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

Donghoon Lee[2] and Kyungchul Song[3]

July 25, 2012

Abstract

Many existing methods of simulated likelihood for discrete choice models require additive errors that have normal or extreme value distributions, with the prominent exception of the original simulated frequency method of Lerman and Manski (1981). This paper proposes a new method of simulated likelihood that is free from simulation bias for each finite number of simulation, and yet flexible enough to accommodate various model specifications beyond those of additive normal or logit errors. The method is flexible in the sense that it applies to almost any discrete choice model where individual choices can be simulated. The method begins with the likelihood function involving simulated frequencies and finds a transform of the likelihood function that identifies the true parameter for each finite simulation number. The transform 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. This paper presents and discusses results from Monte Carlo simulation studies of the new method.

Key words: Simulated MLE, Discrete Choice Models, Simulation Bias, Simulated Frequencies, Cube-Root Asymptotics

JEL Classifications: C12, C14, C52.

[1] A prevoius version was titled, "A Consistent SMLE When the Simulation Number was Finite." We are indebted to Sean D. Campbell, Joel Horowitz, and Eugene Savin for their valuable comments and inputs. We are also grateful to the seminar participants at Cemmap Seminar at University College London, Columbia University, Econometric Society Meeting, Greater New York Econometrics Colloquium, Lehigh University, University of Iowa, University of Rochester, and New York University for useful comments. The views expressed in this paper 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 British Columbia, 997-1873 East Mall, Vancouver, BC, Canada, V6T 1Z1.

1 Introduction

Discrete choice models have long been used in a wide range of empirical fields of economics. While a discrete choice model typically specifies the data generating process up to a parametric family of distributions, maximum likelihood estimation is infeasible in practice except for simple models because the explicit evaluation of the likelihood is not possible. Since the seminal work of Lerman and Manski (1981), the approach of simulation-based inference has been increasingly instrumental for overcoming this difficulty, providing the researcher with a wider spectrum of flexibility in modeling. (See Hajivassiliou and Ruud (1994), Stern (1997), Gouriéroux and Monfort (1997), and Train (2003) for a review of the literature and references therein.)

Most developments of methods of simulated likelihood have been made with a requirement that the original method of Lerman and Manski (1981) was free from: the assumption of additive normal or logit errors in the latent processes. For example, while the method of Stern (1992) and the method of GHK simulator (Geweke (1989), Hajivassiliou (1990) and Keane (1993)) are computationally efficient, these methods rely on the assumption that the random utilities involve additive normal errors. Hajivassiliou (1990) and Hajivassiliou and McFadden (1998) proposed a different method of simulated likelihood that uses simulated scores to construct simulated moment conditions and proved efficiency of the estimators. In particular Hajivassiliou and McFadden (1998) suggested three methods of maximum simulated scores (named MSS-SAR, MSS-SRC, and MSS-GRS in their paper). These methods also require that the random utilities have additive multivariate normal errors. Another increasingly popular class of discrete choice models include mixed multinomial logit models (MMLM) (Ben-Akiva, et. al. (1997) and McFadden and Train (2000) and see references therein). The MMLMs offer a flexible way of modeling heterogeneity through random coefficient specifications and yet requires the presence of additive logit errors.

Many structural econometric models used in labor economics and industrial organization do not admit such simple modeling of random utilities. In these models, unobserved heterogeneity in individual decision making often lies at the center of econometric modeling. Depending on the way one models the role of heterogeneity, one can easily encounter a structural model for which aforementioned simulation methods do not apply. (See Keane and Wolpin (1994), Keane and Wolpin (1997). etc.) In these situations, the original simulated frequency method of Lerman and Manski or their smoothed variants tend to emerge as the sole feasible solution. As is well-documented, however, the simulated frequency method of Lerman and Manski poses several difficulties such as discontinuity of the sample objective function, the zero probability problem, and the simulation bias due to the use of only a finite number of simulations.

This paper proposes a new method of simulated likelihood (MSL from here on) for discrete choices. Our method does not require additive normal or logit specification of random utilities and flexibly applies to all the models that the procedure of Lerman and Manski applies to. At the same time, the method is free from the zero-probability problem and does not accompany simulation bias for each finite simulation number. (See Lee (1995) for the asymptotic bias analysis of simulated discrete choice models and for a bias adjustment method.) The method is easy to implement, accompanying almost no additional computational cost beyond that of the simulated frequency method of Lerman and Manski. To the best of our knowledge, our method is the first simulated likelihood method that does not suffer from simulation bias for finite simulation numbers and yet allows for flexible modeling beyond that of normal or logit additive errors.

Our method is built on the main finding of this paper that there exists a simple and explicit transform of a simulated likelihood function whose maximization delivers a consistent estimator even with a finite simulation number. The transform is algebraically explicit, depending on no unknowns. Furthermore, the use of the transform does not require any restrictions on the specification of the random utilities, and hence flexibly applies to many discrete choice models that have a nonlinear, nonnormal form of heterogeneity. We call this new method *transformed simulated frequencies (TSF) method*. Our approach, however, shares one drawback of other simulation methods that use simulated frequencies, such as MSL of Lerman and Manski (1981) or methods of simulated moments of McFadden (1989): the sample objective function is discontinuous in parameters. The comparison of our method with other existing ones is summarized in Table 1.

Table 1: Comparison of Simulated MLE Methods

	Zero Prob. Prblm.	Sample Object. Func.	Addit. Normal/Logit Specification	Sim. Bias
Our Method (TSF)	No	Discontinuous	Not Required	No Bias
Lerman-Manski	Yes	Discontinuous	Not Required	Bias
Stern (1992)	No	Smooth	Required (Normal)	Bias
GHK	No	Smooth	Required (Normal)	Bias
MSS - SRC/GRS	No	Smooth	Required (Normal)	Bias
MSS - SAR	No	Discontinuous	Required (Normal)	No Bias
Mixed Multinomial Logit	No	Smooth	Required (Logit)	Bias

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. We also demonstrate how our framework can also be adapted to the case where only the cohort-level aggregate data are available under certain conditions. This set-up is relevant to some empirical studies in industrial organization.

Here is the summary of the asymptotic properties of the estimators based on the TSF method. When the simulation number is fixed and the sample size n increases, the estimator is consistent at the rate of $\sqrt[3]{n}$, like the maximum score estimator (Manski (1975) and Kim and Pollard (1990)). 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 faster than $\sqrt{n}$. Under this same condition, the estimator achieves the asymptotic efficiency of MLE.

To illustrate the usefulness of our approach, we performed Monte Carlo simulation studies based on two types of schooling choice models which involve heterogeneity in discount factor and ability. More specifically, the discount factor is assumed to be correlated with other observed individual characteristics and also an unobserved characteristic. In the second type of models, we assume time-varying heterogeneity so that the econometric model is a dynamic discrete choice model. The simulation methods considered in this study are, Lerman and Manski's MSL and its smoothed version, because these are the methods that are applicable in these models. Our estimator mostly dominates Lerman and Manski's simulation method and smoothed versions regardless of the simulation number. The domination is prominent especially when the simulation number is small and the sample size is large.

The remainder of this paper is organized as follows. In Section 2, we define the class of discrete choice models, discuss MSL, and offer a preview of our method. In Section 3, we present the main results of this paper which formally establish identification and consistency of the proposed estimator. It is also shown that the estimator is asymptotically normal when the simulation number goes to infinity fast enough. In Section 4, we present and discuss results from Monte Carlo simulation studies. Section 5 concludes. All the technical proofs are relegated to the appendix.

2 Discrete Choice Models and TSF

2.1 Methods of Simulated Likelihoods

Suppose that a binary decision variable, $D_{ij} \in \{0, 1\}$, of an agent i choosing the j-th choice, is stochastically linked with an observed covariate vector X_i as follows:

$$D_{ij} = \delta_j(X_i, \eta_i; \theta_0), \tag{1}$$

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

$$\delta_j(X_i, \eta_i; \theta_0) = 1\left\{ u_j(X_i, \eta_i; \theta_0) \geq \max_{k \neq j, 1 \leq k \leq J} u_k(X_i, \eta_i; \theta_0) \right\}, \tag{2}$$

where the function $u_j(X_i, \eta_i; \theta)$ is a random utility (McFadden (1974)).

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

$$p_j(X_i, \theta_0) = P\{D_{ij} = 1 | X_i\}.$$

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:

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

So far as $p_j(x, \theta)$ can be evaluated, 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 is large and one wants to admit flexibility in specifying the joint distribution of η_i.

Methods of simulated likelihood substitute a simulated choice probability for the choice probability to construct a simulated log-likelihood,

$$l^*_{n,R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \log p^*_{jR}(X_i, \theta). \tag{3}$$

The number R represents the repetition number of simulated stochastic variables. A simulated maximum likelihood estimator is defined as a maximizer of the simulated log-likelihood function,

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

When R increases with the sample size fast enough, it is known that for most choice probability simulators in the literature, the resulting estimator is consistent.

The original method of Lerman and Manski (1981) uses simulated frequency in constructing $p^*_{jR}(X_i,\theta)$. More specifically, suppose that R number of stochastic errors $\eta^*_{i,r},\ r=1,\cdots,R$, are drawn from the known distribution F of η_i. We let $\delta_j(X_i,\eta^*_{i,r};\theta),\ r=1,\cdots,R$, (taking values of 0 or 1) denote simulated choices for each value of θ. The simulated frequency of each choice j at simulation number R is defined to be

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

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

$$m_{ij}(\theta)=m_{jR}(X_i,\eta^*_i;\theta), \tag{4}$$

and $m_i(\theta)=(m_{1i}(\theta),\cdots,m_{Ji}(\theta))^\top$. Then the simulated choice probability is defined to be

$$p^*_{jR}(X_i,\theta)=\frac{m_{ij}(\theta)}{R}. \tag{5}$$

Plugging this into (3), one obtains the MSL estimator of Lerman and Manski.

This simulated frequency method of Lerman and Manski has been known to suffer from several drawbacks. Among which are the zero probability problem, discontinuity of the sample objective function in the parameters and simulation bias. First, the zero probability problem refers to the fact that for small R and large n, some of $p^*_{jR}(X_i,\theta)$ is likely to assume a zero value, causing a log-of-zero problem in the estimation. Second, the procedure involves a sample objective function that is discontinuous in the parameters. Finally, simulation bias occurs because the choice probability $p_j(X_i,\theta)$ is different from $p^*_{jR}(X_i,\theta)$. Many developments since Lerman and Manski (1981) have focused on overcoming the first two problems, i.e., zero probability problem and the problem of discontinuous objective functions (e.g. GHK method, MSS, and MMNL modeling mentioned in the introduction).

Nevertheless, it is important to note that the original method of Lerman and Manski (1981) can be applied in almost any discrete choice model where one can simulate the individual choices. For example, the method does not require that the random utilities have a certain form additive in normal or logit errors, making contrast with later developments in the literature. For researchers who use structural models that do not satisfy such requirements are often left with the method of Lerman and Manski, along with its disadvantages. This paper's proposal is an alternative that retains the modeling flexibility of the original procedure of Lerman and Manski, and yet is free from the two major disadvantages of Lerman and Manski's method: the zero probability problem and simulation bias. The flexibility of the method comes from the fact that like Lerman and Manski's method, this paper's method also applies to almost any discrete choice model where one can obtain the simulated frequencies of individual choices. The price to pay for this flexibility is that the sample objective function in our method, like that of Lerman and Manski, is discontinuous in the parameter.

2.2 Transformed Simulated Frequency (TSF)-Based Method

Our method begins first by attempting to overcome the problem of simulation bias of Lerman and Manski's simulated frequency method. Certainly, the fact that the simulated choice probabilities are unbiased estimators of the true choice probabilities does not help due to the presence of logarithm in (3). Instead of the logarithmic function, this paper proposes an alternative function that leads to an estimator entirely free from the simulation bias for each finite simulation number, as long as the sample size n is large enough. More specifically, for each fixed R, this paper develops a transform $T_{R,j}(\cdot)$ of simulated frequencies, $j = 1, 2, \cdots, J$, $R = 2, \cdots$, such that

$$\theta_0 = \arg\max_{\theta \in \Theta} \sum_{j=1}^{J} \mathbf{E}\left[D_{ij} T_{R,j}(m_i(\theta)) \right], \tag{6}$$

i.e., the population objective function identifies θ_0 for *each* $R = 2, \cdots$. Then for each R, an estimator of θ_0 is naturally obtained by maximizing its sample analogue:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{R,j}(m_i(\theta)). \tag{7}$$

A transform that satisfies (6) under regular conditions turns out to be of the following form: for each $j = 1, 2, \cdots, J$, and $R = 2, 3, \cdots$,

$$T_{R,j}(m) = -\sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{1}{R} \sum_{k=1,k\neq j}^{J} 1\{m_k > 0\}, \; m = (m_1, \cdots, m_J)^\top, \tag{8}$$

where m_j's are nonnegative integers such that $\sum_{j=1}^{J} m_j = R$. The remarkable aspect of the transform $T_{R,j}(\cdot)$ is that the transform does not depend on any unknown aspects of the data generating process. The transform depends only on J and R which are known. This means that we do not have to estimate the transform $T_{R,j}$ when one solves (7). We call $T_{R,j}(m_i(\theta))$ a *transformed simulated frequency* (*TSF*).

The main advantage of the TSF-based method comes from the fact that it relies only on the elementary method of simulated frequencies, and hence does not require a particular structure of random utilities. Furthermore, the TSF-based method is free from the zero probability problem by design. In other words, when the sample size is large, using only a finite simulation number does not cause a zero probability problem, in contrast to the simulated frequency method of Lerman and Manski. Some simulation results to be presented in the next subsection illustrate this point.

2.3 Illustration

To illustrate how the transform $T_{R,j}$ works, let us consider the following simple simulation example. We consider a binary choice model where the conditional choice probability of the first choice given $X \in \mathbf{R}$ is specified as

$$p(X; \theta_0) = \frac{1}{1 + \exp\left((10 + \theta_0)\, X\right)},$$

where X is drawn from the uniform distribution on $[-1, 1]$ and the true parameter θ_0 is set to be zero.

Figure 1 shows three population objective functions against different values of θ with different simulation numbers R:

$$\begin{aligned} Q_R^{TSF}(\theta) &= \sum_{j=1}^{J} \mathbf{E}\left[D_{ij} T_{R,j}(m_i(\theta))\right] : \text{ (TSF)} \\ Q_R^{L-M}(\theta) &= \sum_{j=1}^{J} \mathbf{E}\left[D_{ij} \log\left(\frac{m_{ij}(\theta)}{R}\right)\right] : \text{ (Lerman and Manski)} \\ Q^{MLE}(\theta) &= \sum_{j=1}^{J} \mathbf{E}\left[D_{ij} \log\left(p_j(X_i;\theta)\right)\right] : \text{ (MLE).} \end{aligned}$$

Each panel plots the three population objective functions over different simulation numbers R. Even when $R = 2$, the maximizers of $Q_R^{TSF}(\theta)$ and $Q^{MLE}(\theta)$ coincide, and this coincidence is maintained as R increases. When R is large, both the objective functions $Q_R^{TSF}(\theta)$ and $Q^{MLE}(\theta)$ coincide for all the values of θ. This makes contrast with the objective function of Lerman and Manski. When R is small, the objective function of Lerman and Manski's is away from the true value $\theta_0 = 0$. This reflects the well-known fact that the MSL estimator of Lerman and Manski is inconsistent for a finite R. Only when R becomes large, Lerman and Manski's objective function becomes close to the true MLE objective function.

