1 - p_3)^{R-1} - \left(1 - \frac{p_{30}}{p_3} \right) (R - 1) (1 - p_3)^{R-2}.$$

.

We also obtain that $\partial (\lambda_1 - \lambda_3) / \partial p_2 = \lambda_{12} - \lambda_{32} = 0$. Likewise, from

$$\lambda_2 - \lambda_3 = \frac{p_{20}}{p_2} + \left(1 - \frac{p_{20}}{p_2} \right) (1 - p_2)^{R-1} - \frac{p_{30}}{p_3} - \left(1 - \frac{p_{30}}{p_3} \right) (1 - p_3)^{R-1},$$

we obtain

$$\begin{aligned}
\lambda_{22}-\lambda_{32} &= -\frac{p_{20}}{p_2^2}+\frac{p_{20}}{p_2^2}\left(1-p_2\right)^{R-1}-\left(1-\frac{p_{20}}{p_2}\right)\left(R-1\right)\left(1-p_2\right)^{R-2}, \\
\lambda_{33}-\lambda_{32} &= -\frac{p_{30}}{p_3^2}+\frac{p_{30}}{p_3^2}\left(1-p_3\right)^{R-1}-\left(1-\frac{p_{30}}{p_3}\right)\left(R-1\right)\left(1-p_3\right)^{R-2}, \text{ and} \\
\lambda_{21}-\lambda_{31} &= 0.
\end{aligned}$$

Note that $\lambda_{13}-\lambda_{33}=\lambda_{23}-\lambda_{33}$. Now it suffices to show that the matrix

$$\begin{aligned}
& \begin{pmatrix} \lambda_{11}-\lambda_{31}-(\lambda_{13}-\lambda_{33}) & \lambda_{21}-\lambda_{31}-(\lambda_{23}-\lambda_{33}) \\ \lambda_{12}-\lambda_{32}-(\lambda_{13}-\lambda_{33}) & \lambda_{22}-\lambda_{32}-(\lambda_{23}-\lambda_{33}) \end{pmatrix} \\
= & \begin{pmatrix} \lambda_{11}-\lambda_{31} & \lambda_{21}-\lambda_{31} \\ \lambda_{12}-\lambda_{32} & \lambda_{22}-\lambda_{32} \end{pmatrix} - \begin{pmatrix} \lambda_{13}-\lambda_{33} & \lambda_{13}-\lambda_{33} \\ \lambda_{13}-\lambda_{33} & \lambda_{13}-\lambda_{33} \end{pmatrix} \\
= & \begin{pmatrix} \lambda_{11}-\lambda_{31} & 0 \\ 0 & \lambda_{22}-\lambda_{32} \end{pmatrix} - (\lambda_{13}-\lambda_{33}) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}
\end{aligned}$$

is negative definite. This matrix is the hessian matrix of $\Lambda_R(p,p_0;\{T_R^j\})$ under the restriction that $p\in S_J$. For this, it suffices to show that $\lambda_{11}-\lambda_{31}<0$ for all p_1 and p_{10}; when this condition is satisfied, we can obtain, by symmetry, $\lambda_{22}-\lambda_{32}<0$ and $\lambda_{13}-\lambda_{33}>0$ as well.

Note that $\lambda_{11}-\lambda_{31}$ is only a function of p_{10} and p_1. We want to show

$$\lambda_{11}-\lambda_{31}=-\frac{p_{10}}{p_1^2}+\frac{p_{10}}{p_1^2}\left(1-p_1\right)^{R-1}-\left(1-\frac{p_{10}}{p_1}\right)\left(R-1\right)\left(1-p_1\right)^{R-2}<0$$

The above is linear in p_{10} and hence bounded by the maximum over the two points, the first one with $p_{10}=1$ and the one with $p_{10}=0$. Therefore, it suffices to check these two extreme cases.

First, when $p_{10}=1$,

$$\begin{aligned}
\lambda_{11}-\lambda_{31} &= -\frac{1}{p_1^2}+\frac{1}{p_1^2}\left(1-p_1\right)^{R-1}+\frac{1}{p_1}\left(R-1\right)\left(1-p_1\right)^{R-1} \\
&= -\frac{1}{p_1^2}+\frac{1}{p_1^2}\left[1+p_1\left(R-1\right)\right]\left(1-p_1\right)^{R-1}.
\end{aligned}$$

However, note that $1+p_1\left(R-1\right)\leq\left(1+p_1\right)^{R-1}$ for any $R\geq 2$ and $p_1>0$ and strictly so when $R\geq 3$. Hence

$$\begin{aligned}
\lambda_{11}-\lambda_{31} &\leq -\frac{1}{p_1^2}+\frac{1}{p_1^2}\left(1+p_1\right)^{R-1}\left(1-p_1\right)^{R-1} \\
&\leq -\frac{1}{p_1^2}+\frac{1}{p_1^2}\left(1-p_1^2\right)^{R-1}<0.
\end{aligned}$$

Second, when $p_{10}=0$, trivially, $\lambda_{11}-\lambda_{31}=-\left(R-1\right)\left(1-p_1\right)^{R-2}<0$. Therefore, Λ is globally concave over $p\in S_J$ for any $p_0\in S_J$ when $J=3$.

