Simulated MLE for Discrete Choices using Transformed Simulated Frequencies

Donghoon Lee and Kyungchul Song

December 8, 2010

Abstract

While many methods of simulated likelihood for discrete choice models have been developed in the literature, most methods require the presence of additive errors that have normal or extreme value distributions. This paper proposes a new method of simulated likelihood that is easy to implement, free from simulation bias for each finite number of simulation, and yet flexible enough to accommodate a variety of model specifications beyond those of additive normal or logit errors. 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, Simulated Frequency, Cube-Root Asymptotics

JEL Classifications: C12, C14, C52.

---

1A previous version was titled, "A Consistent SMLE When the Simulation Number was Finite." We are indebted to Sean D. Campbell for his valuable comments and inputs. We are also grateful to the seminar participants at Columbia University, Econometric Society Meeting, Greater New York Econometrics Colloquium, Lehigh University and New York University for useful comments. We thank a Co-Editor and two anonymous referees for their thoughtful 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.

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

1 Introduction

Discrete choice models have long been very popular in applied researches across a wide range of empirical fields of economics. While a discrete choice model typically specifies the data 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 much wider spectrum of flexibility in modeling. (See Hajivassiliou and Ruud (1994), Stern (1997) and Gouriéroux and Monfort (1997) for a review of the literature and references therein.) Part of the popularity of simulation-based inference is perhaps due to the rise of structural econometric models in labor economics and industrial organizations. Such models involve various forms of random utilities, depending on the way heterogeneity in individual decision-making is modeled.

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 very efficient, these methods rely on the assumption that the latent processes take the form of additive normal errors. As important contributions in the literature, Hajivassiliou (1990) and McFadden and Hajivassiliou (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 McFadden and Hajivassiliou (1998) suggesed 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 take the form of additive multivariate normal errors. Another increasingly popular class of discrete choice models include mixed multinomial logit models (MMLM) (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.

There are many other types of multinomial choice models that do not necessarily have additive logit or probit errors and individual heterogeneity in the random utility takes a much more complicated form. (Keane and Wolpin (1994), Keane and Wolpin (1997). etc.) For example, in the case of dynamic discrete choice models...

When the model does not have additive logit or probit errors in the random utilities, the simulated frequency method of Lerman and Manski or their smoothed variants often
emerge as the sole feasible solution. As is well-documented, however, the use of simulated frequencies as in 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 that complements the existing methods. 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. 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 is summarized in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method (TSF)</td>
<td>No</td>
<td>Discontinuous</td>
<td>Not Required</td>
<td>No Bias</td>
</tr>
<tr>
<td>Lerman-Manski</td>
<td>Yes</td>
<td>Discontinuous</td>
<td>Not Required</td>
<td>Bias</td>
</tr>
<tr>
<td>Stern (1992)</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>GHK</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>MSS - SRC/GRS</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>MSS - SAR</td>
<td>No</td>
<td>Discontinuous</td>
<td>Required (Normal)</td>
<td>No Bias</td>
</tr>
<tr>
<td>Mixed Multinomial Logit</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Logit)</td>
<td>Bias</td>
</tr>
</tbody>
</table>

Table 1: Comparison of Simulated MLE Methods
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 researches in industrial organizations.

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 \( \sqrt{n} \), 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 slightly faster than \( \sqrt{n} \). This latter condition is only slightly stronger than the existing condition for many MSL estimators. (See e.g. Lerman and Manski (1981) and Gourieroux and Monfort (1997).) Under this same condition, the estimator achieves the asymptotic efficiency of MLE.