Unlike Lerman and Manski's method, the approach of TSF does not suffer from the zero-probability problem. With each finite sample size n and finite simulation number R, TSF $T_{R,j}(m_i(\theta))$ always assumes a finite number regardless of the realizations of the simulated frequency $m_i(\theta)$. Hence the finite sample objective function is well defined regardless of the sample size and the simulation number. To illuminate this point, Figure 2 plots $\log(p)$, $\mathbf{E}\log(m_i(\theta_0)/R)$, and $\mathbf{E}T_{R,j}(m_i(\theta_0))$ against p, the choice probability, where the expected value is over the distribution of simulated errors when R is finite. Here the simulated frequencies $m_i(\theta_0)$ are generated according to the given value of the choice probability p. Certainly, in the case of Lerman and Manski, the zero-probability problem is severe when the simulation number is small, as shown by steeply falling curves as we move p close to zero. In contrast, the expected TSF does not suffer from this zero-probability problem. Furthermore, the expected TSF becomes close to $\log(p)$ more quickly than the expected logarithm of simulated probabilities as the simulation number increases.

Figure 1: Population Objective Functions: The objective function of TSF-based MSL has the same maximizer as that of MLE for each simulation number.

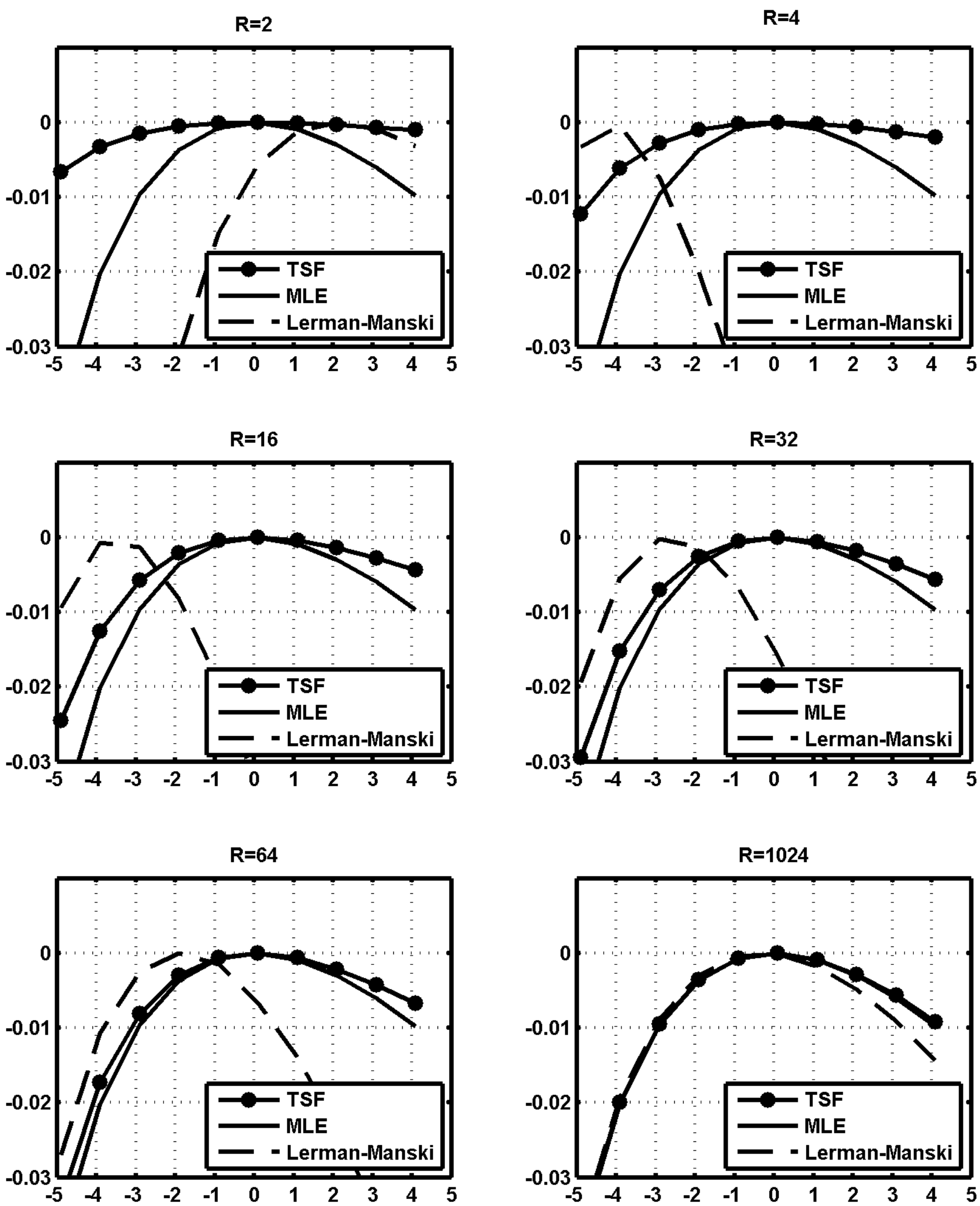
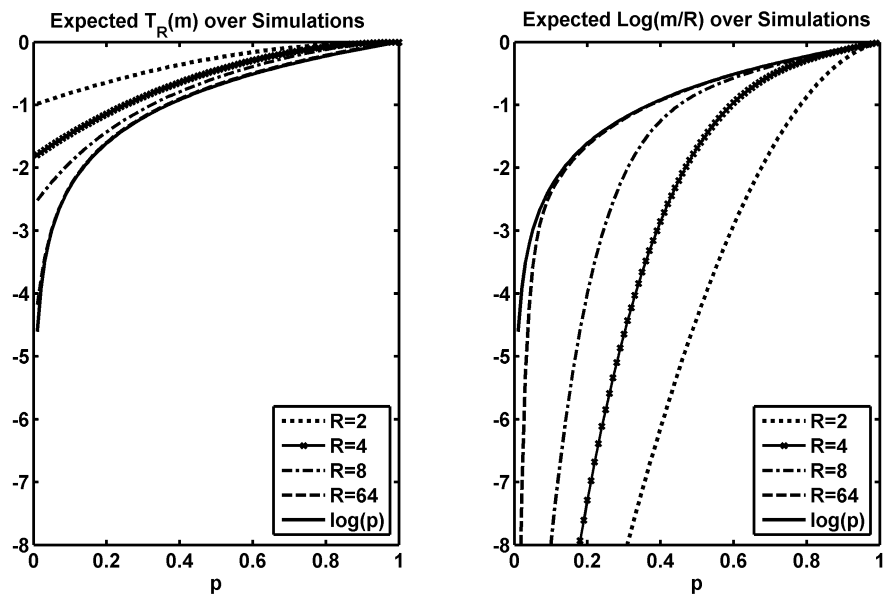


Figure 2: Value of Log of Simulated Probabilities (Lerman and Manski): The plots are against the given true choice probability p. The expected value of TSF is bounded from below when p is close to zero, whereas the expected value of log of simulated probabilities falls to $-\infty$.



3 Main Results

3.1 Identification

In this section, we provide the main result of this paper that the use of TSF in (8) identifies θ_0 for each finite simulation number R. Let $m_{ij}(\theta)$ be as defined in (4) and $\hat{\theta}$ be as defined in (7). We introduce the following regularity conditions. Let $\mathcal{X}$ be the support of X_i.

Assumption 1 : (i) (a) Θ is compact with a nonempty interior containing θ_0, and (b) for all $(\theta, x) \in \Theta \times \mathcal{X}$, $p_j(x;\theta)$ belongs to $S_J \in \{p \in [0,1]^J : \Sigma_j^J p_j = 1\}$.
(ii) For all $\theta \in \Theta$ such that $\theta \neq \theta_0$, there exists $j \in \{1, \cdots, J\}$ such that $P\{p_j(X_i;\theta_0) \neq p_j(X_i;\theta)\} > 0$.
(iii) For some $\delta > 0$, $0 < \varepsilon_p < \inf_{\theta \in B(\theta_0;\delta)} \inf_{x \in \mathcal{X}} p_j(x;\theta)$, $j = 1, \cdots, J$, for some $\varepsilon_p > 0$.
(iv) $(\eta_i)_{i=1}^n$ and $(\eta_{i,r}^*)_{i=1}^n$ are distributionally identical and $(\eta_{i,r}^*)_{i=1}^n$ is i.i.d. across $r = 1, 2, \cdots, R$.

Condition (i) requires that the choice probability function $p(x;\theta)$ is well-defined for all $\theta \in \Theta$ and Lipschitz continuous in $\theta \in \Theta$. Condition (ii) is a condition used to identify θ_0 from the identification of choice probabilities. Condition (iii) requires that the choice probabilities be bounded away from zero, and implies that $p_j(x;\theta_0) < 1 - \varepsilon_p$ for each $j = 1, \cdots, J$. Condition (iv) says that $(\eta_{i,r}^*)_{i=1}^n$ is independently and identically drawn from the distribution of $(\eta_i)_{i=1}^n$.

The following theorem is the main result of this paper.

Theorem 1 (Identification) : *Suppose that* Assumption 1 *holds. Then for each* $\delta > 0$, *and for each* $R = 2, 3, \cdots$,

$$\sum_{j=1}^{J} \mathbf{E}\left[D_{ij} T_{R,j}(m_i(\theta_0))\right] > \max_{\theta \in \Theta \setminus B(\theta_0;\delta)} \sum_{j=1}^{J} \mathbf{E}\left[D_{ij} T_{R,j}(m_i(\theta))\right],$$

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

The identification result in Theorem 1 tells us that in contrast to the use of logarithm, the use of $T_{R,j}$ leads to identification of θ_0 for *each finite* $R = 2, 3, \cdots$. This fact forms the basis on which we develop an MSL estimator that is consistent even if we have a finite simulation number R.

3.2 A Heuristic Exposition on the Discovery of TSF

In this section, we explain how the transform (8) is discovered. Let $\mathbb{N}_R = \{0, 1, 2, \cdots, R\}$ and define

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

The set $\mathbb{N}_{R,J}$ denotes the space of J-tuples where simulated frequencies $m_i(\theta)$ realize. Also we write the conditional choice probability $p_i(\theta) = p(X_i; \theta)$ for brevity. To find the right map $T_{R,j}$, we first focus on some necessary conditions that such a map should satisfy. Given a generic map $T_j(\cdot) : \mathbb{N}_{R,J} \to \mathbf{R}_+,\ j = 1, \cdots, J$, and $R = 2, 3, \cdots$, we introduce a function $\Lambda_R(p, p_0; T) : S_J \times S_J \to \mathbf{R}$, with $T = (T_1, \cdots, T_J)$, as follows:

$$\Lambda_R(p, p_0; T) = \sum_{j=1}^{J} p_{j0} \int T_j(M) dF_{R,p}(M),$$

where $F_{R,p}$ is the CDF of the multinomial distribution on $\mathbb{N}_{R,J}$ with parameter (R, p). The function Λ_R is uniquely determined once R and the transform T are chosen, and it does not depend on any other specifics of the data generating process of (D_i, X_i).

Using Λ_R, we rewrite

$$\sum_{j=1}^{J} \mathbf{E}\left[D_{ij} T_j(m_i(\theta))\right] = \mathbf{E}\left[\Lambda_R(p_i(\theta), p_i(\theta_0); T)\right]. \tag{9}$$

The main idea of this paper is that we extract conditions for Λ_R such that

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbf{E}\left[\Lambda_R(p_i(\theta), p_i(\theta_0); T)\right]. \tag{10}$$

Invoking the interchangeability of the derivative and the expectation, we write the first order condition for (10) as

$$\frac{\partial}{\partial \theta} \mathbf{E}\left[\Lambda_R(p_i(\theta), p_i(\theta_0); T)\right] |_{\theta=\theta_0} = \sum_{j=1}^{J} \mathbf{E}\left[\lambda_j(p_i(\theta_0), p_i(\theta_0); T) \frac{\partial p_j(X_i, \theta_0)}{\partial \theta}\right] = 0, \tag{11}$$

where for each $j = 1, 2, \cdots, J$,

$$\lambda_j(p, p_0; T) = \frac{\partial \Lambda_R(p, p_0; T)}{\partial p_j}. \tag{12}$$

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

at each θ implies

$$\sum_{j=1}^{J} \frac{\partial p_j(x,\theta)}{\partial \theta} = 0. \tag{13}$$

This means that the first order condition in (11) immediately follows if $\lambda_j(p,p;T)$ is the same across j's for all $p \in S_J$, i.e., for each $p \in S_J$,

$$\lambda_j(p,p;T) = \lambda_k(p,p;T), \text{ for all } j,k = 1,2,\cdots,J. \tag{14}$$

The condition in (14) has an important merit of not depending on any aspects of the data generating process other than what is already fully known, i.e., R and J. It remains to search for T such that (14) is satisfied.

We first write out

$$\Lambda_R(p,p_0;T) = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,J}} T_j(m) p_R(m;p), \tag{15}$$

where $p_R(\cdot;p)$ is the probability mass function of $F_{R,p}$ which is given by

$$p_R(m;p) = \binom{R}{m_1,\cdots,m_J} p_1^{m_1}\cdots p_J^{m_J}.$$

From (15), $\Lambda_R(p,p_0;T)$ is a weighted sum of $p_R(m;p)$ over $m \in \mathbb{N}_{R,J}$. Hence for each k, we can write $\lambda_k(p_0,p_0;T)$ also as a weighted sum of $p_R(m,p_0)$ by rearranging the terms. More specifically, suppose that $J=2$, and take the following form of transform, $T_j(m) = \tilde{T}(m_j)$ for a certain map $\tilde{T}$ on $\{0,1,\cdots,R\}$. Then we can write from (12):

$$\begin{aligned}
\lambda_1(p_0,p_0;T) &= \sum_{j=1}^{2} p_{j0} \sum_{m \in \mathbb{N}_{R,2}} \tilde{T}(m_j) \begin{pmatrix} R \\ m_1, m_2 \end{pmatrix} m_1 p_{10}^{m_1-1} p_{20}^{m_2} \\
&= \sum_{m \in \mathbb{N}_{R,2}} c_1(m_1,m_2) p_{10}^{m_1} p_{20}^{m_2} \\
\lambda_2(p_0,p_0;T) &= \sum_{j=1}^{2} p_{j0} \sum_{m \in \mathbb{N}_{R,2}} \tilde{T}(m_j) \begin{pmatrix} R \\ m_1, m_2 \end{pmatrix} m_2 p_{10}^{m_1} p_{20}^{m_2-1} \\
&= \sum_{m \in \mathbb{N}_{R,2}} c_2(m_1,m_2) p_{10}^{m_1} p_{20}^{m_2},
\end{aligned}$$

where, by setting $\tilde{T}(-1) = 0$,

$$\begin{aligned} c_1(m_1, m_2) &= \tilde{T}(m_1) \binom{R}{m_1, m_2} m_1 + \tilde{T}(m_2 - 1) \binom{R}{m_1 + 1, m_2 - 1} (m_1 + 1) \\ &= \binom{R}{m_1, m_2} \left[\tilde{T}(m_1) m_1 + \tilde{T}(m_2 - 1) m_2 \right] \text{ and} \\ c_2(m_1, m_2) &= \tilde{T}(m_2) \binom{R}{m_1, m_2} m_2 + \tilde{T}(m_1 - 1) \binom{R}{m_1 - 1, m_2 + 1} (m_2 + 1) \\ &= \binom{R}{m_1, m_2} \left[\tilde{T}(m_2) m_2 + \tilde{T}(m_1 - 1) m_1 \right]. \end{aligned}$$

From (14), we impose on $\tilde{T}$ the condition that for all m_1, m_2 such that $m_1 + m_2 = R$,

$$c_1(m_1, m_2) = c_2(m_1, m_2)$$

or

$$m_1 \{T(m_1) - T(m_1 - 1)\} = m_2 \{T(m_2) - T(m_2 - 1)\}. \tag{16}$$

Now suppose that T satisfies that $T(R) = T(R - 1)$, and

$$T(m_1) - T(m_1 - 1) = \frac{a}{m_1}, m_1 = 1, 2, \cdots, R - 1, \tag{17}$$

for some $a > 0$. Then this transform T should satisfy the equations in (16). Starting from an arbitrary initial value $T(R)$, we can recursively recover the values of $T(m_1)$ for each $m_1 \in \{0, 1, 2, \cdots, R\}$ from (17). The resulting transform T takes the following form: for each $m_1 = 0, 1, 2, \cdots, R - 1$,

$$\begin{aligned} T(m_1) &= -\sum_{s=1}^{R-m_1-1} \frac{a}{R - s} + T(R) \\ &= -\sum_{s=0}^{R-m_1-1} \frac{a}{R - s} + \frac{a \cdot 1\{m_1 = 0\}}{R} + T(R), \end{aligned}$$

where we take the summation $\sum_{s=0}^{-1}$ to be zero. By setting $T(R) = 0$ and $a = 1$, we obtain the formula in (8) for the case of $J = 2$. For $J \geq 3$, the derivation of T from (16) follows similar arguments but is more involved. The solution for the case of a general J is given in the following algebraic result.