Consider the case $J>3$. First, we get

$$\lambda_j-\lambda_k=\frac{p_{j0}}{p_j}+\left(1-\frac{p_{j0}}{p_j}\right)\left(1-p_j\right)^{R-1}-\frac{p_{k0}}{p_k}-\left(1-\frac{p_{k0}}{p_k}\right)\left(1-p_k\right)^{R-1}$$

for all $j,k=1,2,...,J$. Then it suffices to check the negative definiteness of the matrix

$$\begin{pmatrix} \lambda_{11}-\lambda_{J1} & 0 & ... & 0 \\ 0 & \lambda_{22}-\lambda_{J2} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_{J-1,J-1}-\lambda_{J,J-1} \end{pmatrix} - (\lambda_{1J}-\lambda_{JJ}) \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix}.$$

And as before, it suffices to show that $\lambda_{11}-\lambda_{J1}<0$ for all p_1 and p_{10} because then, by symmetry, $\lambda_{jj}-\lambda_{J,j}<0$ for all $j=1,2,...,J-1$. This can be proved exactly in the same way as before. ■

Proof of Lemma 2 : Fix $\varepsilon>0$, $\tilde{\theta}\in\Theta$ and choose $\theta\in B(\tilde{\theta},\varepsilon)$. By construction, the absolute difference $\left|\delta_j(X_i,\eta^*_{i,r};\tilde{\theta})-\delta_j(X_i,\eta^*_{i,r};\theta)\right|$ is equal to

$$\left|\prod_{k=1,k\neq j}^{J} g_{jk}(\tilde{\theta})-\prod_{k=1,k\neq j}^{J} g_{jk}(\theta)\right|, \tag{25}$$

where $g_{jk}(\theta)=1\left\{u(X_{ij},\eta_{ij};\theta)\geq u(X_{ik},\eta_{ik};\theta)\right\}$. Using the fact that $\sup_{\theta\in B(\tilde{\theta},\varepsilon)}\left|g_{jk}(\theta)-g_{jk}(\tilde{\theta})\right|\leq 1$ for all $k\in\{1,\cdots,J\}$, we bound the difference above by

$$C\max_{1\leq k\leq J}\sup_{\theta\in B(\tilde{\theta},\varepsilon)}\left|g_{jk}(\theta)-g_{jk}(\tilde{\theta})\right|$$

for some $C>0$. Define $\mu^D_{ijk}(\theta)=\mu(X_{ij},\theta)-\mu(X_{ik},\theta)$ and $h_{ijk}(\varepsilon)=\sup_{\theta\in B(\tilde{\theta},\varepsilon)}\left|\mu^D_{ijk}(\theta)-\mu^D_{ijk}(\tilde{\theta})\right|$. We deduce

$$\begin{aligned} & \mathbf{E}\left[\sup_{\theta\in B(\tilde{\theta},\varepsilon)}\left|g_{jk}(\theta)-g_{jk}(\tilde{\theta})\right|^2|X_i\right]\leq \mathbf{P}\left\{\mu^D_{ijk}(\tilde{\theta})-h_{ijk}(\varepsilon)\leq\eta_{ik}-\eta_{ij}\leq\mu^D_{ijk}(\tilde{\theta})+h_{ijk}(\varepsilon)|X_i\right\} \\ = \;& F_{ikj}(\mu^D_{ijk}(\tilde{\theta})+h_{ijk}(\varepsilon)|X_i)-F_{ikj}(\mu^D_{ijk}(\tilde{\theta})-h_{ijk}(\varepsilon)|X_i) \\ \leq \;& C\left(\sup_x\sup_\mu f_{ikj}(\mu|x)\right)\sup_{\theta\in B(\tilde{\theta},\varepsilon)}||\theta-\tilde{\theta}||\leq C\varepsilon, \end{aligned}$$

where $F_{ikj}(\cdot|X_i)$ is the conditional cdf of $\eta_{ik}-\eta_{ij}$ given X_i and $f_{ikj}(\cdot|X_i)$ its conditional density function. Hence Assumption 2(ii) is satisfied. ■

Proof of Theorem 2 : We first show the consistency of the estimator. Given the identification result, it suffices for consistency to show that

$$\sup_{\theta\in\Theta}\left|l^*_{n,R}(\theta)-l_R(\theta)\right|\to_p 0 \text{ as } n\to\infty,$$

where $l_R(\theta)=\mathbf{E}l^*_{n,R}(\theta)$. Since Θ is compact and for each $\theta\in\Theta$ we have

$$l^*_{n,R}(\theta)\to_p l_R(\theta),$$

by the Law of Large Numbers, it suffices to show the following stochastic equicontinuity condition, i.e., for

any $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that for each $\tilde{\theta} \in \Theta$,

$$\mathbf{P}\left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right| > \eta \right\} < \varepsilon.$$

Define T_R as in (24). Recall that we can write $l^*_{n,R}(\theta)$ as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_R(m^*_{ij}(\theta), m^*_{-ij}(\theta)) \\ = \quad & -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \left[\sum_{m=0}^{R} \frac{1\{m \leq R - m_{jR}(X_i, \eta^*_i; \theta) - 1\}}{R - m} \right] + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \frac{\nu(m^*_{-ij}(\theta))}{R}. \end{aligned}$$

Hence, $\left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right|$ is bounded by $C \sum_{j=1}^{J} 1\{|m_{jR}(X_i, \eta^*_i; \theta) - m_{jR}(X_i, \eta^*_i; \tilde{\theta})| \geq \eta\}$, for any small $\eta > 0$, because $m_{jR}(X_i, \eta^*_i; \theta)$ is an integer. (Note that C may depend on R.) Therefore, for each $\tilde{\theta} \in \Theta$,

$$\begin{aligned} & \mathbf{P}\left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right| > \eta \right\} \\ \leq \quad & J\mathbf{P}\left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \sup_{1 \leq j \leq J} \left| m_{jR}(X_i, \eta^*_i; \theta) - m_{jR}(X_i, \eta^*_i; \tilde{\theta}) \right| \geq \eta / (CJ) \right\}. \end{aligned}$$

However, note that $\mathbf{E}\left[\sup_{\theta \in B(\tilde{\theta}, \delta)} \left| m_{jR}(X, \eta; \theta) - m_{jR}(X, \eta; \tilde{\theta}) \right| \right]$ is bounded by

$$\sum_{r=1}^{R} \mathbf{E}\left[\sup_{\theta \in B(\tilde{\theta}, \delta)} \left| \delta_j(X_i, \eta^*_{i,r}; \tilde{\theta}) - \delta_j(X_i, \eta^*_{i,r}; \theta) \right| \right] \leq R\delta^{1/2}.$$

This yields the stochastic equicontinuity of the process $l^*_{n,R}(\tilde{\theta})$ and thereby, completes the proof for the consistency of $\hat{\theta}$.