To illustrate the usefulness of our approach, we also performed a Monte Carlo simulation study based on a schooling choice model which involves heterogeneity in discount factor and ability. More specifically, the discount factor is assumed to be correlated with other observed individual characteristic and also an unobserved characteristic. The simulation methods considered in this study are, Lerman-Manski’s MSL and its smoothed version. 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 and discuss MSL. In section 3, we introduce transformed simulated frequency (TSF) and present our main result of identification of parameters for each finite simulation number. Section 4 establishes the asymptotic properties of the estimator. Section 5 is devoted to two examples. The first example concerns with static random utility models and the second one discusses the case when only cohort-level aggregate data are available. In Section 6, we present and discuss results from a Monte Carlo simulation study. Section 7 concludes. All the technical proofs are relegated to the appendix.
2 Discrete Choice Models and TSF

2.1 Methods of Simulated Likelihoods

We introduce a discrete choice model and notations. 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:

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

where \( X_i = (X_{i1}, \ldots, X_{iJ})' \) represents a vector of observed covariates, \( \eta_i = (\eta_{i1}, \ldots, \eta_{iJ})' \) a vector of unobserved variables, and \( \theta_0 \in \Theta \subset \mathbb{R}^d \) 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, \( \delta_j \) can be specified as follows,

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

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

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

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

In the case that \( p_j(x, \theta) \) has a closed form representation, the maximum likelihood estimation is straightforward. (e.g. Amemiya (1985).) However, the choice probability is often hard to evaluate, in particular when the number of choices is large and one wants to admit flexibility in specifying the joint distribution of \( \eta_i \).

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

\[
l_{n,R}^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \log(p_{jR}^*(X_i, \theta)).
\]
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.

### 2.2 TSF-Based Method

Many literatures have focused on developing a method to simulate the choice probabilities $p_j^*(X_i, \theta)$ in (3). When $u_k(X_i, \eta_i; \theta_0) = X_i\theta_0 + \eta_i$ with $\eta_i$ following a multivariate normal distribution, the method of GHK is known to be computationally very efficient. An alternative specification is to specify $\eta_i$ to have an extreme value distribution with coefficient $\theta_0$ replaced by a random coefficient that can be simulated from a known distribution, and apply the mixed multinomial logit simulation methods. Both specifications are convenient as the simulated choice probabilities are smooth in $\theta$, although they suffer from simulation bias for finite simulation numbers.

This paper uses simulated frequency to construct the simulated choice probabilities like Lerman and Manski (1981). While using simulated frequencies has disadvantage of involving discontinuous objective functions like other simulated frequency-based methods, it allows for a flexible modeling of random utilities that goes beyond additive normal or logit specifications. As long as one can generate simulated errors $\eta_i^* = (\eta_{i,1}^*, \ldots, \eta_{i,R}^*)'$ for $\eta_i$, one can simulate individual choices through the specification of $\delta_i$ in (1) and obtain simulated frequencies.

To define simulated frequencies, suppose that $R$ number of stochastic errors $\eta_{i,r}^*$, $r = 1, \ldots, R$, are drawn from the known distribution $F$ of $\eta_i$. We let $\delta_j(X_i, \eta_{i,r}^*; \theta)$, $r = 1, \ldots, R$, (taking values of 0 or 1) denote simulated choices for each value of $\theta$. The simulated frequency of each choice $j$ at simulation number $R$ is defined to be

$$
m_{j,R}(X_i, \eta_i^*; \theta) = \sum_{r=1}^{R} \delta_j(X_i, \eta_{i,r}^*; \theta)
$$

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

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

and $m_i(\theta) = (m_{1i}(\theta), \cdots, m_{ji}(\theta))'$. One of the difficulties that this paper addresses is the presence of simulation bias which arises when we use the simulated choice probabilities in (3). 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). This paper proposes an alternative function different from logarithm that eliminates the simulation bias entirely for each finite simulation number. More specifically, for each fixed $R$, this paper develops a transform $T_{R,j}(\cdot)$ of simulated frequencies, $j = 1, 2, \cdots, J$, $R = 1, 2, \cdots$, such that

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

i.e., the population objective function identifies $\theta_0$ for each $R$. Then for each $R$, an estimator of $\theta_0$ is naturally obtained by maximizing its sample analogue:

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

A transform that satisfies (5) under regular conditions turns out to be of the following form: for each $j = 1, 2, \cdots, R$, and $R = 2, 3, \cdots$,

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

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 transforms depend only on $J$ and $R$ which are known. This means that we do not have to estimate the transform $T_{R,j}$. 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 is not confined to a specific class of distributions of $\eta_i$. Furthermore, as we shall show later, the method overcomes two major difficulties of the simulated frequency method of Lerman-Manski: the zero probability problem and the simulation bias for a finite simulation number.
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 $X \in \mathbb{R}$ is specified as

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

where $X$ is drawn from the uniform distribution over $[-1, 1]$ and the true parameter $\theta_0$ is set to be zero.

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

$$Q^{TSF}_R(\theta) = \sum_{j=1}^{J} \mathbb{E}[D_{ij}T_{R,j}(m_i(\theta))]: \text{(TSF)}$$

$$Q^{L-M}_R(\theta) = \sum_{j=1}^{J} \mathbb{E}\left[D_{ij} \log\left(\frac{m_{ij}(\theta)}{R}\right)\right]: \text{(Lerman-Manski)}$$

$$Q^{MLE}_R(\theta) = \sum_{j=1}^{J} \mathbb{E}\left[D_{ij} \log(p(X_i; \theta))\right]: \text{(MLE)}.$$

Each panel represents results with a different simulation number $R$. Even when $R = 2$, the maximizer of $Q^{TSF}_R(\theta)$ and $Q^{MLE}_R(\theta)$ coincide, and this coincidence is maintained as $R$ increases. When $R$ is large, both the objective functions $Q^{TSF}_R(\theta)$ and $Q^{MLE}(\theta)$ coincide for all the values of $\theta$. This makes contrast with the objective function of Lerman-Manski. When $R$ is small, the objective function of Lerman-Manski is away from the true value $\theta_0 = 0$. This reflects the well-known fact that the SMLE of Lerman-Manski is inconsistent for a finite $R$. Only when $R$ becomes large, Lerman-Manski objective function becomes close to the true MLE objective function.

Unlike Lerman-Manski, the approach of TSF does not suffer from the zero-probability problem. With each finite sample size $n$ and finite simulation number $R$, TSF $T_R(m_j(\theta); j)$ always assumes a finite number regardless of the realizations of the simulated frequency $m_j(\theta)$. 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 $\log(p)$, the expected value of logarithm of simulated probabilities according to Lerman-Manski (i.e. $\mathbb{E}\log(m_i(\theta_0)/R)$), and the expected value of transformed simulated frequency (TSF) (i.e. $\mathbb{E}T_{R,j}(m_i(\theta_0))$) 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
Figure 1: Population Objective Functions: The objective function of TSF-based MSL has the same maximizer as that of MLE for each simulation number.
the given value of the choice probability $p$. Certainly, in the case of Lerman-Manski, the zero-probability problem is severe when the simulation number is small $R$, as is shown by steeply falling curves as we move $p$ close to zero. However, 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.

### 3 Main Results

#### 3.1 Identification

In this section, we provide the main result of this paper that the use of TSF in (7) identifies $\theta_0$ for each finite simulation number $R$. Let $m_{ij}(\theta)$ be as defined in (4) and $\hat{\theta}$ be as defined in (6). Theorem 1 below shows the use of $T_{R,j}$ in (7) leads to the identification of $\theta_0$ for each finite $R$. We introduce the following regularity conditions.

**Assumption 1**: (i) $\Theta$ is compact with an interior containing $\theta_0$ and for all $\theta \in \Theta$ and $x$ in the support of $X_i$, the choice probability $p_j(x; \theta)$ belongs to $S_j$. 
(ii) For each \( x \) in the support of \( X_i \) and for each \( j \in \{1, \cdots, J\} \), \( p_j(x; \theta) \) is twice-continuously differentiable at \( \theta = \theta_0 \) and for some \( \delta > 0 \),

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

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

Conditions in Assumption 1 are standard in the MLE of discrete choice models. Condition (i) requires that the choice probability function \( p(x, \theta) \) is well-defined for all \( \theta \in \Theta \). Condition (ii) is stronger than needed for identification. The twice differentiability is assumed envisaging the derivation of asymptotic normality in the next section. Condition (iii) is a minimal necessary condition used to identify \( \theta_0 \) from the choice probabilities.