LEMMA 1: *Transform vector $T = (T_1, \cdots, T_J)$ satisfies (14), if for each $j = 1, \cdots, J$ and*

$R = 2, 3, \cdots,\ T_j(m) = T_{R,j}(m)$, *where*

$$T_{R,j}(m) = -\sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\sum_{k=1,k\neq j}^{J} 1\{m_k > 0\}}{R}. \tag{18}$$

Lemma 1 is a pure algebraic result that does not involve any unknown specifics of the data generating process in the model. While the existence of such an explicit transform is novel in our view, the transform is based on the condition (14) that is only a necessary condition for (10). Now the sufficient second order condition for the optimization problem in (10) stems from the result of Lemma 2 below. Given $p \in S_J$, we write $\tilde{p} = (p_1, \cdots, p_{J-1})^\top$. We define

$$\tilde{\Lambda}_R(\tilde{p}, p_0; T_R) = \Lambda_R((p_1, \cdots, p_{J-1}, 1 - \Sigma_{j=1}^{J-1} p_j), p_0; T_R). \tag{19}$$

Hence $\tilde{\Lambda}_R(\tilde{p}, p_0; T_R)$ is $\Lambda_R(p, p_0; T_R)$ with the imposition of the constraint $\Sigma_{j=1}^{J} p_j = 1$.

Lemma 2: *Let* $T_R = (T_{R,1}, \cdots, T_{R,J})$ *with* $T_{R,j}$ *given by* (18). *Then for any* $a = (a_1, \cdots, a_{J-1})^\top \in \mathbf{R}^{J-1}$,

$$a^\top \left(\frac{\partial^2 \tilde{\Lambda}_R(\tilde{p}, p_0; T_R)}{\partial \tilde{p} \partial \tilde{p}^\top} \right) a \leq -2\varepsilon \left(\sum_{j=1}^{J-1} a_j^2 \right)$$

where $\varepsilon > 0$ *is such that* $p_{j0} \in [\varepsilon, 1-\varepsilon]$, *for all* $j = 1, \cdots, J$.

Lemma 2 says that the function $\tilde{\Lambda}_R(\tilde{p}, p_0; T_R)$ for all $\tilde{p} = (p_1, \cdots, p_{J-1})^\top$ with $p \in S_J$ is globally strictly concave in $\tilde{p}$ if p_0 is such that for some $\varepsilon > 0$, $p_{j0} \in [\varepsilon, 1-\varepsilon]$, for all $j = 1, \cdots, J$. Combined with Lemma 1, the result of Lemma 2 shows that $\Lambda_R(p, p_0; T_R)$ (under constraint $\Sigma_{j=1}^{J} p_j = 1$) is uniquely maximized at $p = p_0$. Formal proofs of Lemmas 1 and 2 are algebraically involved and provided in the appendix.

It may not be immediately clear how the choice of (18) is related to MLE with sufficiently large R. To see this connection, note first that the simulated frequencies $m_{ij}(\theta)/R \to_P p_{ij}(\theta) \in (0,1)$ with $R \to \infty$, by the law of large numbers, where $p_{ij}(\theta) = p_j(X_i; \theta)$. Also, note that

$$0 \leq \frac{1}{R} \sum_{k=1,k\neq j}^{J} 1\{m_k > 0\} \leq \frac{J-1}{R} \to 0$$

as $R \to \infty$. Finally, observe that for large R,

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

This latter convergence is immediate as the sum on the left-hand side is a Riemann sum of

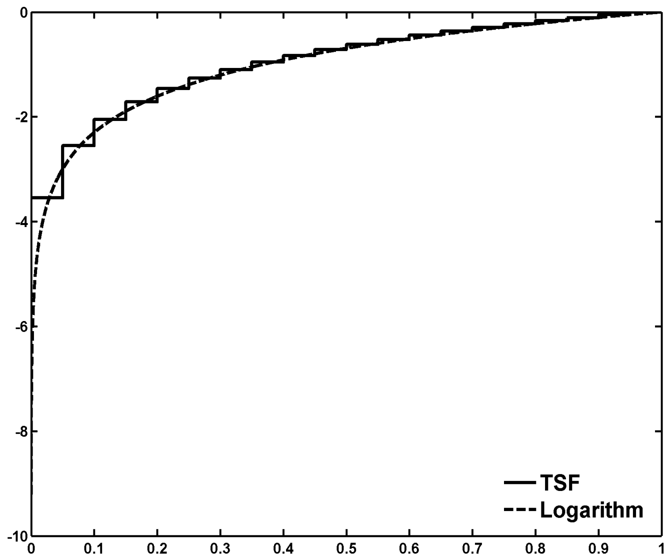


Figure 3: The Comparison of TSF and Logarithmic Function. The solid line depicts $T_{R,j}(\lfloor Rp \rfloor)$ against $p \in (0,1)$ in the case of $J = 2$ and $R = 20$ and the dashed line depicts $\log(p)$, where $\lfloor Rp \rfloor$ denotes the greatest integer that is not larger than Rp. As R increases, $T_{R,j}(\lfloor Rp \rfloor)$ becomes closer to $\log(p)$.

$-\int_0^{1-m_{ij}(\theta)/R-1/R}(1/(1-u))du$ and $m_{ij}(\theta)/R \approx p_{ij}(\theta)$ for large R. Therefore,

$$\mathbf{E}[D_{ij}T_{R,j}(m_{ij}(\theta))] \approx \mathbf{E}[D_{ij}\log p_{ij}(\theta)],$$

for large R. Hence the TSF population objective function is close to that of MLE for large R. That the TSF approximates the logarithmic function is clearly seen in Figure 3.

3.3 Asymptotic Properties

In this section we investigate the asymptotic properties of the estimator $\hat{\theta}$ defined in (7). The asymptotic properties of $\hat{\theta}$ are developed for two separate cases: the case with R fixed and the case with R tending to infinity jointly with the sample size n. 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 and $\{\eta_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ are independent.
(ii) There exists $C > 0$ such that $\sup_{x\in\mathcal{X}}\mathbf{E}[\sup_{\tilde{\theta}\in\Theta}\sup_{\theta\in B(\tilde{\theta};\varepsilon)}|\delta_j(x,\eta_i;\tilde{\theta})-\delta_j(x,\eta_i;\theta)|^2] \leq C\varepsilon$, for any $\varepsilon > 0$.
(iii) (a) For each $j \in \{1,\cdots,J\}$, $p_j(x;\theta)$ is bounded away from zero uniformly over $x \in \mathcal{X}$

and $\theta \in \Theta$, and (b) there exists $C > 0$ such that

$$\sum_{j=1}^{J} \mathbf{E}\left[(p_j(X_i, \theta) - p_j(X_i, \theta_0))^2\right] \geq C||\theta - \theta_0||^2,$$

for all $\theta \in \Theta$.
(iv) There exists $\delta > 0$ such that 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 in $\theta \in B(\theta_0; \delta)$ with derivatives bounded uniformly over $\theta \in B(\theta_0; \delta)$.

Condition (ii) controls the manner the random decision rule $\delta_j(X_i, \eta_i; \theta)$ depends on θ and (X_i, η_i). The condition requires that the decision rule δ is locally uniformly L_2-continuous in θ (e.g. Chen, Linton, and 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. See Example 1 below for lower level conditions in the case of a random utility framework. Condition (iii)(a) ensures that $\log p_j(x; \theta)$ is well defined for all the values of $\theta \in \Theta$ and $x \in \mathcal{X}$. For example, see the fourth condition in Assumption 1 of Lee (1995) for a similar condition. One may weaken this condition by introducing a trimming sequence in the estimator as in Klein and Spady (1993). Condition (b) is a regularity condition that is used to ensure $\sqrt{n}$-consistency of MLE. Condition (iv) requires the smoothness of the conditional choice probabilities in the parameter θ local around θ_0.

Example 1: 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 $u_j(X_i, \eta_{ij}; \theta)$. Then she makes the j-th choice when

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

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

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

Suppose that for each $\theta \in \Theta$, and for each x in the support of X, and for $j = 1, 2, \cdots, J$,

$$\left|u_j(X_i, \eta_{ij}; \theta) - u_j(X_i, \eta_{ij}; \tilde{\theta})\right| \leq C||\theta - \tilde{\theta}||.$$

Then the condition of Assumption 2(ii) holds. ■

The following theorem establishes the rate of convergence for the TSF-MLE estimator $\hat{\theta}$ for finite R.

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

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

Theorem 2 tells us that the TSF-based estimator becomes more accurate as the sample size becomes large, even when the simulation number is finite. In fact, the estimator $\hat{\theta}$ follows the cube-root asymptotics of Kim and Pollard (1990) with fixed R and $n \to \infty$. The cube-root asymptotics occurs precisely in the same way it occurs in Manski's maximum score estimation. 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) : *Suppose that* Assumptions 1-2 *hold. As $n, R \to \infty$ jointly, with $\sqrt{n}/R \to 0$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, V),$$

where $V = \Omega^{-1}$ and

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

The rate condition $\sqrt{n}/R \to 0$ is typically used in the simulated likelihood literature. (See Gouriéroux and Monfort (1991) and Stern (1997).)

When R is small, the asymptotic distribution of $\hat{\theta}$ is not normal, and hence one cannot use the usual standard error formula of MLE. This fact applies commonly to all the simulated likelihood based estimators in the literature to our best knowledge. It appears to us that the computation of a confidence set with a small simulation number is still an open problem both in theory and in practice. One may suggest using subsampling that is known to work for an estimator that follows cube-root asymptotics (Delgado, Rodríguez-Poo, and Wolf (2001)). However, in order for the computation of a confidence set to be effective, it should not take long; otherwise, it would be better to use the time to simply increase the simulation number and resort to asymptotic normal approximation. This is why we believe the subsampling approach in our situation is not a desirable option.

Needless to say, the finite R asymptotics does not suggest choosing a small R over a large R. It is always better to use a large simulation number as far as the computation cost is not too large. Therefore, the choice of R is by no means an issue of delicacy like bandwidth choice in nonparametric estimation. In the latter case, bandwidth should be chosen to be not too small and not too large relative to the sample size in order to ensure consistency of the estimator. On the other hand, choice of the simulation number R does not affect the consistency of the TSF-based estimator, and it is always better to use larger R for efficiency.

Example 2: Cohort-Level Aggregate Data: 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 $D(k) = (D_1(k), \cdots, D_J(k))^\top$. 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

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

When the conditional distribution of the stochastic error $\eta_{ij}(k)$ given $X(k)$ is identical for each individual i, the conditional probability $P\left\{D_{ij}(k) = 1 | X(k), \theta\right\}$ 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

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

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

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

Note that

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

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.

4 Monte Carlo Study

We performed a Monte Carlo simulation study based on three different models. The first model is a simple multinomial logit model where we can evaluate the choice probability explicitly and compare simulation-based estimators with MLE. The second model is a dynamic schooling choice model where the unobserved heterogeneity in the payoff functions is time invariant. In this case, the model can be written as a static discrete choice model. The third model is a dynamic schooling model where unobserved heterogeneity in the payoff functions is time-varying. In this case, one cannot reduce the dynamic model to a static discrete choice model.

4.1 Multinomial Logit Models

We consider the following standard logit models, where we can explicitly compare the true MLE, the Lerman and Manski's simulated frequency method, and our transformed simulated frequency method. The choice probability is specified as follows:

$$\begin{aligned}
P\{D_{i1} &= 1|X_i\} = \frac{1}{1+\exp(X_{1i}\beta_1)+\exp(X_{2i}\beta_2)}, \\
P\{D_{i2} &= 1|X_i\} = \frac{\exp(X_{1i}\beta_1)}{1+\exp(X_{1i}\beta_1)+\exp(X_{2i}\beta_2)}, \text{ and} \\
P\{D_{i3} &= 1|X_i\} = \frac{\exp(X_{2i}\beta_2)}{1+\exp(X_{1i}\beta_1)+\exp(X_{2i}\beta_2)}.
\end{aligned}$$

As for the distribution of $(X_{1i}, X_{2i})^{\top}$, the study considered the following three different specifications. Let $V_i \sim \text{Uniform}[0,1]$, $Z_i \sim N(0,1)$, $B_{1i} \sim$Binomial$(2, 0.3)$, $B_{2i} \sim$Binomial$(2, 0.5)$, and $W_i = Z_{1i} + (B_{3i} - 1/2)/4$, where $B_{3i} \sim$Bernoulli(0.3) and $Z_{1i} \sim N(0,1)$. The random variables $V_i, Z_i, Z_{1i}, B_{1i}, B_{2i}$, and B_{3i} were drawn independently. Using these random variables, we specified X_{1i} and X_{2i} as follows:

$$
\begin{aligned}
\text{Specification A:} \quad & X_{1i} = Z_{1i} + V_i \\
& X_{2i} = Z_{2i} + V_i \\
\text{Specification B:} \quad & X_{1i} = \Phi\left(Z_{1i} + V_i\right) - 1 + V_i \\
& X_{2i} = 2U_{1i} - 4U_{2i}^2 + V_i \\
\text{Specification C:} \quad & X_{1i} = \Phi\left(Z_{1i}/2 + (B_{1i}/2 - 1)/4\right) - 1/2 + W_i \\
& X_{2i} = Z_i/2 + \Phi(B_{2i}/2 - 1)/4 + W_i) - 1/2.
\end{aligned}
$$

The Monte Carlo simulation number was taken to be 5000.

Figure 4 reports the average of the mean absolute deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$. Both in the cases of the sample size equal to 300 and 1000, the TSF dominates the Lerman and Manski's simulated frequency method when the simulation number is small ranging from 10 to 100. When the sample size is 300, this order of dominance is slightly reversed when the simulation number is larger, although both converging to the mean absolute deviation of the MLE. However, when the sample size is 1000, the estimator from the TSF method dominates that of the Lerman and Manski's method uniformly over all the simulation numbers considered. As expected, the dominance is prominent when the simulation number is small. This is because the TSF method delivers a consistent estimator even for a small simulation number while the Lerman and Manski's simulated frequency method does not.