Now we turn to the rate of convergence. Following the arguments used to prove Claim 1 in the proof of Theorem 3 below, we can show that

$$|\mathbf{E} l^*_{n,R}(\theta) - \mathbf{E} l^*_{n,R}(\theta_0)| \leq C||\theta - \theta_0||^2,$$

for some constant C. Hence, in view of Theorem 3.2.5 of van der Vaart and Wellner (1996), it suffices to investigate the continuity modulus of the process $\sqrt{n} l^*_{n,R}(\theta)$. Given our definition of T_R, the objective function $l^*_{n,R}(\theta)$ can be rewritten as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_j(X_i, \eta_i; \theta_0) h_R(p^*_{jR}(X_i, \theta), \nu(m^*_{-ij}(\theta))),$$

where $h_R(p,\nu) = -\frac{1}{R}\sum_{m=0}^{R-1} 1\{1-m/R > p\}/(1-(m/R)) + \nu/R$. In the meanwhile,

$$\begin{aligned} & \mathbf{E}\left[\sup_{\theta:||\theta-\theta_0||\le\delta}\left|h_R(p^*_{jR}(X_i,\theta),\nu(m^*_{-ij}(\theta))) - h_R(p^*_{jR}(X_i,\theta_0),\nu(m^*_{-ij}(\theta_0)))\right|^2\right] \qquad (26)\\ \le\ & CR\mathbf{P}\left\{\sup_{\theta:||\theta-\theta_0||\le\delta}\sup_{1\le r\le R}\left|\delta_j(X_i,\eta^*_{i,r};\theta)-\delta_j(X_i,\eta^*_{i,r};\theta_0)\right|\ge 1\right\}\\ \le\ & CR\mathbf{E}\left[\sup_{\theta:||\theta-\theta_0||\le\delta}\sup_{1\le r\le R}\left|\delta_j(X_i,\eta^*_{i,r};\theta)-\delta_j(X_i,\eta^*_{i,r};\theta_0)\right|^2\right]\le C\delta, \end{aligned}$$

by Assumption 2(ii). Let us define $\gamma_j(D,X,\eta;\theta) = Dh_R(m_j(X,\eta;\theta)/R, \nu(m_{-j}(X,\eta;\theta)))$ and $\mathcal{G} = \{\gamma(\cdot,\cdot,\cdot;\theta) : \theta\in\Theta\}$. From the proof of Theorem 3.1 in Chen, Linton, and van Keilegom (2003), the result of (26) gives us

$$\int_0^1 \sqrt{1+\log N_{[]}(\varepsilon||G||_2,\mathcal{G},||\cdot||_2)}d\varepsilon \le \int_0^1 \sqrt{1+\log N_{[]}((C\varepsilon||G||_2)^2,\Theta,||\cdot||)}d\varepsilon<\infty,$$

where G is an envelope of $\mathcal{G}$. We define $\mathcal{G}_\delta = \{\gamma_1-\gamma_2 : \gamma_1,\gamma_2\in\mathcal{G}, ||\gamma_1-\gamma_2||_2<\delta\}$. Then by the maximal inequality in terms of the bracketing entropy (e.g. Pollard (1989), van der Vaart (1996)), we have

$$\begin{aligned} & \mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\sqrt{n}\left|l^*_{n,R}(\theta)-l^*_{n,R}(\theta_0)-\mathbf{E}l^*_{n,R}(\theta)+\mathbf{E}l^*_{n,R}(\theta_0)\right|\right]\\ \le\ & C\int_0^1\sqrt{1+\log N_{[]}(\varepsilon||G_\delta||_2,\mathcal{G}_\delta,||\cdot||_2)}d\varepsilon||G_\delta||_2\le C||G_\delta||_2, \end{aligned}$$

where G_δ indicates the envelope of $\mathcal{G}_\delta$. The second inequality follows from the fact that

$$N_{[]}(\varepsilon||G_\delta||_2,\mathcal{G}_\delta,||\cdot||_2)\le N_{[]}(2\varepsilon||G||_2,\mathcal{G}-\mathcal{G},||\cdot||_2)$$

By the result of (26), we can take G_δ such that $||G_\delta||_2\le C\delta^{1/2}$, and deduce that the continuity modulus of $l^*_{n,R}(\theta)$ in θ turns out to be $O(\delta^{1/2})$. Now, following Kim and Pollard (1990) (e.g. see van der Vaart and Wellner (1996), p.323.), the rate of convergence r_n for $\hat\theta$ satisfies $r_n^{2-1/2}\le\sqrt{n}$. Hence $r_n\sim n^{1/3}$, yielding the result of the theorem.