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

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

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

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

### 3.2 A Heuristic Exposition on the Discovery of TSF

In this section, we explain the way the transform (7) is discovered. Let \( \mathbb{N}_R = \{0, 1, 2, \cdots, R\} \) and define

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

The set \( \mathbb{N}_{R,J} \) denotes the space of \( J \)-tuples where simulated frequencies \( m_i(\theta) \) realizes in. Define a simplex \( S_J = \{p \in [0,1]^J : \Sigma_j p_j = 1\} \). Also we write the conditional choice probability \( p_i(\theta) = p(X_i; \theta) \) for brevity, where \( p(X_i; \theta) = (p_1(X_i; \theta), \cdots, p_J(X_i; \theta))' \).

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 \( T_j(\cdot) : \mathbb{N}_{R,J} \rightarrow \mathbb{R}_+ \), we introduce a function
\[ \Lambda(p, p_0; T) : S_J \times S_J \rightarrow \mathbb{R}, \ T = (T_1, \cdots, T_J) , \] as follows:

\[ \Lambda(p, p_0; T) = \sum_{j=1}^{J} p_{j0} \mathbb{E}[T_j(M)], \]

where \( M \) is a random vector in \( \mathbb{R}^J \) having the multinomial distribution on \( \mathbb{N}_{R,J} \) with parameter \( (R, p) \). The function \( \Lambda \) 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.

Using \( \Lambda \), we write

\[ \sum_{j=1}^{J} \mathbb{E}[D_{ij}T_j(m_i(\theta))] = \sum_{j=1}^{J} \mathbb{E}[\Lambda(p_i(\theta), p_i(\theta_0); T)]. \]

The main idea of this paper is that we extract conditions for \( \Lambda \) such that

\[ \theta_0 = \arg \max_{\theta \in \Theta} \sum_{j=1}^{J} \mathbb{E}[\Lambda(p_i(\theta), p_i(\theta_0); T)]. \] (8)

First, we assume the interchangeability of the derivative and expectation. Then, the first order condition for (8) is written as

\[ \frac{\partial}{\partial \theta} \sum_{j=1}^{J} \mathbb{E}[\Lambda(p_i(\theta), p_i(\theta_0); T)] |_{\theta = \theta_0} \]

\[ = \sum_{j=1}^{J} \mathbb{E} \left[ \lambda_j(p_i(\theta_0), p_i(\theta_0); T) \frac{\partial p_j(X_i, \theta_0)}{\partial \theta} \right] = 0, \]

where for each \( j = 1, 2, \cdots, J, \)

\[ \lambda_j(p, p_0; T) = \frac{\partial \Lambda(p, p_0; T)}{\partial p_j}. \] (10)

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

\[ \sum_{j=1}^{J} \frac{\partial p_j(x, \theta)}{\partial \theta} = 0. \] (11)

This means that the first order condition in (9) immediately follows if \( \lambda_j(p_0, p_0; T) \) is the
same across j’s. In other words, for each \( p_0 \in S_J \),

\[ \lambda_j(p_0, p_0; T) = \lambda_k(p_0, p_0; T), \text{ for all } j, k = 1, 2, \cdots, J. \]  

(12)

The condition in (12) has an important merit of not depending on any aspect of the data generating process. It remains to search for \( T \) such that (12) is satisfied. The solution is given in the following algebraic result.

**Lemma 1:** Transforms \( T \) satisfy (12) if for each \( j = 1, \cdots, J \) and \( R = 2, 3, \cdots, T_j(m) = T_{R,j}(m) \), where

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

(13)

Here we take the summation \( \sum_{s=0}^{R-1} \) to be zero.