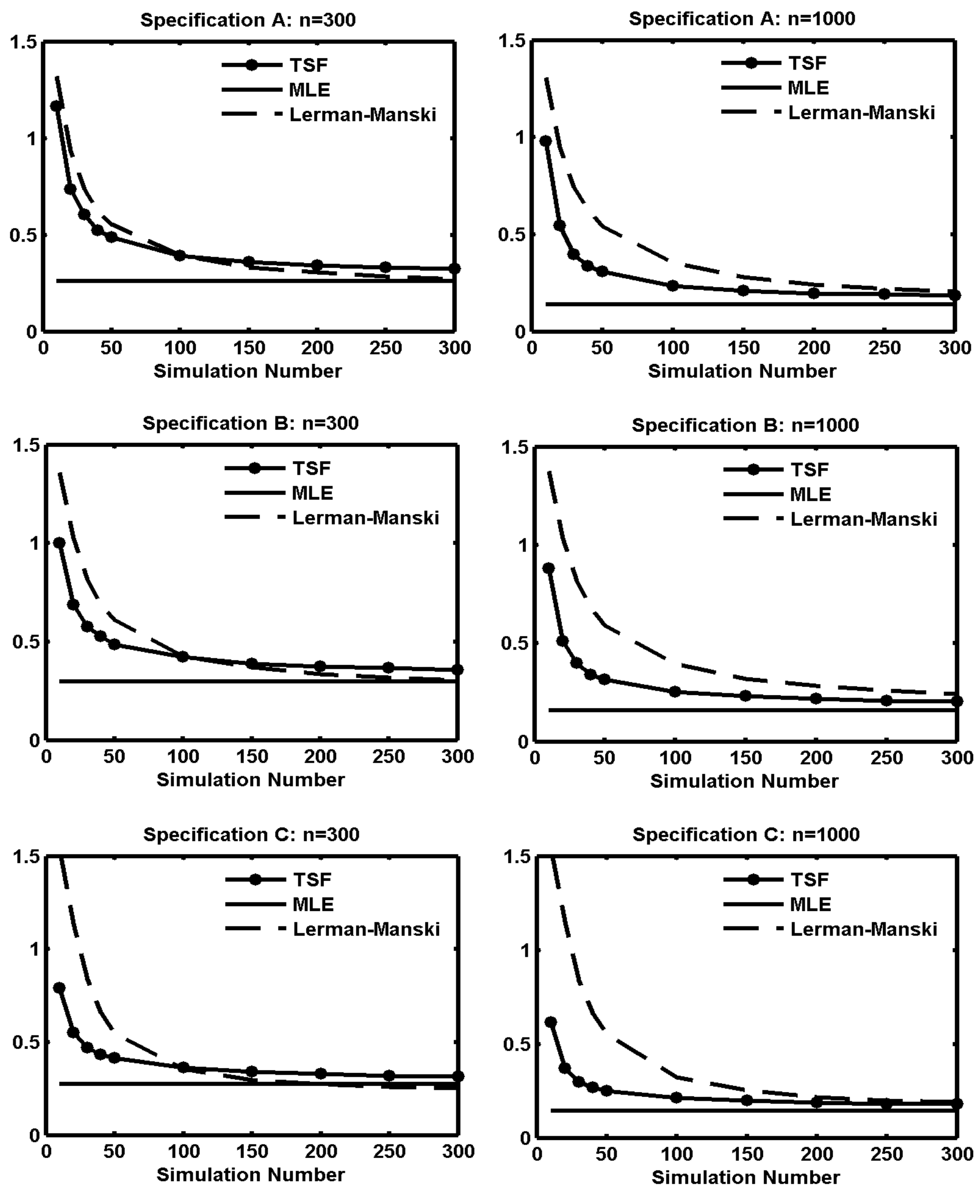
4.2 Schooling Choice with Time Invariant Unobserved Heterogeneity

4.2.1 The Data Generating Process

In this section, we present and discuss results from a Monte Carlo simulation study based on a model of schooling choice. The model involves observed ability affecting each agent's labor market outcome, and unobserved heterogeneity in discount factor and preference. See Willis and Rosen (1979) and Keane and Wolpin (1997) for structural models of education with 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

Figure 4: Average Mean Absolute Deviation of Estimators of β_1 and β_2



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 specification:

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

where E is the years of education taking values of 10, 12, 14, and 16 and ε_w is normal, i.i.d., across individuals and periods with standard deviation of σ_w. Once an individual enters the labor market and starts working, going back to school is not permitted in the model.

In each period t, the utility is given by $U_{w,t}$ if the individual works, and $U_{s,t}$ 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 from entering the labor market and from school:

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

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 is 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_s 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_s is denoted by σ_s.

We assume that the discount factor δ is heterogeneous across people and is correlated with an observed covariate X_3. It is specified as

$$\delta = \frac{1}{1+\exp(\varepsilon_\delta + \rho_0 + \rho_1 X_3)}, \tag{20}$$

where ε_δ is normally distributed with mean 0 and standard deviation σ_δ and does not change over time for each individual. The errors $\varepsilon_w, \varepsilon_s,$ and ε_δ are independent.

Let E be the total amount of schooling and $U(E=a)$ be the discounted utility from schooling choice $E=a$ at the beginning of one's life cycle. Given that working is an absorbing state, we can represent this multi period dynamic programming model as a 4-choice static model with the following random utilities:

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

Given the model structure, we expect people with higher ability X_1 and X_2, higher discount factor δ and higher utility value of schooling ε_s 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 δ and the utility value of schooling ε_s are not observed. For simplicity, 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 are as follows:

$$\begin{aligned}
\text{schooling utility parameters}: &\quad \gamma_1, \gamma_2, \gamma_3, \sigma_s \text{ and} \\
\text{discount factor parameters}: &\quad \rho_0, \rho_1, \sigma_\delta.
\end{aligned}$$

For comparison with our TSF-MLE, we considered MSL estimator following Lerman and Manski (1981)'s simulated frequency method, and its smoothed version (McFadden (1989)). The Lerman and Manski 's simulated frequency method 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 MSL estimator which is computed by using the following smoothed simulated choice probability:

$$\frac{1}{R}\sum_{r=1}^{R}\frac{\exp(U_{j,r,t}/\lambda)}{\Sigma_{j=1}^{J}\exp(U_{j,r,t}/\lambda)}. \tag{21}$$

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

Note that except for the simulated frequency method of Lerman and Manski (1981) or its smoothed version, we cannot apply the existing simulation methods that require the presence of additive normal or logit errors in the random utility due to nonlinear unobserved heterogeneity in discount factor.

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. The parameter values are chosen as follows. As for wage parameters, $\alpha_0 = 8$, $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 0.07$, and $\sigma_w = 0.3$. As for the schooling utilities, $\gamma_1 = 0$, $\gamma_2 = -5000$, $\gamma_3 = -20,000$, and $\sigma_s = 5000$. And finally, as for discount factors, $\rho_0 = -0.25$ and $\rho_1 = 0.2$.

4.2.2 The Results

The results are reported in Tables 5-8. In Table 5 we compare 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 and Manski's simulated frequency method is different from its smoothed version. The simulation results show that the use of smoothing does not necessarily improve the performance, and sometimes, even worsen the quality of the estimator.

Table 5: True Log Likelihood Evaluated at the Estimators

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	-857.99	-853.07	-848.21	-845.89
$n = 300$	Lerman-Manski	-912.13	-859.81	-848.93	-849.82
	Smoothed Lerman-Manski ($\lambda = 0.1$)	-905.51	-871.45	-849.19	-845.93
	Smoothed Lerman-Manski ($\lambda = 0.01$)	-903.98	-871.36	-849.19	-846.00
	TSF-MLE	-855.51	-852.52	-850.57	-849.78
$n = 1000$	Lerman-Manski	-907.59	-858.35	-851.15	-849.82
	Smoothed Lerman-Manski ($\lambda = 0.1$)	-937.18	-906.90	-856.37	-850.81
	Smoothed Lerman-Manski ($\lambda = 0.01$)	-935.56	-906.91	-856.39	-850.85
	TSF-MLE	-853.35	-852.44	-851.85	-851.57
$n = 5000$	Lerman-Manski	-905.72	-856.25	-852.12	-851.58
	Smoothed Lerman-Manski ($\lambda = 0.1$)	-962.27	-949.84	-872.51	-855.93
	Smoothed Lerman-Manski ($\lambda = 0.01$)	-960.47	-948.46	-872.59	-855.99

Table 6: MAE of Estimated γ_1

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	0.038	0.036	0.031	0.031
$n = 300$	Lerman-Manski	0.086	0.041	0.030	0.030
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.068	0.043	0.027	0.029
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.067	0.042	0.027	0.029
	TSF-MLE	0.026	0.023	0.020	0.019
$n = 1000$	Lerman-Manski	0.080	0.029	0.019	0.018
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.098	0.083	0.023	0.018
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.097	0.082	0.024	0.017
	TSF-MLE	0.016	0.015	0.012	0.012
$n = 5000$	Lerman-Manski	0.086	0.021	0.011	0.011
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.121	0.123	0.046	0.018
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.121	0.121	0.047	0.017

Table 7: MAE of Estimated γ_2.

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	0.059	0.052	0.042	0.038
	Lerman-Manski	0.080	0.058	0.044	0.038
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.110	0.082	0.045	0.036
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.110	0.083	0.045	0.037
$n = 1000$	TSF-MLE	0.039	0.032	0.025	0.022
	Lerman-Manski	0.067	0.043	0.026	0.023
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.084	0.084	0.051	0.028
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.084	0.085	0.051	0.027
$n = 5000$	TSF-MLE	0.022	0.018	0.014	0.012
	Lerman-Manski	0.049	0.030	0.015	0.012
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.042	0.042	0.086	0.046
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.045	0.041	0.085	0.046

Table 8: MAE of Estimated γ_3.

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	0.040	0.040	0.037	0.036
	Lerman-Manski	0.081	0.046	0.037	0.036
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.053	0.042	0.032	0.034
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.051	0.042	0.033	0.033
$n = 1000$	TSF-MLE	0.029	0.026	0.023	0.022
	Lerman-Manski	0.078	0.034	0.024	0.022
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.080	0.055	0.027	0.020
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.080	0.056	0.026	0.020
$n = 5000$	TSF-MLE	0.019	0.015	0.012	0.012
	Lerman-Manski	0.088	0.028	0.013	0.011
	Smoothed Lerman-Manski ($\lambda = 0.1$)	0.129	0.104	0.036	0.022
	Smoothed Lerman-Manski ($\lambda = 0.01$)	0.128	0.105	0.036	0.022

When the sample size is small, the performance of Lerman and Manski's simulated frequency method and its smoothed version becomes comparable to our methods. However, when the sample size is large, the improved performance of our estimator becomes prominent over that of the competing procedures. This confirms our theoretical result that our estimator is consistent even when the simulation number is small, but the Lerman and Manski's procedures do not possess this property.

A similar pattern of performance comparison is obtained in terms of mean absolute errors (MAE) of individual estimators, as shown in Tables 6-8. While not reported here, we observed a similar pattern of performance for other parameters.

Lastly, in Table 9, we report a sample of computing time for each method. Overall, it is easily seen that as one increases n and R, the computation time increases. This suggests that an estimator that maintains good quality for a smaller simulation number R is also computationally more convenient. The computing times for the TSF method and Lerman and Manski's simulated frequency method turned out to be similar, except when $R = 100$ and $n = 100$ or 200. In our simulation study, the smoothed version of Lerman and Manski's method does not show computational efficiency.

Table 9: Computing Time for Obtaining a Point Estimate (in Median Seconds from 1000 Simulations)

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	2	4	8	16
$n = 100$	Lerman-Manski	2	6	7	9
	Smoothed Lerman-Manski ($\lambda = 0.1$)	4	3	6	14
	Smoothed Lerman-Manski ($\lambda = 0.01$)	3	7	7	27
	TSF-MLE	4	11	11	18
$n = 200$	Lerman-Manski	5	8	10	13
	Smoothed Lerman-Manski ($\lambda = 0.1$)	7	12	11	21
	Smoothed Lerman-Manski ($\lambda = 0.01$)	8	15	11	23
	TSF-MLE	9	9	17	32
$n = 500$	Lerman-Manski	10	14	16	31
	Smoothed Lerman-Manski ($\lambda = 0.1$)	18	16	26	51
	Smoothed Lerman-Manski ($\lambda = 0.01$)	20	24	25	54
	TSF-MLE	12	15	32	65
$n = 1000$	Lerman-Manski	13	15	30	62
	Smoothed Lerman-Manski ($\lambda = 0.1$)	20	28	57	105
	Smoothed Lerman-Manski ($\lambda = 0.01$)	32	43	56	109

4.3 Schooling Choice with Time-Varying Unobserved Heterogeneity

4.3.1 The Data Generating Process

The schooling choice model considered here is the same as the previous model except that the utility of schooling involves an idiosyncratic shock each period which is observed by the individual before he makes a choice but is not observed by the econometrician. More specifically, the utility from schooling at period t is given by

$$U_{st}(E_t) = \gamma_0 + \gamma_1 1\{E_t \geq 4\} + \varepsilon_{s,t},$$

where $\varepsilon_{s,t}$ is the idiosyncratic shock. Here γ_0 is utility of schooling in the first 4 years of schooling and γ_1 is the additional utility of schooling, potentially associated with the tuition cost of college in which case $\gamma_1 < 0$.

The decision rule for an individual is written in a standard dynamic programming framework. Given that leaving schooling is an absorbing state, we look at the decisions at t who have continued to school up to $t-1$ with $E = t-1$. The value of attending school and the value of working at education level E and at period t are given by

$$\begin{aligned} V_s(t, E) &= U_{st}(E) + \delta \mathbf{E}\left[\max\left\{V_s(t+1, E+1), V_w(E+1)\right\} | X, E, \varepsilon_\delta\right] \\ V_w(t, E) &= U_{w,t}(E) + \sum_{\tau=1}^{T-t} \delta^\tau U_{w,t}(E), \end{aligned}$$

where ε_δ is the error term in (20) and $X = (X_1, X_2, X_3)^\top$. Here

$$U_{w,t}(E) = \mathbf{E}\left(w_t | X_1, X_2, E\right) = \exp\left(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \frac{1}{2}\sigma_w^2\right).$$

Individuals will attend school if $V_s(t, E) > V_w(t, E)$.

While the expected value function $V_s(t, E)$ depends not only on the observable variables of X's, but also on the unobservable component of ε_δ, the choice probability of education requires integration over ε_δ. In this situation, the method of smoothed simulated frequencies such as (21) is quite cumbersome, because one then needs to smooth choice probabilities of the binary decision on schooling or work from $t = 0$ to $t = E$, as well as the probability of leaving school at $t = E + 1$. On the other hand, the method of TSF and Lerman and Manski's simulated frequencies is computationally efficient, because one can directly count the number of simulated outcomes that match the given amount of schooling observed for each individual.

The simulation was performed as follows. First, we drew realizations from the distributions of ε_δ and X that were set to be time-invariant. Then, for each simulated period, we drew realizations from the distribution of $\varepsilon_{s,t}$, and calculated simulated versions of $V_s(t, E)$ and $V_w(t, E)$. These simulated versions of value functions as well as the given levels of X's and ε_δ constitute individual decision rules. Following these decision rules, we simulated individual work/schooling choices and saved the results. We performed the same steps now beginning with another set of realizations from the distributions of ε_δ and X. By repeating this process, we obtained simulated frequencies of each individual's choice corresponding to different values of X's and ε_δ. Using these simulated frequencies, we performed simulated maximum likelihood estimation in two different ways: one using our TSF-based method and the other following Lerman and Manski's simulated frequency method. Note that there are 11 discrete choices ($E = 0$ to $E = 10$) which come from the binary choices between work and school at each point of time over the life cycle. (Recall that in this model, once an individual decides to work, she cannot come back to school for the rest of her life.)