Proof of Theorem 3 : Observe that by Assumption 2(ii),

$$\begin{aligned} & P\left\{\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)\right|>\varepsilon_p/2|X_i=x\right\} \qquad (27)\\ \le\ & \frac{2}{\varepsilon_p}\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)\right||X_i=x\right]\\ \le\ & \frac{2}{\varepsilon_p\sqrt{R}}\int_0^C\sqrt{1+\log N_{[]}((\varepsilon/C)^2,\Theta,||\cdot||)}d\varepsilon = O(R^{-1/2}). \end{aligned}$$

The last inequality uses Assumption 2(ii) and the maximal inequality of Pollard (1989). Hence, for sufficiently small $\delta>0$,

$$\begin{aligned} & P\left\{\inf_{\theta\in B(\theta_0,\delta)}p^*_{ij}(\theta)>\varepsilon_p/2|X_i\right\}\\ \ge\ & P\left\{\inf_{\theta\in B(\theta_0,\delta)}p_{ij}(\theta)>\varepsilon_p/2+\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)\right||X_i\right\}\\ \ge\ & P\left\{\varepsilon_p/2>\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)\right||X_i\right\}\to 1 \text{ almost everywhere,} \end{aligned}$$

because $P\left\{\inf_{\theta\in B(\theta_0,\delta)} p_{ij}(\theta) > \varepsilon_p | X_i\right\} = 1$ by Assumption 2(iii). Hence $\inf_{\theta\in B(\theta_0,\delta)} p^*_{ij}(\theta) > \varepsilon_p/2$ with conditional probability given X_i converges to one almost surely. We assume that $\inf_{\theta\in B(\theta_0,\delta)} p^*_{ij}(\theta) > \varepsilon_p/2$ for the rest of the proof.

Claim 1 : $\sup_{\theta\in B(\theta_0;M\delta)} \mathbf{E} l_{n,R}(\theta) - \mathbf{E} l_{n,R}(\theta_0) \leq C\log(R)\delta^2$.

Claim 2 : $\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)} \sqrt{n}\left|l_{n,R}(\theta) - l_{n,R}(\theta_0) - \mathbf{E} l_{n,R}(\theta) + \mathbf{E} l_{n,R}(\theta_0)\right|\right] \leq \mu_R(\delta)$, where

$$\mu_R(\delta) = C\left\{\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R} + \delta\right\} \times \left\{\sqrt{\log(R)} + \sqrt{\log(-\log\delta)}\right\} + C\sqrt{n}R^{-1}.$$

Combining the results, we can establish the $\sqrt{n}$-rate of convergence as we demonstrate now. Suppose we have shown Claims 1-2. Take a sequence $r_n = n^{1/2}$ and partition Θ into "shells" $R_{j,n} = \{\theta : 2^{j-1} < r_n||\theta - \theta_0|| \leq 2^j\}$ with j ranging over integers. For any $\eta, M > 0$, we have

$$\mathbf{P}\left\{r_n||\hat{\theta} - \theta_0|| > 2^M\right\} \leq \sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} \mathbf{P}\left\{\sup_{\theta\in R_{j,n}} l_{n,R}(\theta) - l_{n,R}(\theta_0) \geq \eta\right\} + \mathbf{P}\left\{2||\hat{\theta} - \theta_0|| \geq \eta\right\}. \tag{28}$$

The second probability on the right-hand side vanishes because $\hat{\theta}$ is consistent by Claim 1. For each $\theta \in R_{j,n}$, we have

$$l_R(\theta) - l_R(\theta_0) \leq \frac{C2^{2j-2}\log(R)}{r_n^2}$$

by Claim 1. By using Claim 2, the sum of probabilities on the right-hand side in (28) is bounded by

$$\begin{aligned}
&\sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} \mathbf{P}\left\{\sup_{\theta\in R_{j,n}} \|l_{n,R}(\theta) - \mathbf{E} l_{n,R}(\theta) - l_{n,R}(\theta_0) + \mathbf{E} l_{n,R}(\theta)\| \geq \frac{C2^{2j-2}\log(R)}{r_n^2}\right\} \\
\leq\ & C\sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} \frac{C\sqrt{n}}{2^{2j-2}\log(R)} \mu_R\left(\frac{2^{j-1}}{\sqrt{n}}\right) \\
\leq\ & \frac{Cn^{1/4}\sqrt{\log n}}{\sqrt{R\log(R)}} \times \sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} 2^{-3j/2} + \frac{C}{\sqrt{\log R}} \times \sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} 2^{-j+1} + \frac{C\sqrt{n}}{2^{2j-2}R\log(R)} \sum_{\substack{j\geq M \\ 2^j\leq \eta r_n}} 2^{-2j+2} \to 0,
\end{aligned}$$

as $M \to \infty$, because $n^{1/4}\sqrt{\log n}/\sqrt{R\log(R)} \leq Cn^{1/4}\sqrt{\log R}/\sqrt{R\log(R)} = Cn^{1/4}/\sqrt{R} < \infty$. Therefore, $\sqrt{n}$-consistency of the estimator $\hat{\theta}$ follows.

Having established the $\sqrt{n}$-consistency of $\hat{\theta}$, we establish the asymptotic normality of the estimator as follows. Define

$$\begin{aligned}
T^*_{ij}(\theta) &= T_R(m^*_{ij}(\theta), m^*_{-ij}(\theta)), \\
\Delta_{ij}(\theta) &= T^*_{ij}(\theta) - T^*_{ij}(\theta_0), \text{ and } V(X_i) = \sum_{j=1}^J D_{ij}\frac{\partial}{\partial\theta} p_j(X_i,\theta_0)/p_j(X_i,\theta_0).
\end{aligned}$$

We first show the following.

Claim 3 :

$$\begin{aligned} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left[D_{ij} \Delta_{ij}(\theta) - \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right) \right] \\ = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ \frac{D_{ij}(p_j(X_i,\theta) - p_j(X_i,\theta_0))}{p_j(X_i,\theta_0)} - \mathbf{E} \left[\frac{D_{ij}(p_j(X_i,\theta) - p_j(X_i,\theta_0))}{p_j(X_i,\theta_0)} \right] \right\} + o_P(1), \end{aligned} \tag{29}$$

uniformly over $\theta \in B(\theta_0, Mn^{-1/2})$.