The parameters used in the simulation study are as follows. As for schooling parameters, $\gamma_0 = 10,000$, $\gamma_1 = -10,000$, and $\sigma_s = 5,000$. Regarding observable characteristics vector X, it is specified as follows.

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix},$$

where ν's are i.i.d. standard normal.

In the actual simulation, we fix all the wage parameters, and estimate the six of the schooling parameters and discount factor parameters. The likelihood values based on the estimators and the mean absolute deviations from the true parameters are reported.

4.3.2 The Results

The results are reported in Tables 10-16. The performance of TSF-MLE overall performs better than that of Lerman and Manski's simulated frequency method. In terms of true log-likelihood evaluated at the estimators, the estimator based on TSF-MLE is closer to the true MLE than the estimator of Lerman and Manski's simulated frequency method (Table 10). Also the MAEs of parameter estimates from TSF-MLE are mostly smaller than those from Lerman and Manski's simulated frequency method.

Table 10: True Log Likelihood Evaluated at the Estimators

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	-1695.69	-1690.51	-1685.33	-1682.72
	Lerman-Manski	-1829.93	-1711.97	-1686.86	-1686.24
$n = 1000$	TSF-MLE	-1692.47	-1690.45	-1687.17	-1685.97
	Lerman-Manski	-1837.39	-1713.67	-1688.82	-1686.24
$n = 5000$	TSF-MLE	-1689.55	-1688.54	-1687.55	-1687.12
	Lerman-Manski	-1840.09	-1713.68	-1689.38	-1687.36

Table 11: MAE of Estimated ρ_0

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	0.0307	0.0283	0.0249	0.0234
	Lerman-Manski	0.0580	0.0396	0.0274	0.0235
$n = 1000$	TSF-MLE	0.0198	0.0184	0.0159	0.0156
	Lerman-Manski	0.0562	0.0344	0.0208	0.0159
$n = 5000$	TSF-MLE	0.0124	0.0111	0.0099	0.0088
	Lerman-Manski	0.0600	0.0341	0.0169	0.0107

Table 12: MAE of Estimated ρ_1

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	0.0331	0.0304	0.0278	0.0243
	Lerman-Manski	0.0959	0.0548	0.0326	0.0249
$n = 1000$	TSF-MLE	0.0207	0.0197	0.0167	0.0157
	Lerman-Manski	0.1042	0.0513	0.0231	0.0167
$n = 5000$	TSF-MLE	0.0132	0.0114	0.0099	0.0088
	Lerman-Manski	0.1087	0.0482	0.0170	0.0104

Table 13: MAE of Estimated σ_δ

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	0.0107	0.0107	0.0105	0.0102
	Lerman-Manski	0.0110	0.0097	0.0095	0.0098
$n = 1000$	TSF-MLE	0.0097	0.0099	0.0091	0.0095
	Lerman-Manski	0.0093	0.0079	0.0078	0.0078
$n = 5000$	TSF-MLE	0.0090	0.0085	0.0080	0.0075
	Lerman-Manski	0.0083	0.0068	0.0057	0.0058

Table 14: MAE of Estimated γ_0

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	286.15	269.50	251.72	241.27
	Lerman-Manski	908.56	378.80	298.93	268.83
$n = 1000$	TSF-MLE	228.25	220.92	201.30	200.21
	Lerman-Manski	1159.10	347.30	287.80	239.13
$n = 5000$	TSF-MLE	161.32	156.11	151.45	144.59
	Lerman-Manski	1652.87	406.66	305.05	219.30

Table 15: MAE of Estimated γ_1

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	279.91	242.48	221.05	222.13
	Lerman-Manski	1427.69	471.45	251.70	232.17
$n = 1000$	TSF-MLE	210.76	194.12	179.55	179.12
	Lerman-Manski	1632.24	450.25	229.41	192.51
$n = 5000$	TSF-MLE	145.05	135.32	139.53	131.75
	Lerman-Manski	2094.36	497.22	237.65	168.65

Table 16: MAE of Estimated σ_s

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
$n = 300$	TSF-MLE	823.64	764.97	645.38	592.98
	Lerman-Manski	1493.10	839.97	569.71	560.06
$n = 1000$	TSF-MLE	633.43	579.57	448.05	390.52
	Lerman-Manski	1709.34	863.23	382.59	371.73
$n = 5000$	TSF-MLE	398.67	337.53	262.41	211.04
	Lerman-Manski	1797.97	961.55	256.57	192.67

In Table 17, we report the computing time taken for this simulation study. The computing time is substantially longer than that from the previous dynamic schooling model. This again emphasizes the fact that using a smaller simulation number in the case of a large sample size is computationally more convenient. The computing times for the TSF method and Lerman and Manski's method are overall similar. This affirms our claim that using TSF does not cause much additional computational cost beyond that of Lerman and Manski.

Table 17: Computing Time for Obtaining a Point Estimate (in Seconds on Average from 1000 Simulations)

Sample Size	Simulation Methods	$R = 10$	$R = 20$	$R = 50$	$R = 100$
	TSF-MLE	85.6	107.9	193.6	306.3
$n = 300$	Lerman-Manski	102.7	134.7	242.4	302.2
	TSF-MLE	199.4	252.9	438.0	759.1
$n = 1000$	Lerman-Manski	268.8	331.6	465.7	762.4
	TSF-MLE	839.6	1066.8	1902.1	3296.2
$n = 5000$	Lerman-Manski	969.1	1187.4	1959.3	3387.8

5 Conclusion

In this paper we propose an alternative method of MSL for discrete choice models that is applicable in various specifications of random utilities. While the method is as easy to apply as Lerman and Manski's simulated frequency method, it is also free from the issue of zero simulated choice probabilities and the issue of simulation bias. Furthermore, when the simulation number is large, the estimator is bound to achieve the efficiency of the infeasible MLE. This advantage is demonstrated both through the asymptotic result that shows consistency of the estimator with a finite simulation number and through various Monte Carlo simulation results.

6 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 : Let $T = (T_1, \cdots, T_J)$ be given collection of maps $T_j : \mathbb{N}_{R,J} \to \mathbf{R}$, $j = 1, \cdots, J$, such that for each $j = 1, 2, \cdots, J$,

$$T_j(m) = \tilde{T}(m_j, m_{-j}), \tag{22}$$

for a fixed map $\tilde{T} : \mathbb{N}_{R,J} \to \mathbf{R}$, where $m_{-j} = (m_1, \cdots, m_{-j-1}, m_{-j+1}, \cdots, m_J)$. Using this map, we write

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

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

$$\begin{aligned}
\lambda_k(p_0, p_0; T) &= \frac{\partial}{\partial p_k}\Lambda_R(p, p_0; T)|_{p=p_0} \\
&= \sum_{j=1}^{J} p_{j0} \sum_{m\in\mathbb{N}_{R,J}} \tilde{T}(m_j, m_{-j}) \left(\begin{array}{c} R \\ m_1, \cdots, m_J \end{array}\right) m_k p_{10}^{m_1} p_{20}^{m_2} \cdots p_{k0}^{m_k-1} \cdots p_{J0}^{m_J} \\
&= \sum_{m\in\mathbb{N}_{R,J}} c_k(m_1, \cdots, m_J) p_{10}^{m_1} \cdots p_{J0}^{m_J},
\end{aligned}$$

where $c_k(m_1, \cdots, m_J)$ are coefficients of $p_{10}^{m_1} \cdots p_{J0}^{m_J}$. For brevity, put $p = p_0$ so that we write $\lambda_1(p) \equiv \lambda_1(p, p; \{T_{R,j}\})$ as

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

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 $\tilde{T}$ in (22) and Λ_R. Hence it suffices to show that

$$c_1(m_1, m_2, \cdots, m_J) = c_1(m_2, m_1, \cdots, m_J). \tag{23}$$

To show this, first note that

$$c_1(m_1, \cdots, m_J) = \tilde{T}(m_1, m_{-1}) \left(\begin{array}{c} R \\ m_1, \cdots, m_J \end{array}\right) m_1 + U_R \tag{24}$$

where

$$\begin{aligned}
U_R &= \sum_{j=2}^{J} \tilde{T}(m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J) \\
&\quad \times \left(\begin{array}{c} R \\ m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J \end{array}\right) (m_1 + 1).
\end{aligned}$$

The relation in (24) holds for any $(m_1, \cdots, m_J) \in \mathbb{N}_{R,J}$ and we can simply extend the domain of $\tilde{T}(m_j, m_{-j})$ to negative numbers by taking $\tilde{T}(m_j, m_{-j}) = 0$ if $m_j < 0$. By noting

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

we write the coefficient $c_1(m_1, \cdots, m_J)$ in (24) as

$$\left(\begin{array}{c} R \\ m_1, \cdots, m_J \end{array}\right) \left[m_1 \tilde{T}(m_1, m_{-1}) + \sum_{j\neq 1} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \right] \tag{25}$$

and similarly, write $c_1(m_2, m_1, m_3, \cdots, m_J)$ as

$$\binom{R}{m_1, \cdots, m_J} \left[m_2 \tilde{T}(m_2, m_{-2}) + \sum_{j \neq 2} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \right] \tag{26}$$

Since the factor in front of the above brackets in (25) and (26) are the same, it suffices for (23) to show that

$$\begin{aligned} & m_1 \tilde{T}(m_1, m_{-1}) + \sum_{j \neq 1} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \\ = \quad & m_2 \tilde{T}(m_2, m_{-2}) + \sum_{j \neq 2} m_j \tilde{T}(m_j - 1, m_2 + 1, m_1, , \cdots, m_{j-1}, m_{j+1}, \cdots, m_J). \end{aligned}$$

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

$$\begin{aligned} & m_1 \left[\tilde{T}(m_1, m_2, \cdots, m_J) - \tilde{T}(m_1 - 1, m_2 + 1, m_3, \cdots, m_J) \right] \\ & + \sum_{j=3}^{J} m_j \left[\tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_J) - \tilde{T}(m_j - 1, m_2 + 1, m_1, \cdots, m_J) \right] \\ = \quad & m_2 \left[\tilde{T}(m_2, m_1, m_3, \cdots, m_J) - \tilde{T}(m_2 - 1, m_1 + 1, m_3, \cdots, m_J) \right]. \end{aligned} \tag{27}$$

Therefore, the proof is complete once we show that the above equality is satisfied by our choice of (36). 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 (36) was discovered. In the proof we generate sufficient conditions for the equality in (27). Then these sufficient conditions lead to the solution of (36).

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 (27) is trivially satisfied.

Case 1) $m_1, m_2 > 0$. Then, the condition (27) is satisfied if

$$m_1 \left[\tilde{T}(m_1, m_2, \cdots, m_J) - \tilde{T}(m_1 - 1, m_2 + 1, m_3, \cdots, m_J) \right] = 1, \tag{28}$$

and

$$\tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_J) - \tilde{T}(m_j - 1, m_2 + 1, m_1, \cdots, m_J) = 0. \tag{29}$$

Restriction (29) implies that $\tilde{T}(m_1, m_2, ..., m_J)$ 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

$$\tilde{T}(m_1, m_2, ..., m_J) = \tilde{T}(m_1, m_2', ..., m_J'),$$

by repeating the process in (29) with adding and subtracting by 1 between two non-zero members from $\{m_2, ..., m_J\}$.

Therefore we write

$$\tilde{T}(m_1, m_2, ..., m_J) = \tilde{T}(m_1, \nu(m_2, ..., m_J)),$$

where ν denotes the number of non-zero elements in the non-choice set. Using the observation in (29), (28) can be re-written as

$$m_1\left[\tilde{T}\left(m_1,\nu\left(m_2,\cdots,m_J\right)\right)-\tilde{T}\left(m_1-1,\nu\left(m_2+1,m_3,\cdots,m_J\right)\right)\right]=1, \tag{30}$$

and note that $\nu\left(m_2,\cdots,m_J\right)=\nu\left(m_2+1,m_3,\cdots,m_J\right)$. Hence we extract one condition for $\tilde{T}$ that leads to (30):

$$\tilde{T}\left(m,\nu\right)-\tilde{T}\left(m-1,\nu\right)=\frac{1}{m}\text{ for all possible }m. \tag{31}$$

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

$$\begin{aligned} & m_1\left[\tilde{T}\left(m_1,m_2,\cdots,m_J\right)-\tilde{T}\left(m_1-1,m_2+1,m_3,\cdots,m_J\right)\right] \\ = \; & R\left[\tilde{T}\left(R,m_2=0,\cdots,m_J=0\right)-\tilde{T}\left(R-1,m_2+1,m_3=0,\cdots,m_J=0\right)\right]=0 \end{aligned} \tag{32}$$

or

$$\tilde{T}\left(R,0\right)=\tilde{T}\left(R-1,1\right). \tag{33}$$

If on the other hand $m_3>0$, we have from (14)

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

By subtracting and adding back $\tilde{T}\left(m_1-1,m_2,m_3+1,\cdots,m_J\right)$, we can write the above equation as

$$\begin{aligned} & m_1\left[\tilde{T}\left(m_1,m_2,m_3,\cdots,m_J\right)-\tilde{T}\left(m_1-1,m_2,m_3+1,\cdots,m_J\right)\right] \\ = \; & m_1\left[\tilde{T}\left(m_1-1,m_2+1,m_3,\cdots,m_J\right)-\tilde{T}\left(m_1-1,m_2,m_3+1,\cdots,m_J\right)\right] \\ & +\sum_{j=3}^{J}m_j\left[\tilde{T}\left(m_j-1,m_2+1,m_1,\cdots,m_J\right)-\tilde{T}\left(m_j-1,m_1+1,m_2,\cdots,m_J\right)\right] \end{aligned} \tag{34}$$

Note that the left hand side in (34) is 1 by (30) and the difference in the number of non-zero elements in $\tilde{T}$ for each difference term on the right-hand side is exactly 1. For example, $\nu\left(m_2+1,m_3,\cdots,m_J\right)=\nu\left(m_2,m_3+1,\cdots,m_J\right)+1$ and $\nu\left(m_2+1,m_1,\cdots,m_J\right)=\nu\left(m_1+1,m_2,\cdots,m_J\right)+1$. Therefore, if

$$\tilde{T}\left(m,\nu\right)-\tilde{T}\left(m,\nu-1\right)=c$$

for some c independent of m and ν, (34) is satisfied. In this case, (34) becomes

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

Therefore, we extract a condition for (34):

$$\tilde{T}\left(m,\nu\right)-\tilde{T}\left(m,\nu-1\right)=\frac{1}{R} \tag{35}$$

for all m and ν. To summarize, conditions (31), (33), and (35) are sufficient for (27).

Now, if we define

$$\tilde{T}(m_j, m_{-j}) = -\sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\nu(m_{-j})}{R}, \tag{36}$$

this choice of $\tilde{T}$ satisfies conditions (31), (33), and (35), and hence the equation (27) follows, completing the proof. On the other hand, it is also worth noting that the conditions (31), (33), and (35) for $\tilde{T}$ also lead to the form of (36) up to an affine transform. This is the way the transform T_R is determined. ■

Proof of Lemma 2 : We first consider the case of $J = 3$. Recall

$$\Lambda_R(p, p_0; T_R) = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \begin{pmatrix} R \\ m_1, m_2, m_3 \end{pmatrix} T_{R,j}(m_1, m_2, m_3)\, p_1^{m_1} p_2^{m_2} p_3^{m_3}$$

where $T_{R,j}(m_1, m_2, m_3) = \tilde{T}(m_j; m_{-j})$ with $\tilde{T}$ as defined in (36). Recall that λ_j denotes the derivative of $\Lambda_R(p, p_0; \{T_{R,j}\})$ with respect to p_j, so that

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

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

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

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

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

However, by the definition of $T_{R,j}$, we have

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

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

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

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

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

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

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

or

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

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,

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

We define $\lambda_{ij} = \partial\lambda_i/\partial p_j$. Then $\partial\left(\lambda_1 - \lambda_3\right)/\partial p_1$ is equal to

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

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

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

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

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

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 also that $\lambda_{13}-\lambda_{33}=\lambda_{23}-\lambda_{33}$. We write

$$\begin{aligned}
\Lambda_R(p,p_0;T_R) &= \Lambda_R((p_1,p_2,p_3),p_0;T_R) \\
&= \Lambda_R((p_1,p_2,1-p_1-p_2),p_0;T_R) \\
&= \tilde{\Lambda}_R(\tilde{p},p_0;T_R).
\end{aligned}$$

Viewing this as a function of (p_1,p_2), we find that its Hessian matrix is given by

$$\begin{aligned}
H_3 &\equiv \begin{pmatrix} \lambda_{11}-\lambda_{31}-(\lambda_{13}-\lambda_{33}) & \lambda_{21}-\lambda_{31}-(\lambda_{23}-\lambda_{33}) \\ \lambda_{12}-\lambda_{32}-(\lambda_{13}-\lambda_{33}) & \lambda_{22}-\lambda_{32}-(\lambda_{23}-\lambda_{33}) \end{pmatrix} \\
&= \begin{pmatrix} \lambda_{11}-\lambda_{31} & \lambda_{21}-\lambda_{31} \\ \lambda_{12}-\lambda_{32} & \lambda_{22}-\lambda_{32} \end{pmatrix}-\begin{pmatrix} \lambda_{13}-\lambda_{33} & \lambda_{13}-\lambda_{33} \\ \lambda_{13}-\lambda_{33} & \lambda_{13}-\lambda_{33} \end{pmatrix} \\
&= \begin{pmatrix} \lambda_{11}-\lambda_{31} & 0 \\ 0 & \lambda_{22}-\lambda_{32} \end{pmatrix}-(\lambda_{13}-\lambda_{33})\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.
\end{aligned} \tag{37}$$

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

$$\begin{aligned}
\lambda_{11}-\lambda_{31} &= -\frac{p_{10}}{p_1^2}+\frac{p_{10}}{p_1^2}\left(1-p_1\right)^{R-1}-\left(1-\frac{p_{10}}{p_1}\right)\left(R-1\right)\left(1-p_1\right)^{R-2} \\
&= \frac{p_{10}}{p_1^2}h(R)-(R-1)(1-p_1)^{R-2},
\end{aligned} \tag{38}$$

where $h(R)=(1-p_1)^{R-1}+p_1(R-1)(1-p_1)^{R-2}-1$. We rewrite

$$h(R)=e^{(R-2)\log(1-p_1)}\left(1+(R-2)p_1\right)-1.$$

Now the first order derivative of $h(R)$ is given by

$$\begin{aligned} h'(R) &= e^{(R-2)\log(1-p_1)}\{\log(1-p_1)(1+(R-2)p_1)+p_1\} \\ &\leq e^{(R-2)\log(1-p_1)}\{\log(1-p_1)+p_1\} \leq 0 \end{aligned}$$

when $R \geq 2$, because $\log(1-p_1)+p_1 \leq 0$ for $p_1 \in [0,1]$. Therefore, $h(R)$ is decreasing in R. In view of (38), this implies that

$$\lambda_{11} - \lambda_{31} \leq \max\left\{\frac{p_{10}}{p_1^2}h(2)-1, \frac{p_{10}}{p_1^2}h(3)\right\} = \max\{-1,-p_{10}\} \leq -p_{10} \leq -\varepsilon.$$

Similarly, $\lambda_{22} - \lambda_{32} \leq -\varepsilon$ and $\lambda_{33} - \lambda_{13} \leq -\varepsilon$. Therefore, from (37),

$$a^\top H_3 a \leq -2\varepsilon\left(\sum_{j=1}^{J-1} a_j^2\right).$$

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

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

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 Theorem 1 : Take $\theta \in \Theta$ such that $||\theta - \theta_0|| > \eta$, for some $\eta > 0$. Then

$$P\{||p_i(\theta) - p_i(\theta_0)|| > 0\} > 0.$$

Now, we focus on $\mathbf{E}[\Lambda_R(p_i(\theta), p_i(\theta_0); T_R)]$. We write $\tilde{p}_i(\theta) = (p_1(X_i;\theta), \cdot\cdot\cdot, p_{J-1}(X_i;\theta))^\top$ and $\tilde{\lambda}_j = \partial\tilde{\Lambda}_R(\tilde{p}, p_0; T_R)/\partial\tilde{p}_j$, where $\tilde{\Lambda}_R$ is defined in (19). Then

$$\begin{aligned} & \mathbf{E}[\Lambda_R(p_i(\theta), p_i(\theta_0); T_R)] - \mathbf{E}[\Lambda_R(p_i(\theta_0), p_i(\theta_0); T_R)] \\ = \; & \mathbf{E}[\tilde{\Lambda}_R(\tilde{p}_i(\theta), p_i(\theta_0); T_R)] - \mathbf{E}[\tilde{\Lambda}_R(\tilde{p}_i(\theta_0), p_i(\theta_0); T_R)]. \end{aligned}$$

By the mean value theorem, the last difference is written as

$$\begin{aligned} & \sum_{j=1}^{J-1} \mathbf{E}\left[\tilde{\lambda}_j(\tilde{p}_i(\theta_0), p_i(\theta_0); T_R)(\tilde{p}_j(X_i;\theta) - \tilde{p}_j(X_i;\theta_0))\right] \\ & + \mathbf{E}\left[(\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))^\top \left(\frac{\partial^2\tilde{\Lambda}_R(p_i^*, p_i(\theta_0); T_R)}{\partial p \partial p^\top}\right)(\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))\right], \end{aligned}$$

where p_i^* lies on the line segment connecting $\tilde{p}_i(\theta)$ and $\tilde{p}_i(\theta_0)$. The first term is zero because for all $1 \leq j < J$,

$$\tilde{\lambda}_j(\tilde{p}_i(\theta_0), p_i(\theta_0); T_R) = \lambda_j(p_i(\theta_0), p_i(\theta_0); T_R) - \lambda_J(p_i(\theta_0), p_i(\theta_0); T_R) = 0$$

and with the transform T_R, it is shown in Lemma 1 that (11) is satisfied. By Lemma 2, we have

$$\begin{aligned}
&\mathbf{E}\left[(\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))^\top \left(\frac{\partial^2 \tilde{\Lambda}_R(p_i^*, p_i(\theta_0); T_R)}{\partial p \partial p^\top}\right)(\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))\right] \\
\geq &\left(\sum_{j=1}^{J-1} \mathbf{E}\left[(\tilde{p}_{ij}(\theta) - \tilde{p}_{ij}(\theta_0))^2\right]\right)\varepsilon_p.
\end{aligned}$$

The last term is positive because $P\{\tilde{p}_{ij}(\theta) \neq \tilde{p}_{ij}(\theta_0)\} > 0$. ■

The proofs of Theorems 2-3 below require the following preliminary results. Define for $\theta \in \Theta$,

$$T_{ij}^*(\theta) \equiv T_R(m_{ij}^*(\theta), m_{-ij}^*(\theta)) \text{ and } \Delta_{ij}(\theta) \equiv T_{ij}^*(\theta) - T_{ij}^*(\theta_0). \tag{39}$$

LEMMA A1: *Suppose that Assumptions 1 and 2 hold. Then there exists $C > 0$ that does not depend on R such that for any $\delta > 0$,*

$$\inf_{\theta \in \Theta \setminus B(\theta_0;\delta)} \mathbf{E} l_{n,R}^*(\theta) - \mathbf{E} l_{n,R}^*(\theta_0) \geq C\delta^2. \tag{40}$$

PROOF: From (9) and from the proof of Theorem 1, it is satisfied that

$$\inf_{\theta \in \Theta \setminus B(\theta_0;\delta)} \mathbf{E} l_{n,R}^*(\theta) - \mathbf{E} l_{n,R}^*(\theta_0) \geq \inf_{\theta \in \Theta \setminus B(\theta_0;\delta)} \left(\sum_{j=1}^{J-1} \mathbf{E}\left[(\tilde{p}_{ij}(\theta) - \tilde{p}_{ij}(\theta_0))^2\right]\right) C\varepsilon_p,$$

where $\tilde{p}_i(\theta) = (p_1(X_i;\theta), \cdots, p_{J-1}(X_i;\theta))^\top$ and $\varepsilon_p > 0$ is the constant in Assumption 1(iii). The desired result follows by applying Assumption 2(iii)(b) to the above expectation. ■

PROOF OF THEOREM 2 : We first show the consistency of the estimator. Given the identification result in Theorem 1, it suffices for consistency to show that for each $\varepsilon > 0$,

$$\text{limsup}_{n\to\infty} P\left\{\sup_{\theta \in \Theta} \left|l_{n,R}^*(\theta) - l_R(\theta)\right| > \varepsilon\right\} = 0, \tag{41}$$

where $l_R(\theta) = \mathbf{E} l_{n,R}^*(\theta)$. Since Θ is compact, for each $\delta > 0$, we have a finite, say, $C\delta^{-d}$ number of δ-balls centered at θ_m, $m = 1, \cdots, M_\delta$, which cover Θ, where $C > 0$ is a fixed constant and $M_\delta = C\delta^{-d}$. First, bound

$$P\left\{\sup_{\theta \in \Theta} \left|l_{n,R}^*(\theta) - l_R(\theta)\right| > \varepsilon\right\} \leq A_{n,R}(\varepsilon) + B_{n,R}(\varepsilon),$$

where

$$\begin{aligned}
A_{n,R}(\varepsilon) &= P\left\{\sup_{\tilde{\theta} \in \Theta} \sup_{\theta \in B(\tilde{\theta};\delta)} \left|l_{n,R}^*(\theta) - l_R(\theta) - \{l_{n,R}^*(\tilde{\theta}) - l_R(\tilde{\theta})\}\right| > \frac{\varepsilon}{2}\right\} \text{ and} \\
B_{n,R}(\varepsilon) &= P\left\{\max_{1 \leq m \leq M_\delta} \left|l_{n,R}^*(\theta_m) - l_R(\theta_m)\right| > \frac{\varepsilon}{2}\right\}.
\end{aligned}$$

As for $A_{n,R}(\varepsilon)$, let $\tilde{T}$ be as in (36) and write $l^*_{n,R}(\theta)$ as

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

Since $m_{jR}(X_i,\eta^*_i;\theta)$ is an integer,

$$|l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta})| \leq \frac{CR}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} 1\left\{\left|\frac{m_{jR}(X_i,\eta^*_i;\theta)}{R} - \frac{m_{jR}(X_i,\eta^*_i;\tilde{\theta})}{R}\right| \geq \frac{1}{R}\right\},$$

where $C > 0$ does not depend on R. Therefore, for each $\tilde{\theta} \in \Theta$,

$$\begin{aligned}
&P\left\{\sup_{\tilde{\theta}\in\Theta}\sup_{\theta\in B(\tilde{\theta};\delta)}\left|l^*_{n,R}(\theta) - l_R(\theta) - \{l^*_{n,R}(\tilde{\theta}) - l_R(\tilde{\theta})\}\right| > \frac{\varepsilon}{2}\right\} \\
\leq \quad & 2JP\left\{\sup_{\tilde{\theta}\in\Theta}\sup_{\theta\in B(\tilde{\theta};\delta)}\sup_{1\leq j\leq J}\frac{R}{n}\sum_{i=1}^{n} 1\left\{\left|\frac{m_{jR}(X_i,\eta^*_i;\theta)}{R} - \frac{m_{jR}(X_i,\eta^*_i;\tilde{\theta})}{R}\right| \geq \frac{1}{R}\right\} \geq \frac{\varepsilon}{CJ}\right\} \\
\leq \quad & \frac{2CRJ^2}{\varepsilon}P\left\{\sup_{\tilde{\theta}\in\Theta}\sup_{\theta\in B(\tilde{\theta};\delta)}\sup_{1\leq j\leq J}\left|\frac{m_{jR}(X_i,\eta^*_i;\theta)}{R} - \frac{m_{jR}(X_i,\eta^*_i;\tilde{\theta})}{R}\right| \geq \frac{1}{R}\right\},
\end{aligned}$$

for some $C > 0$ that does not depend on R. The second inequality uses Markov's inequality. By Assumption 2(ii), the last probability is bounded by

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

for some $C > 0$. Hence by sending $\delta \to 0$, we obtain that $\limsup_{n\to\infty}A_{n,R}(\varepsilon) = 0$.

As for $B_{n,R}(\varepsilon)$, we bound it by

$$\begin{aligned}
&M_\delta\max_{1\leq m\leq M_\delta} P\left\{\left|l^*_{n,R}(\theta_m) - l_R(\theta_m)\right| > \frac{\varepsilon}{2}\right\} \qquad (42)\\
\leq \quad & M_\delta\max_{1\leq m\leq M_\delta}\frac{4}{\varepsilon^2}\mathbf{E}\left[\left(l^*_{n,R}(\theta_m) - l_R(\theta_m)\right)^2\right] \leq \max_{1\leq m\leq M_\delta}\frac{4M_\delta}{n\varepsilon^2}\mathbf{E}\left[D_{ij}T^2_{R,j}(m_i(\theta_m))\right] \leq \frac{CM_\delta R^2}{n\varepsilon^2}.
\end{aligned}$$

The last bound vanishes as $n \to \infty$ (while R is fixed.) Hence we have established (41). This completes the proof of the consistency of $\hat{\theta}$.