Now, note that by the mean-value theorem,

$$\sum_{j=1}^{J} \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right) = \sum_{j=1}^{J} \psi_j(\theta_0)'(\theta - \theta_0) + \sum_{j=1}^{J} (\theta - \theta_0) \Omega_j(\theta_*)'(\theta - \theta_0)$$

where θ_* lies on the line segment between θ and θ_0 and $\psi_j(\theta) = \frac{\partial}{\partial \theta} \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right)$ and $\Omega_j(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right)$. However,

$$\begin{aligned} \sum_{j=1}^{J} \psi_j(\theta_0) &= \frac{\partial}{\partial \theta} \sum_{j=1}^{J} \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right) |_{\theta = -\theta_0} = \frac{\partial}{\partial \theta} \sum_{j=1}^{J} \mathbf{E} \left(D_{ij} (T_{ij}^*(\theta) - T_{ij}^*(\theta_0) \right) |_{\theta = -\theta_0} \\ &= \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \mathbf{E} \left(\Lambda_R(p_{ij}(\theta), p_{ij}(\theta_0)) - \Lambda_R(p_{ij}(\theta_0), p_{ij}(\theta_0)) \right) |_{\theta = -\theta_0} = 0 \end{aligned}$$

by the identification result in Theorem 1. Hence we have

$$\sum_{j=1}^{J} \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right) = \sum_{j=1}^{J} (\theta - \theta_0) \Omega_j(\theta_*)'(\theta - \theta_0).$$

Therefore, using the fact that $\mathbf{E} V(X_i) = 0$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left[D_{ij} \Delta_{ij}(\theta) - \mathbf{E} \left(D_{ij} \Delta_{ij}(\theta) \right) \right] = (\theta - \theta_0)' Z_n + o_P(n^{-1/2}), \tag{30}$$

uniformly over $\theta \in B(\theta_0, Mn^{-1/2})$, where $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(X_i)$. Now, we follow similar steps in the proof of Theorem 3.2.16 in van der Vaart and Wellner (1996). Combined with Claim 3, the result of (30) yields

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta_0) \\ = & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \Delta_{ij}(\hat{\theta}) = \frac{1}{2} (\hat{\theta} - \theta_0)' \Omega (\hat{\theta} - \theta_0) + \frac{1}{\sqrt{n}} (\hat{\theta} - \theta_0)' Z_n + o_P(n^{-1}). \end{aligned} \tag{31}$$

Similarly

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}T_{ij}^{*}(\theta_0 - n^{-1/2}\Omega^{-1}Z_n) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}T_{ij}^{*}(\theta_0) \qquad (32) \\ = \quad & -\frac{1}{2n}Z_n'\Omega^{-1}Z_n + o_P(n^{-1}). \end{aligned}$$

By the definition of $\hat{\theta}$, the left-hand side of (31) is larger than the left-hand side of (32). We subtract the second equation from the first equation to obtain

$$\frac{1}{2}(\hat{\theta} - \theta_0 + n^{-1/2}\Omega^{-1}Z_n)'\Omega(\hat{\theta} - \theta_0 + n^{-1/2}\Omega^{-1}Z_n) \geq -o_P(n^{-1}).$$

Since Ω is negative definite, we conclude that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \Omega^{-1}Z_n + o_P(1).$$

The wanted result follows by the usual CLT. The proof of the theorem is complete.

Proof of Claim 1 : First, observe that $l_R(\theta) - l_R(\theta_0)$ is less than or equal to

$$\frac{1}{2}\left[\sup_{\theta\in B(\theta_0,\delta)}(\theta-\theta_0)'\frac{\partial^2\mathbf{E}\left[\sum_{j=1}^{J} D_{ij}T_{ij}^{*}(\theta)\right]}{\partial\theta\partial\theta'}(\theta-\theta_0)\right].$$

Note that

$$\frac{\partial^2\mathbf{E}[D_{ij}T_{ij}^{*}(\theta)]}{\partial\theta\partial\theta'} = -\frac{1}{R}\sum_{m=0}^{R-1}\frac{\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}\left(D_{ij}1\{1-m/R > p_{jR}^{*}(\theta)\}\right)}{1-m/R} + \frac{1}{R}\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}[\nu(m_{ij}^{*}(\theta))]. \qquad (33)$$

We consider the first term. We can write it as

$$\begin{aligned} & -\frac{1}{R}\sum_{m=0}^{R-1}\frac{\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}\left(D_{ij}1\{R-m > \sum_{m=1}^{R}\delta_j(X_i,\eta_{i,m}^{*};\theta)\}\right)}{1-m/R.} \qquad (34) \\ = \quad & \mathbf{E}\left[-\frac{p_j(X_i,\theta_0)}{R}\sum_{m=0}^{R-1}\frac{F_{R,p_j(X_i,\theta)}^{(1)}(R-m-1)}{1-m/R.}\frac{\partial^2 p_j(X_i,\theta)}{\partial\theta\partial\theta'}\right] \\ & +\mathbf{E}\left[-\frac{p_j(X_i,\theta_0)}{R}\sum_{m=0}^{R-1}\frac{F_{R,p_j(X_i,\theta)}^{(2)}(R-m-1)}{1-m/R.}\frac{\partial p_j(X_i,\theta)}{\partial\theta}\frac{\partial p_j(X_i,\theta)}{\partial\theta'}\right] \end{aligned}$$

where $F_{R,p_j(X_i,\theta)}^{(1)}(\cdot)$ and $F_{R,p_j(X_i,\theta)}^{(2)}(\cdot)$ are the first order and the second order derivatives of the binomial distribution function with parameter $(R, p_j(X_i,\theta))$. By Assumption 1(ii), we have $p(X_i,\theta) \in B(p(X_i,\theta_0), C(X_i)\delta)$ for all $\theta \in B(\theta_0,\delta)$ where $C(X_i)$ is square integrable and does not depend on $\theta \in B(\theta_0,\delta)$. By taking δ small, we have eventually $1 \geq p_j(X_i,\theta) > \varepsilon > 0$ for the constant ε_p in Assumption 2(iii) with large probability. For this $p(X_i,\theta)$, the derivatives $F_{R,p_j(X_i,\theta)}^{(1)}(\cdot)$ and $F_{R,p_j(X_i,\theta)}^{(2)}(\cdot)$ are bounded uniformly over $\theta \in B(\theta_0,\delta)$ with

large probability. We can bound the Euclidean norm of the first term in (33) by

$$\sup_{\theta\in B(\theta_0,\delta)}\sqrt{\mathbf{E}\left\|\frac{\partial^2 p_j(X_i,\theta)}{\partial\theta\partial\theta'}\right\|^2+C}\times\frac{1}{R}\sum_{m=0}^{R-1}\frac{1}{1-m/R}.$$