Now we turn to the rate of convergence. Since we have Lemma A1, in view of Theorem 3.2.5 of van der Vaart and Wellner (1996), it suffices for the completion of the proof to investigate the continuity modulus of the process $\sqrt{n}\{l^*_{n,R}(\theta) - \mathbf{E}l^*_{n,R}(\theta)\}$. Given our definition of $\tilde{T}$, 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, for each $\theta_1 \in \Theta$ and for $\delta > 0$,

$$\begin{aligned}
&\mathbf{E}\left[\sup_{\theta:||\theta-\theta_1||\le\delta}\left|h_R(p^*_{jR}(X_i,\theta),\nu(m^*_{-ij}(\theta)))-h_R(p^*_{jR}(X_i,\theta_1),\nu(m^*_{-ij}(\theta_1)))\right|^2\right] &(43)\\
\le\ & CR\cdot P\left\{\sup_{\theta:||\theta-\theta_1||\le\delta}\sup_{2\le r\le R}\left|\delta_j(X_i,\eta^*_{i,r};\theta)-\delta_j(X_i,\eta^*_{i,r};\theta_1)\right|\ge 1\right\}\\
\le\ & CR\cdot\mathbf{E}\left[\sup_{\theta:||\theta-\theta_1||\le\delta}\sup_{2\le r\le R}\left|\delta_j(X_i,\eta^*_{i,r};\theta)-\delta_j(X_i,\eta^*_{i,r};\theta_1)\right|^2\right]\le CR\delta,
\end{aligned}$$

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 $\mathcal{G}_\delta = \{\gamma(\cdot,\cdot,\cdot;\theta) : \theta\in B(\theta_0;\delta)\}$. Let G_δ be an envelope of $\mathcal{G}_\delta$. By the maximal inequality in terms of the bracketing entropy (e.g. Pollard (1989), van der Vaart (1996)), we have

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

for some $C>0$. From the proof of Theorem 3 in Chen, Linton, and van Keilegom (2003) and the result of (43), the last integral is bounded by

$$\begin{aligned}
&\int_0^{||G_\delta||_2}\sqrt{1+\log N((C\varepsilon/||G_\delta||_2)^2,\Theta,||\cdot||)}d\varepsilon &(44)\\
=\ & ||G_\delta||_2\int_0^1\sqrt{1+\log N((C\varepsilon)^2,\Theta,||\cdot||)}d\varepsilon\le C||G_\delta||_2.
\end{aligned}$$

By the result of (43), we can take G_δ such that $||G_\delta||_2\le C\delta^{1/2}$, and conclude that the continuity modulus of $l^*_{n,R}(\theta)$ in θ turns out to be $O(\delta^{1/2})$. Now, by Theorem 3.2.5 of van der Vaart and Wellner (1996), 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 desired result of the theorem. ■

Lemma A2 : *Suppose that Assumptions 1 and 2 hold. Then for each $\varepsilon>0$, we have*

$$\sup_{x\in\mathcal{X}}P\left\{\sup_{\theta\in\Theta}\left|\frac{m_{jR}(x,\theta)}{R}-p_j(x,\theta)\right|>\varepsilon\right\}\le C\varepsilon^{2d}R^d\exp\left(-\varepsilon^2 R\right),$$

for some $C>0$ that does not depend on R or n.

Proof : Fix x in the support $\mathcal{X}$ of X_i. Define $\mathcal{F}_x=\{\delta_j(x,\cdot,\theta):\theta\in\Theta\}$ and let $||f||_2=\sqrt{\int fdP_\eta}$, where P_η is the distribution of η_i. Then by Assumption 2(ii) and from the proof of Theorem 3 in Chen, Linton, and van Keilegom (2003), for all $\varepsilon>0$,

$$N_{[]}(\varepsilon||F_x||_2,\mathcal{F}_x,||\cdot||_2)\le N((C\varepsilon||F_x||_2)^2,\Theta,||\cdot||)\le C\varepsilon^{-2d}||F_x||_2^{-2d},$$

where F_x is an envelope of $\mathcal{F}_x$. We can take F_x to be a constant function at 1. Hence the above bound does not depend on x in the support of X_i. The desired bound of the lemma now follows from Theorem 2.14.9 of van der Vaart and Wellner (1996). ■

LEMMA A3: *For any* $p \in (0,1)$, *there exists* $C > 0$ *that does not depend on* R *or* p *such that*

$$\left| -\sum_{s=0}^{R-Rp-1} \frac{1}{R-s} - \log\left(p + R^{-1}\right) \right| \leq \frac{C}{(p+R^{-1})^2 R} \text{ and}$$

$$\sup_{p,p' \in (\varepsilon, 1-\varepsilon)} \left| -\sum_{s=0}^{R-Rp-1} \frac{1}{R-s} + \sum_{s=0}^{R-Rp'-1} \frac{1}{R-s} - \left\{\log\left(p+R^{-1}\right) - \log\left(p'+R^{-1}\right)\right\} \right| \leq \frac{C|p-p'|}{(\min\{p,p'\}+R^{-1})^2 R}.$$

PROOF: As for the first statement, note that when $m \in \{0, 1, \cdots, R\}$,

$$-\sum_{s=0}^{R-Rp-1} \frac{1}{R-s} = -\frac{1}{R}\sum_{s=0}^{R} \frac{1\{s/R \leq 1-p-1/R\}}{1-s/R}.$$

The last sum is a Riemann sum. From the error bound for the Riemann sum approximation of an integral (e.g. Kythe and Schäferkrotter (2005, p.46)), we have

$$\left| -\frac{1}{R}\sum_{s=0}^{R} \frac{1\{s/R \leq 1-p-1/R\}}{1-s/R} + \int_0^{1-p-1/R} \frac{1}{1-u} du \right| \leq \frac{C}{(p+R^{-1})^2 R}.$$

As for the second statement, we obtain similarly the following bound:

$$\left| -\frac{1}{R}\sum_{s=0}^{R} \frac{1\{s/R \leq 1-p-1/R\} - 1\{s/R \leq 1-p'-1/R\}}{1-s/R} + \int_{1-p'-1/R}^{1-p-1/R} \frac{1}{1-u} du \right| \leq \frac{C|p-p'|}{(\min\{p,p'\}+R^{-1})^2 R}.$$

■

LEMMA A4: *Suppose that Assumptions 1 and 2 hold. Then for any* $\delta > 0$,

$$\mathbf{E}\left[\sup_{\theta \in B(\theta_0;\delta)} \sqrt{n} \left| l_{n,R}^*(\theta) - l_{n,R}^*(\theta_0) - \mathbf{E} l_{n,R}^*(\theta) + \mathbf{E} l_{n,R}^*(\theta_0) \right| \right] \leq C v_R(\delta),$$

where $C > 0$ *is a constant that does not depend on* R, *and*

$$v_R(\delta) \equiv \sqrt{\frac{\delta}{R}} + \delta + \frac{R^{d/2+1}}{\exp(CR)}. \tag{45}$$

PROOF: Let $p_{ij}^*(\theta) = p_{jR}^*(X_i, \theta)$ where $p_{jR}^*(X_i, \theta)$ is defined in (5), and

$$\tilde{T}_{ij}^*(\theta) \equiv \log(p_{ij}^*(\theta) + R^{-1}) - \log(p_{ij}^*(\theta_0) + R^{-1}).$$

Fix small $C_0 > 0$ such that

$$\begin{aligned} \sup_{x\in\mathcal{X}} P\left\{\sum_{j=1}^{J} p_{ij}(\theta) \in (C_0, 1-C_0) \text{ for all } \theta\in\Theta\ |X_i = x\right\} &= 1 \text{ and} \quad (46)\\ \sup_{x\in\mathcal{X}} P\left\{\sup_{\theta\in\Theta}\sum_{j=1}^{J} \left|p_{ij}^*(\theta) - p_{ij}(\theta)\right| > C_0/2\ |X_i = x\right\} &\leq CR^d\exp(-CR), \end{aligned}$$

for some $C > 0$ that does not depend on R or n. Such constants exist by Assumption 2(iii)(a) and Lemma A2. We fix R^* and n^* such that for all $R > R^*$, $C_0/2 + R^{-1} \leq C_0$.

Define

$$1_R(x,\theta) = 1\left\{\sum_{j=1}^{J}\left|p_{jR}^*(x,\theta) - p_j(x,\theta)\right| \leq \frac{C_0}{2}\right\}. \quad (47)$$

Recalling $\Delta_{ij}(\theta)$ defined in (39), we find that

$$\begin{aligned} \mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|\Delta_{ij}(\theta)|^2\right] &\leq 2\int \mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|\Delta_{ij}(\theta)|^2 1_R(x,\theta)|X_i = x\right] dF_X(x) \quad (48)\\ &\quad + CR^2\int \mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}(1 - 1_R(x,\theta))\,|X_i = x\right] dF_X(x). \end{aligned}$$

By (46), the last term is bounded by $CR^{d+2}\exp(-CR)$ for some $C > 0$. As for the first integral, using Lemma A3, we bound the integral by

$$\int \mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|\tilde{T}_{ij}^*(\theta)|^2 1_R(x,\theta)|X_i = x\right] dF_X(x) + \frac{C\delta}{(C_0/2 + R^{-1})^2 R}, \quad (49)$$

for some $C > 0$. By applying the mean value theorem for the logarithmic function, the above integral is bounded by

$$\begin{aligned} &\frac{C}{C_0/2 + R^{-1}}\cdot\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|p_{ij}^*(\theta) - p_{ij}^*(\theta_0)|^2\right]\\ \leq\ &\frac{2C}{C_0}\cdot\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|p_{ij}^*(\theta) - p_{ij}(\theta) - \{p_{ij}^*(\theta_0) - p_{ij}(\theta_0)\}|^2\right]\\ &+\frac{2C}{C_0}\cdot\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}|p_{ij}(\theta) - p_{ij}(\theta_0)|^2\right]. \end{aligned}$$

The last expectation is bounded by $C\delta^2$ for some $C > 0$ by Assumption 2(iv). Conditional on $X_i = x$, $\sqrt{R}\{p_{ij}^*(\theta) - p_{ij}(\theta)\}$ is an empirical process indexed by $\theta \in B(\theta_0;\delta)$. We use Theorem 2.14.5 of van der Vaart and Wellner (1996) to deduce that

$$\begin{aligned} &\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}(p_{ij}^*(\theta) - p_{ij}(\theta) - \{p_{ij}^*(\theta_0) - p_{ij}(\theta_0)\})^2|X_i = x\right]\\ \leq\ &2\left(\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{\theta\in B(\theta_0;\delta)}\left|p_{ij}^*(\theta) - p_{ij}(\theta) - \{p_{ij}^*(\theta_0) - p_{ij}(\theta_0)\}\right|\,|X_i = x\right]\right)^2 + \frac{C\delta}{R}, \end{aligned}$$

because we can take an envelope of the functions indexing the empirical process $p_{ij}^*(\theta) - p_{ij}(\theta) - \{p_{ij}^*(\theta_0) - p_{ij}(\theta_0)\}$ whose L_2-norm is bounded by $C\delta^{1/2}$ for some $C > 0$ using Assumption 2(ii). As for the leading term, utilizing the same envelope with L_2-norm bounded by $C\delta^{1/2}$ and the arguments in the proof of Theorem 3 of Chen, Linton, and van Keilegom (2003), we find that

$$\begin{aligned} & \sup_{x\in\mathcal{X}} \mathbf{E}\left[\sup\nolimits_{\theta\in B(\theta_0;\delta)}|p_{ij}^*(\theta)-p_{ij}(\theta)-\{p_{ij}^*(\theta_0)-p_{ij}(\theta_0)\}||X_i=x\right] \\ \leq\ & \frac{C_1}{\sqrt{R}}\int_0^{C\sqrt{\delta}}\sqrt{1+\log N((\varepsilon/\{C\sqrt{\delta}\})^2,\Theta,||\cdot||)}d\varepsilon \leq C_2\sqrt{\frac{\delta}{R}}, \end{aligned} \tag{50}$$

for some constants $C_1, C_2 > 0$, using Assumption 2(ii) and the maximal inequality. Therefore, for some $C > 0$,

$$\int \mathbf{E}\left[\sup\nolimits_{\theta\in B(\theta_0;\delta)}\left|D_{ij}\tilde{T}_{ij}^*(\theta)\right|^2 1_R(x,\theta)|X_i=x\right]dF_X(x) \leq C\left\{\frac{\delta}{R}+\delta^2\right\}.$$

Subsuming the last term in (49) above into $C\delta$ in the above bound, we conclude that

$$\sqrt{\mathbf{E}\left[\sup\nolimits_{\theta\in B(\theta_0;\delta)}|\Delta_{ij}(\theta)|^2\right]} \leq C\left\{\sqrt{\frac{\delta}{R}}+\delta+\frac{R^{d/2+1}}{\exp(CR)}\right\},$$

for some $C > 0$. This bound reveals an L_2-bound for an envelope for the class of functions indexing the empirical process, $\frac{1}{\sqrt{n}}\sum_{i=1}^n\{\Delta_{ij}(\theta)-\mathbf{E}\Delta_{ij}(\theta)\}$. Since we can obtain the same result replacing $\theta_0\in\Theta$ by any arbitrary $\theta\in\Theta$, the bound also reveals the local uniform L_2-continuity condition for this process. Using the maximal inequality and following the proof of Theorem 3 of Chen, Linton, and van Keilegom (2003) as in (44), we deduce that

$$\begin{aligned} & \mathbf{E}\left[\sup\nolimits_{\theta\in B(\theta_0;\delta)}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{j=1}^J\left(D_{ij}\tilde{T}_{ij}^*(\theta)-\mathbf{E}[D_{ij}\tilde{T}_{ij}^*(\theta)]\right)\right|\right] \\ \leq\ & C_1\int_0^{C_1 v_R(\delta)}\sqrt{1-C_1\log(\varepsilon/C_1 v_R(\delta))}d\varepsilon \leq C_1 v_R(\delta)\int_0^{C_1}\sqrt{1-C_1\log\varepsilon}d\varepsilon \leq C_2 v_R(\delta), \end{aligned}$$

for some constants $C_1, C_2 > 0$. ■

PROOF OF THEOREM 3 : Define $\bar{l}_{n,R}(\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^J D_{ij}\log(p_{ij}^*(\theta)+1/R)$. Let $p_{ij}^*(\theta) = m_{ij}(\theta)/R$ and $m_{ij}(\theta)$ is the simulated frequency with simulation number R. First, we show that $\hat{\theta}$ is consistent. For this, it suffices to show that

$$\sup_{\theta\in\Theta}\left|l_{n,R}^*(\theta)-l(\theta)\right| = o_P(1), \tag{51}$$

where $l(\theta) = \mathbf{E}l_n(\theta)$ and $l_n(\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^J D_{ij}\log p_j(X_i,\theta)$. For each $\varepsilon > 0$, we write

$$\begin{aligned} P\left\{\sup_{\theta\in\Theta}\left|l_{n,R}^*(\theta)-l(\theta)\right|>\varepsilon\right\} \leq\ & P\left\{\sup_{\theta\in\Theta}\left|l_{n,R}^*(\theta)-\bar{l}_{n,R}(\theta)\right|>\frac{\varepsilon}{3}\right\} \\ & +P\left\{\sup_{\theta\in\Theta}\left|\bar{l}_{n,R}(\theta)-l_n(\theta)\right|>\frac{\varepsilon}{3}\right\}+P\left\{\sup_{\theta\in\Theta}|l_n(\theta)-l(\theta)|>\frac{\varepsilon}{3}\right\}. \end{aligned} \tag{52}$$

We take $C_0 > 0$ and $1_R(x,\theta)$ as in (46) and (47). As for the first leading probability,

$$P\left\{\sup_{\theta\in\Theta}\left|l_{n,R}^*(\theta)-\bar{l}_{n,R}(\theta)\right|>\frac{\varepsilon}{3}\right\} \leq \frac{3}{\varepsilon}\mathbf{E}\left[\sup_{\theta\in\Theta}\left|l_{n,R}^*(\theta)-\bar{l}_{n,R}(\theta)\right|\right].$$

Similarly as in (48), we bound the last expectation by

$$\sum_{j=1}^{J}\int \mathbf{E}\left[\sup_{\theta\in\Theta}\left|T_{R,j}(m_{ij}^*(\theta))-\log\left(\frac{m_{ij}^*(\theta)+1}{R}\right)\right|1_R(x,\theta)|X_i=x\right]dF_X(x)+\frac{CR^{d+1}}{\exp(CR)},$$

for some $C>0$. The last term certainly vanishes as $R\to\infty$. By Lemma A2, the first term is bounded by CR^{-1} which again vanishes as $R\to\infty$.