And note that

$$\frac{C}{R}\sum_{m=0}^{R-1}\frac{1}{1-m/R}\leq C\int_0^{1-1/R}\frac{1}{1-u}du+O(R^{-1})=\log(R)+O(R^{-1}).$$

Hence the first term on the right-hand side of (33) is $O(\log(R))$.

Now we consider the second term in (33). Note that

$$\begin{aligned}
\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}[\nu(m^*_{-ij}(\theta))] &= \sum_{k=1,k\neq j}^{J}\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}\left[\mathbf{P}\left\{\sum_{m=1}^{R}\delta_k(X_i,\eta^*_{i,m};\theta)>0|X_i\right\}\right]\\
&= \sum_{k=1,k\neq j}^{J}\frac{\partial^2}{\partial\theta\partial\theta'}\mathbf{E}\left[\left\{1-F_{R,p_k(X_i,\theta)}(0)\right\}\right]\\
&= \sum_{k=1,k\neq j}^{J}\mathbf{E}\left[-F^{(2)}_{R,p_k(X_i,\theta)}(0)\frac{\partial p_j(X_i,\theta)}{\partial\theta}\frac{\partial p_j(X_i,\theta)}{\partial\theta'}-F^{(1)}_{R,p_k(X_i,\theta)}(0)\frac{\partial^2 p_j(X_i,\theta)}{\partial\theta\partial\theta'}\right].
\end{aligned}$$

As argued before, we can take δ small so that for all $\theta\in B(\theta_0,\delta)$, $1>p(X_i,\theta)>\varepsilon>0$ for some ε. And this leads to the fact that $F^{(s)}_{R,p_j(X_i,\theta)}(0)$ and $F^{(s)}_{R,p_j(X_i,\theta)}(0)$, $s=1,2$, are bounded uniformly over $\theta\in B(\theta_0,\delta)$. Therefore, the Euclidean norm of the second term in (33) is again bounded by

$$\sup_{\theta\in B(\theta_0,\delta)}\sqrt{\mathbf{E}\left\|\frac{\partial^2 p_j(X_i,\theta)}{\partial\theta\partial\theta'}\right\|^2+C}\times\frac{1}{R}.$$

Hence we conclude $\sup_{\theta\in B(\theta_0,\delta)}\left[l_R(\theta)-l_R(\theta_0)\right]\leq C\log(R)\delta^2$.

Proof of Claim 2 : First, observe that for $p,p_0\in(\varepsilon_p/2,1)$,

$$\begin{aligned}
&\frac{1}{R}\sum_{m=0}^{R-1}\frac{1\{1-m/R>p\}-1\{1-m/R>p_0\}}{1-(m/R)} \qquad (35)\\
=\ &\int_{1-p_0}^{1-p}\frac{1}{1-u}du+O\left(R^{-1}\right)\\
=\ &\log(p)-\log(p_0)+O\left(R^{-1}\right)=\frac{p-p_0}{p_0}+O(|p-p_0|^2+R^{-1}).
\end{aligned}$$

Let $\tilde{T}^*_{ij}(\theta)=\log(p^*_{ij}(\theta))-\log(p^*_{ij}(\theta_0))$. Since $\inf_{\theta\in B(\theta_0,\delta)}p^*_{ij}(\theta)>\varepsilon_p/2$, we deduce that

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\left(D_{ij}T^*_{ij}(\theta)-D_{ij}T^*_{ij}(\theta_0)-\mathbf{E}\left[D_{ij}T^*_{ij}(\theta)-D_{ij}T^*_{ij}(\theta_0)\right]\right)\right|\right]\\
=\ &\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{J}\left(D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}T^*_{ij}(\theta_0)-\mathbf{E}\left[D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}\tilde{T}^*_{ij}(\theta_0)\right]\right)\right|\right]\\
&+O(\sqrt{n}R^{-1}).
\end{aligned}$$

We focus on the last expectation. Observe that

$$\begin{aligned}
&\sqrt{\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}\tilde{T}^*_{ij}(\theta_0)\right|^2\right]} \leq C\sqrt{\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p^*_{ij}(\theta_0)\right|^2\right]}\\
\leq\ & C\sqrt{\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)-\{p^*_{ij}(\theta_0)-p_{ij}(\theta_0)\}\right|^2\right]}\\
&+C\sqrt{\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p_{ij}(\theta)-p_{ij}(\theta_0)\right|^2\right]}.
\end{aligned}$$

Using Theorem 2.14.5 of van der Vaart and Wellner (1996), the leading term is bounded by

$$C\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)-\{p^*_{ij}(\theta_0)-p_{ij}(\theta_0)\}\right|\right]+C\delta^{1/2}/\sqrt{R}.$$

As for the leading expectation above, we proceed similarly as in (27):

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|p^*_{ij}(\theta)-p_{ij}(\theta)-\{p^*_{ij}(\theta_0)-p_{ij}(\theta_0)\}\right|\,|X_i=x\right]\\
\leq\ & \frac{C}{\sqrt{R}}\int_0^{C\delta^{1/2}}\sqrt{1+\log N_{[]}((\varepsilon/C)^2,\Theta,||\cdot||)}d\varepsilon \leq C\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R},
\end{aligned}$$

using Assumption 2(ii) and the maximal inequality. Therefore,

$$\sqrt{\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}\tilde{T}^*_{ij}(\theta_0)\right|^2\right]} \leq C\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R}+C\delta.$$

This inequality reveals both a bound for an envelope for the class of functions indexing the empirical process in Claim 2 and the local uniform L_2-continuity condition for this process. (e.g. Chen, Linton, and van Keilegom (2003).) Using the maximal inequality and after some algebra,

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\theta\in B(\theta_0,\delta)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{j=1}^J\left(D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}T^*_{ij}(\theta_0)-\mathbf{E}\left[D_{ij}\tilde{T}^*_{ij}(\theta)-D_{ij}\tilde{T}^*_{ij}(\theta_0)\right]\right)\right|\right]\\
\leq\ & C\int_0^{C\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R}+C\delta}\sqrt{1-C\log\{\varepsilon\sqrt{R}/\sqrt{-\log\delta}\}}d\varepsilon\\
&+C\int_0^{C\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R}+C\delta}\sqrt{1-C\log\varepsilon}d\varepsilon\\
\leq\ & C\left\{\delta^{1/2}\sqrt{-\log\delta}/\sqrt{R}+\delta\right\}\times\left\{\sqrt{\log(R)}+\sqrt{\log(-\log\delta)}\right\}.
\end{aligned}$$

Proof of Claim 3 : Let $\delta_n = Mn^{-1/2}$ for some large M. Define $Y_i = (X_i, \eta_{i,1}, \cdots, \eta_{i,R})$ and

$$g_R(Y_i,\theta) = \frac{p^*_{ij}(\theta)-p^*_{ij}(\theta_0)}{p^*_{ij}(\theta_0)} - \frac{p_{ij}(\theta)-p_{ij}(\theta_0)}{p_{ij}(\theta_0)}.$$

Since we have (similarly as in the proof of Claim 2)

$$\sup_{\theta\in B(\theta_0,\delta_n)}|p^*_{ij}(\theta)-p^*_{ij}(\theta_0)| = O_P(\delta_n^{1/2}R^{-1/2}\sqrt{-\log\delta_n}+\delta_n) = O_P(\delta_n\sqrt{-\log\delta_n})$$

we write (using (35)),

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{ij}\Delta_{ij}(\theta)-\mathbf{E}\left(D_{ij}\Delta_{ij}(\theta)\right)\right]\\
=\;&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{D_{ij}\left[\frac{p_{ij}(\theta)-p_{ij}(\theta_0)}{p_{ij}(\theta_0)}\right]-\mathbf{E}\left(D_{ij}\left[\frac{p_{ij}(\theta)-p_{ij}(\theta_0)}{p_{ij}(\theta_0)}\right]\right)\right\}\\
&+D_n(\theta)+O_P(\delta_n^2(-\log\delta_n))+O_P(\sqrt{n}R^{-1})
\end{aligned}$$

where

$$D_n(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{ij}g_R(Y_i,\theta)-\mathbf{E}\left(D_{ij}g_R(Y_i,\theta)\right)\right].$$

Now, we show that $D_n(\theta)=o_P(1)$ uniformly over $\theta\in B(\theta_0,\delta_n)$. Then the proof is complete.

Define $\mathcal{G}_R(\delta_n)=\{g_R(\cdot,\theta):\theta\in B(\theta_0,\delta_n)\}$. Then note that

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}|g_R(Y_i,\theta)-g_R(Y_i,\theta_0)|^2\right]\\
=\;&\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{\left(p_{ij}^*(\theta)-p_{ij}^*(\theta_0)\right)p_{ij}(\theta_0)-\left(p_{ij}(\theta)-p_{ij}(\theta_0)\right)p_{ij}^*(\theta_0)}{p_{ij}^*(\theta_0)p_{ij}(\theta_0)}\right|^2\right]\\
\le\;&\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{\left(p_{ij}(\theta_0)-p_{ij}^*(\theta_0)\right)\left(p_{ij}(\theta)-p_{ij}(\theta_0)\right)}{p_{ij}^*(\theta_0)p_{ij}(\theta_0)}\right|^2\right]\\
&+\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{\left(p_{ij}^*(\theta)-p_{ij}^*(\theta_0)-\left(p_{ij}(\theta)-p_{ij}(\theta_0)\right)\right)p_{ij}(\theta_0)}{p_{ij}^*(\theta_0)p_{ij}(\theta_0)}\right|^2\right].
\end{aligned}$$

Since we have $\sup_{\theta\in B(\theta_0,\delta_n)}\left\|p_{ij}^*(\theta)-p_{ij}(\theta)\right\|^2=O_P(R^{-1})$, the first term is bounded by $CR^{-1}\delta_n^2$. Define $d_{i,r}^*(\theta)=\delta(X_i,\eta_{i,r}^*,\theta)-\delta(X_i,\eta_{i,r}^*,\theta_0)$. Then by Theorem 2.14.5 in van der Vaart and Wellner (1996),

$$\begin{aligned}
&\left(\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{1}{R}\sum_{r=1}^{R}d_{i,r}^*(\theta)-\mathbf{E}\left[d_{i,r}^*(\theta)|X_i\right]\right|^2|X_i\right]\right)^{1/2}\\
=\;&C\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{1}{R}\sum_{r=1}^{R}d_{i,r}^*(\theta)-\mathbf{E}\left[d_{i,r}^*(\theta)|X_i\right]\right||X_i\right]+O(R^{-1/2}\delta_n^{1/2})=O(R^{-1/2}\delta_n^{1/2}\sqrt{-\log\delta_n}).
\end{aligned}$$