We turn to the second probability on the right hand side of (52). We bound the second probability in (52) by

$$\begin{aligned}
&\frac{3}{\varepsilon}\sum_{j=1}^{J}\mathbf{E}\left[\sup_{\theta\in\Theta}\left|\log\left(\frac{m_{ij}^*(\theta)+1}{R}\right)-\log(p_{ij}(\theta))\right|\right]\\
\leq\ &\frac{3}{\varepsilon}\sum_{j=1}^{J}\int \mathbf{E}\left[\sup_{\theta\in\Theta}\left|\log\left(\frac{m_{ij}^*(\theta)+1}{R}\right)-\log(p_{ij}(\theta))\right|1_R(x,\theta)|X_i=x\right]dF_X(x)\\
&+\frac{3\log R}{\varepsilon}\sum_{j=1}^{J}\int \mathbf{E}\left[\sup_{\theta\in\Theta}\left(1-1_R(x,\theta)\right)|X_i=x\right]dF_X(x).
\end{aligned}$$

By (46), the last term is bounded by $CR^d(\log R)\exp(-CR)$ for some constant $C>0$, and hence vanishes as $R\to\infty$. On the other hand, the leading term is bounded by

$$\begin{aligned}
&C\cdot\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{\theta\in\Theta}\left|\{p_{ij}^*(\theta)+1/R\}-p_{ij}(\theta)\right||X_i=x\right] &(53)\\
=\ &C\cdot\sup_{x\in\mathcal{X}}\int_0^\infty P\left\{\sup_{\theta\in\Theta}\left|\{p_{ij}^*(\theta)+1/R\}-p_{ij}(\theta)\right|>v|X_i=x\right\}dv\\
\leq\ &C\int_0^\infty v^{2d}R^d\exp\left(-v^2R\right)dv=\frac{C}{\sqrt{R}}\int_0^\infty v^{2d}\exp\left(-v^2\right)dv,
\end{aligned}$$

by Lemma A2 and by change of variables. The last integral vanishes as $R\to\infty$. We conclude that the second probability on the right hand side of (52) vanishes as $R\to\infty$.

Consider the third probability on the right hand side of (52). One can easily show that this probability vanishes by the uniform law of large numbers applied to $\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^J D_{ij}\log p_j(X_i,\theta):\theta\in\Theta\}$, using Assumption 2. Hence we have established (51), and the estimator $\hat{\theta}$ is consistent.

Now, we show that $\hat{\theta}$ is $\sqrt{n}$-consistent. We follow the proof of Theorem 3.2.5 of van der Vaart and Wellner (1996). First, 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

$$P\left\{r_n||\hat{\theta}-\theta_0||>2^M\right\}\leq\sum_{\substack{j\geq M\\ 2^j\leq\eta r_n}}P\left\{\inf_{\theta\in R_{j,n}}l_{n,R}^*(\theta)-l_{n,R}^*(\theta_0)\leq 0\right\}+P\left\{2||\hat{\theta}-\theta_0||\geq\eta\right\}. \qquad (54)$$

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

$$\mathbf{E}l_{n,R}^*(\theta)-\mathbf{E}l_{n,R}^*(\theta_0)\geq\frac{C2^{2j-2}}{r_n^2}$$

by Lemma A1. By using Lemma A4, the sum of probabilities on the right-hand side in (54) is bounded by

$$
\begin{aligned}
& \sum_{\substack{j \geq M \\ 2^j \leq \eta r_n}} P\left\{ \sup_{\theta \in R_{j,n}} |l_{n,R}^*(\theta) - \mathbf{E} l_{n,R}^*(\theta) - l_{n,R}^*(\theta_0) + \mathbf{E} l_{n,R}^*(\theta)| \geq \frac{C 2^{2j-2}}{r_n^2} \right\} \qquad (55) \\
\leq \quad & C \sum_{\substack{j \geq M \\ 2^j \leq \eta r_n}} \frac{C\sqrt{n}}{2^{2j-2}} v_R \left(\frac{2^{j-1}}{\sqrt{n}} \right).
\end{aligned}
$$

Using (45), we obtain that for j such that $2^j \leq \eta r_n$,

$$
v_R \left(\frac{2^{j-1}}{\sqrt{n}} \right) \leq C \left\{ \frac{2^{j/2-1/2}}{n^{1/4}\sqrt{R}} + \frac{2^{j-1}}{\sqrt{n}} \right\},
$$

for some $C > 0$ from some large n on. Therefore, the last sum in (55) vanishes as $n \to \infty$ and then $M \to \infty$, and $\sqrt{n}$-consistency of the estimator $\hat{\theta}$ follows.

It remains to show that $\hat{\theta}$ has the same asymptotic linear representation as that of the MLE. Let $\bar{q}_R(W_i; \theta) = \sum_{j=1}^J D_{ij} \log p_{ij}^*(\theta)$, $q_R(W_i; \theta) = \sum_{j=1}^J D_{ij} T_{R,j}(m_{ij}(\theta))$, and $q(W_i; \theta) = \sum_{j=1}^J D_{ij} \log p_j(X_i; \theta)$. For $u \in \mathbf{R}^d$, let

$$
\begin{aligned}
Z_{n,R}(u) &= \sum_{i=1}^n q_R(W_i; \theta_0 + u/\sqrt{n}) - \sum_{i=1}^n q_R(W_i; \theta_0), \\
\bar{Z}_{n,R}(u) &= \sum_{i=1}^n \bar{q}_R(W_i; \theta_0 + u/\sqrt{n}) - \sum_{i=1}^n \bar{q}_R(W_i; \theta_0), \text{ and} \\
Z_n(u) &= \sum_{i=1}^n q(W_i; \theta_0 + u/\sqrt{n}) - \sum_{i=1}^n q(W_i; \theta_0).
\end{aligned}
$$

Note that for any compact U, uniformly over $u \in U$,

$$
Z_n(u) = \frac{u^\top}{\sqrt{n}} \sum_{i=1}^n \frac{\partial q(W_i; \theta_0)}{\partial \theta} + \frac{1}{2} u^\top \frac{\partial^2 \mathbf{E}\left[q(W_i; \theta_0)\right]}{\partial \theta \partial \theta^\top} u + o_P(1).
$$

Therefore, it suffices to show that for any fixed compact set U,

$$
\begin{aligned}
\sup_{u \in U} \left| Z_{n,R}(u) - \bar{Z}_{n,R}(u) \right| &= o_P(1) \text{ and} \\
\sup_{u \in U} \left| \bar{Z}_{n,R}(u) - Z_n(u) \right| &= o_P(1),
\end{aligned} \qquad (56)
$$

because we have already established the $\sqrt{n}$-consistency of $\hat{\theta}$ and, after sending $n, R \to \infty$, we can choose U arbitrarily large.

We take $C_0 > 0$ and $1_R(x, \theta)$ as in (46) and (47). We write simply $1_{iR} = 1_R(X_i, \theta_0)$. Using Lemmas A2 and A3 and following the arguments in (48), we can write

$$
\begin{aligned}
& \sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}\left\{T_{R,j}(m_{ij}(\theta_0+u/\sqrt{n}))-T_{R,j}(m_{ij}(\theta_0))\right\} \\
= & \sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}\left\{T_{R,j}(m_{ij}(\theta_0+u/\sqrt{n}))-T_{R,j}(m_{ij}(\theta_0))\right\} 1_{iR}+o_P(1) \\
= & \sum_{i=1}^{n}\sum_{j=1}^{J} D_{ij}\left\{\log(m_{ij}(\theta_0+u/\sqrt{n})/R)-\log(m_{ij}(\theta_0)/R)\right\} 1_{iR}+o_P(1),
\end{aligned}
$$

where $o_P(1)$ is uniform over $u \in U$. The last equality follows from Lemma A4 by bounding the expected absolute value of the difference in the sum involving $T_{R,j}$ and the sum involving log by

$$
\begin{aligned}
& \frac{Cn}{R}\sup_{x\in\mathcal{X}}\mathbf{E}\left[|p_{ij}^*(\theta_0+u/\sqrt{n})-p_{ij}^*(\theta_0)||X_i=x\right] \\
\leq & \frac{Cn}{R}\sup_{x\in\mathcal{X}}\mathbf{E}\left[|\Delta p_{ij}^*(u)-\Delta p_{ij}(u)||X_i=x\right]+\frac{Cn}{R}\sup_{x\in\mathcal{X}}\mathbf{E}\left[|\Delta p_{ij}(u)||X_i=x\right] \\
\leq & \frac{Cn}{R\sqrt{R}n^{1/4}}+\frac{Cn}{R\sqrt{n}}\rightarrow 0,
\end{aligned}
$$

as $n, R \rightarrow \infty$ with $\sqrt{n}/R \rightarrow 0$, where $\Delta p_{ij}^*(u) = p_{ij}^*(\theta_0+u/\sqrt{n}) - p_{ij}^*(\theta_0)$ and $\Delta p_{ij}(u) = p_{ij}(\theta_0+u/\sqrt{n}) - p_{ij}(\theta_0)$. Hence the first statement of (56) follows.

As for the second statement, define

$$
\begin{aligned}
\Delta\bar{q}_R(W_i;u/\sqrt{n}) & \equiv \bar{q}_R(W_i;\theta_0+u/\sqrt{n})-\bar{q}_R(W_i;\theta_0) \\
\Delta q(W_i;u/\sqrt{n}) & \equiv q(W_i;\theta_0+u/\sqrt{n})-q(W_i;\theta_0), \text{ and} \\
\Delta\Delta\bar{q}_R(W_i;u/\sqrt{n}) & \equiv \Delta\bar{q}_R(W_i;u/\sqrt{n})-\Delta q(W_i;u/\sqrt{n}).
\end{aligned}
$$

Using Lemma A2 and following the arguments in (48), we write

$$
\begin{aligned}
\bar{Z}_{n,R}(u)-Z_n(u) & = \sum_{i=1}^{n}\Delta\Delta\bar{q}_R(W_i;u/\sqrt{n}) \\
& = \sum_{i=1}^{n}\Delta\Delta\bar{q}_R(W_i;u/\sqrt{n})1_{iR}+O_P(nR^{d+1}\exp(-CR)).
\end{aligned}
\tag{57}
$$

Since $\sqrt{n}/R \rightarrow 0$, the last term vanishes as $n \rightarrow \infty$. As for the leading sum above, by expanding the logarithm in $\Delta\Delta\bar{q}_R(W_i;u/\sqrt{n})$, we find that

$$
\begin{aligned}
\sqrt{n}\Delta\Delta\bar{q}_R(W_i;u/\sqrt{n})1_{iR} & = \frac{\sqrt{n}\{p_{ij}^*(\theta_0+u/\sqrt{n})-p_{ij}(\theta_0+u/\sqrt{n})\}}{p_{ij}(\theta_0)}1_{iR} \\
& \quad -\frac{\sqrt{n}\{p_{ij}^*(\theta_0)-p_{ij}(\theta_0)\}}{p_{ij}(\theta_0)}1_{iR}+R_{1,i,n}(u),
\end{aligned}
\tag{58}
$$

for some remainder term $R_{1,i,n}(u)$. Accordingly, we write the leading sum in (57) as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\sqrt{n}\Delta\Delta p_{ij}^{*}(u)}{p_{ij}(\theta_0)}1_{iR}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{R_{1,i,n}(u)-\mathbf{E}\left[R_{1,i,n}(u)\right]\right\}+\sqrt{n}\mathbf{E}\left[R_{1,i,n}(u)\right], \tag{59}$$

where $\Delta\Delta p_{ij}^{*}(u)=\Delta p_{ij}^{*}(u)-\Delta p_{ij}(u)$. As for the leading sum, note that

$$\sqrt{n}\Delta\Delta p_{ij}^{*}(u)=\frac{\sqrt{n}}{R}\sum_{r=1}^{R}\left(\delta_{j,r}(X_i;u)-\mathbf{E}[\delta_{j,r}(X_i;u)|X_i]\right), \tag{60}$$

where $\delta_{j,r}(X_i;u)=\delta_j(X_i,\eta_{i,r}^{*};\theta_0+u/\sqrt{n})-\delta_j(X_i,\eta_{i,r}^{*};\theta_0)$. Using Assumption 2(ii), we find that

$$\begin{aligned}\sqrt{\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{u\in U}\left|\delta_{j,r}(X_i;u)\right|^2|X_i=x\right]} &\leq \frac{C}{n^{1/4}} \text{ and}\\ \sqrt{\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{v\in(0,\varepsilon)}\left|\delta_{j,r}(X_i;u+v)-\delta_{j,r}(X_i;u)\right|^2|X_i=x\right]} &\leq \frac{C\varepsilon^{1/2}}{n^{1/4}},\end{aligned}$$

and by Theorem 2.14.5 of van der Vaart and Wellner (1996), also that

$$\begin{aligned}&\sqrt{\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{u\in U}\left|\sqrt{n}\Delta\Delta p_{ij}^{*}(u)\right|^2|X_i=x\right]}\\ \leq\ & C_1\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{u\in U}\left|\sqrt{n}\Delta\Delta p_{ij}^{*}(u)\right||X_i=x\right]+\frac{C_1\sqrt{n}}{n^{1/4}\sqrt{R}}\leq\frac{C_2\sqrt{n}}{n^{1/4}\sqrt{R}},\end{aligned}$$

for some constants $C_1, C_2>0$. The last equality follows from the maximal inequality and the arguments similar to (50). Similarly, we also have for any $\varepsilon>0$ and any $u\in U$,

$$\begin{aligned}&\sqrt{\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{v\in(0,\varepsilon)}\left|\sqrt{n}\Delta\Delta p_{ij}^{*}(u+v)-\sqrt{n}\Delta\Delta p_{ij}^{*}(u)\right|^2|X_i=x\right]}\\ \leq\ & C_1\sup_{x\in\mathcal{X}}\mathbf{E}\left[\sup_{v\in(0,\varepsilon)}\left|\sqrt{n}\Delta\Delta p_{ij}^{*}(u+v)-\sqrt{n}\Delta\Delta p_{ij}^{*}(u)\right||X_i=x\right]+\frac{C_1\sqrt{n\varepsilon}}{n^{1/4}\sqrt{R}}\leq\frac{C_2\sqrt{n\varepsilon}}{n^{1/4}\sqrt{R}},\end{aligned}$$

for some constants $C_1, C_2>0$. This gives the local uniform L_2-continuity of the functions indexing the leading empirical process in (59). Hence we have

$$\mathbf{E}\left[\sup_{u\in U}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\sqrt{n}\Delta\Delta p_{ij}^{*}(u)}{p_{ij}(\theta_0)}\right|\right]\leq\frac{C\sqrt{n}\log n}{n^{1/4}\sqrt{R}}=\frac{Cn^{1/4}\log n}{\sqrt{R}}\to 0,$$

as $\sqrt{n}/R\to 0$ as $n, R\to\infty$.

As for the second and the third sums in (59), we follow a similar arguments to show that it is of smaller order than the leading sum (59). Details are omitted. ■

References

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

[2] Ben-Akiva, M., D. McFadden, M. Abe, U. Böckenholt, D. Bolduc, D. Gopinath, T. Morikawa et. al. (1997), "Modelling Methods for Discrete Choice Analysis. *Marketing Letters* 8, 273–286.

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

[4] Delgado, M. A., J. A. Rodriguez-Poo, and M. Wolf (2001), "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator," *Economics Letters* 73, 241-250.

[5] 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.

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

[7] 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.

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

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

[10] 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.

[11] Keane, M. and K. Wolpin (1994), "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence," *Review of Economics and Statsitics* 76, 648-672.

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

[13] Klein, R. and R. H. Spady (1993), "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421.

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

[15] Kythe, P. K. and M. R. Schäferkrotter (2005), *Handbook of Computational Integration*, Chapmall and Hall, New York.

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

[17] 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.

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

[19] 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.

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

[21] McFadden, D. L. and K. Train (2000), "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15, 447-470.

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

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

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

[25] Train, K. E. (2003), *Discrete Choice Methods with Simulations*, Cambridge University Press.

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

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

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