Hence $\left(\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}|g_R(Y_i,\theta)-g_R(Y_i,\theta_0)|^2\right]\right)^{1/2}\le CR^{-1/2}\delta_n^{1/2}\sqrt{-\log\delta_n}$. From this it follows that

$$N_{[]}(\varepsilon,\mathcal{G}_R(\delta_n),||\cdot||_2)\le N_{[]}(C(\varepsilon R^{1/2}/\sqrt{-\log\delta_n})^2,\Theta,||\cdot||)\le C(\varepsilon^{-1}R^{-1/2}\sqrt{-\log\delta_n})^{2d}.$$

Now, we write

$$\begin{aligned}
\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}|D_n(\theta)|\right]&=\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta_n)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{ij}g_R(Y_i,\theta)-\mathbf{E}\left(D_{ij}g_R(Y_i,\theta)\right)\right]\right|\right]\\
&\le\int_0^{CR^{-1/2}\delta_n^{1/2}\sqrt{-\log\delta_n}}\sqrt{1+\log N_{[]}(\varepsilon,\mathcal{G}_R(\delta_n),||\cdot||_2)}d\varepsilon\to 0,
\end{aligned}$$

because $R^{-1/2}\delta_n^{1/2}\sqrt{-\log\delta_n} \leq R^{-1/2}\delta_n^{1/2}\sqrt{\log(R)} \to 0$. Therefore, $\sup_{\theta\in B(\theta_0,\delta_n)} |D_n(\theta)| = o_P(1)$. We conclude that

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[D_{ij}\Delta_{ij}(\theta) - \mathbf{E}\left(D_{ij}\Delta_{ij}(\theta)\right)\right] \\ = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{D_{ij}\left[\frac{p_{ij}(\theta)-p_{ij}(\theta_0)}{p_{ij}(\theta_0)}\right] - \mathbf{E}\left(D_{ij}\left[\frac{p_{ij}(\theta)-p_{ij}(\theta_0)}{p_{ij}(\theta_0)}\right]\right)\right\} + o_P(1). \end{aligned}$$

We obtain the wanted result. ■

References

[1] Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press, Cambridge, Massachusetts.

[2] Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

[3] Börsch-Supan, A., and V. A. Hajivassiliou (1993), "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics* 58, 347-368.

[4] Chen, X. and T. G. Conley (2001), "A New Semiparametric Spatial Model for Panel Time Series," *Journal of Econometrics* 105, 59-83.

[5] Chen, X., O. Linton, and I. van Keilegom (2003), "Estimation of Semiparametric Models When the Criterion Function is Not Smooth," *Econometrica* 71, 1591-1608.

[6] Geweke, J. (1989), "Efficient Simulation from the Multivariate Normal Distribution Subject to Linear Inequality Constraints and the Evolution of Constraint Probabilities," Discussion Paper, Duke University.

[7] Gourieroux, C. and A. Monfort (1997), *Simulation-Based Econometric Methods*, Oxford University Press.

[8] Hajivassiliou, V. A. (1990), "The Method for Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises," Discussion Paper 697, Cowles Foundation, Yale University.

[9] Hajivassiliou, V. A. (1993), "Simulation Estimation Methods for Limited Dependent Variable Models," in *Handbook of Statistics*, ii. G. S. Maddala, C. R. Rao, and H. Vinod (eds), North-Holland, Amsterdam, 519-543.

[10] Hajivassiliou, V. A. and D. L. McFadden (1998), "The Method of Simulated Scores for the Estimation of LDV Models," *Econometrica* 66, 863-896.

[11] Hajivassiliou, V. A. and P. Ruud (1994), "Estimation by Simulation," in *Handbook of Econometrics*, iv. C. Engle, and D. McFadden (eds.), North-Holland, Amsterdam.

[12] Keane, M. (1993), "Simulation Estimation for Panel Data Models with Limited Dependent Variable Models," in *Handbook of Statistics*, ii., G. S. Maddala, C. R. Rao, and H. Vinod (eds), North-Holland, Amsterdam, 545-570.

[13] Keane, M. (1994), "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica* 62, 95-116.

[14] Keane, M. and K. Wolpin (1997), "The Career Decisions of Young Men," *Journal of Political Economy* 105, 473-522.

[15] Kim, J. and D. Pollard (1990), "Cube Root Asymptotics," *Annals of Statistics* 18, 191-219.

[16] Lee, L-F. (1992), "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood," *Econometric Theory* 8, 518-552.

[17] Lee, L-F. (1995), "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models," *Econometric Theory* 11, 437-483.

[18] Lerman S. R., and C. Manski (1981), "On the Use of Simulated Frequencies to Approximate Choice Models," In C. Manski and D. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*, pp. 305-319, Cambridge, Massachusetts: MIT Press.

[19] Manski, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3, 205-228.

[20] McFadden, D. L. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zaremba, ed., *Frontiers in Econometrics*, pp. 105-142. New York: Academic Press.

[21] McFadden, D. L. (1989), "A Method of Simulated Moments of Estimation of Discrete Choice Models without Numerical Integration," *Econometrica* 57, 995-1026.

[22] McFadden, D. L. and P. Ruud (1994), "Estimation by Simulation," *Review of Economics and Statistics* 76, 591-608.

[23] Pakes, A. (1986), "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica* 54, 755-785.

[24] Pakes, A. and D. Pollard (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57, 1027-1057.

[25] Pollard, D. (1989), "A Maximal Inequality for Sums of Independent Processes under a Bracketing Condition," Unpublished manuscript.

[26] Sherman, P. (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica* 61, 123-137.

[27] Stern, S. (1992), "A Method of Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models," *Econometrica* 60, 943-952.

[28] Stern, S. (1997), "Simulation-based Methods," *Journal of Economic Literature* 35, 2006-2039.

[29] van der Vaart, A. (1996), "New Donsker classes," *Annals of Probability* 24, 2128-2140.

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

[31] Willis T. and S. Rosen (1979), "Education and Self-Selection," *Journal of Political Economy* 87, S7-S36.