Lemma 1 is a pure algebraic result that does not involve any unknown specifics of the data generating process in the model. The condition (12) is only a necessary condition for (8). Later, we show that the choice of \( T = T_R \) guarantees the sufficient second order condition for the optimization problem in (8). The proof of Lemma 1 is algebraically involved and relegated to the appendix.

It may not be immediately clear how the choice of (13) is related to MLE with sufficiently large \( R \). To see this connection, note first that the simulated frequencies \( m_{ij}(\theta)/R \rightarrow p \) \( p_{ij}(\theta) \in (0, 1) \) with \( R \rightarrow \infty \), by the law of large numbers, where \( p_{ij}(\theta) = p_j(X_i; \theta) \). Also, note that

\[ 0 \leq \frac{1}{R} \sum_{k=1, k \neq j}^{J} 1 \{ m_k > 0 \} \leq \frac{J - 1}{R} \rightarrow 0 \]

as \( R \rightarrow \infty \). Finally, observe that

\[ \sum_{s=0}^{R-m_j(\theta)-1} \frac{1}{R - s} \rightarrow p \log(p_{ij}(\theta)). \]

This latter convergence is immediate as the sum on the left-hand side is a Rieman lower sum of \(- \int_{m_{ij}(\theta)/R}^{1} (1/x) dx \) and \( m_{ij}(\theta)/R \rightarrow p \) \( p_{ij}(\theta) \). Therefore,

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

as \( R \rightarrow \infty \) under regularity conditions. Hence the TSF population objective function converges to that of MLE as \( R \rightarrow \infty \).
3.3 Asymptotic Properties

In this section we investigate the asymptotic properties of the estimator \( \hat{\theta} \) defined in (6). The asymptotic properties of \( \hat{\theta} \) are developed for two separate cases: when \( R \) is fixed and when \( R \) tends to infinity jointly with the sample size \( n \). Let \( \mathcal{X} \) be the support of \( X_i \). We introduce the following assumptions.

**Assumption 2**: (i) \( \{(D_i, X_i, \eta_i)\}_{i=1}^n \) is i.i.d. from a common distribution.

(ii) For each \( \tilde{\theta} \in \Theta \), \( \sup_{x \in \mathcal{X}} E[\sup_{\theta \in B(\tilde{\theta}, \varepsilon)} |\delta_j(X_i, \eta_i; \tilde{\theta}) - \delta_j(X_i, \eta_i; \theta)|^2 | X_i = x] \leq C\varepsilon \), for some \( C > 0 \).

(iii) For some \( \delta > 0 \), \( \inf_{\theta \in B(\theta_0, \delta)} \inf_{x \in \mathcal{X}} p_j(x; \theta_0) > \varepsilon_p > 0 \), \( j = 1, \cdots, J \), for some \( \varepsilon_p > 0 \).

(iv) \( (X_i, \eta_i)_{i=1}^n \) and \( (X_i, \eta_i^*)_{i=1}^n \) are distributionally identical.

Condition (ii) controls the manner the random decision rule \( \delta_j(X_i, \eta_i; \theta) \) depends on \( \theta \) and \( (X_i, \eta_i, \gamma) \). The condition requires that the decision rule \( \delta \) is locally uniformly \( L_2 \)-continuous in \( \theta \) (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 \( \delta \). We present lower-level conditions in the case of random utility models in a later section (See Lemma 2 below.) Condition (iii) requires that the choice probabilities be bounded away from zero, and implies that \( p_j(X_i; \theta_0) < 1 - \varepsilon_p \) for each \( j = 1, \cdots, J \). Conditions in Assumption 2(ii) and (iii) can be weakened at the cost of introducing a more complicated procedure that involves a trimming sequence. Condition (iv) is certainly satisfied when \( X_i \) and \( \eta_i \) are independent and \( \eta_i^* \) are drawn i.i.d from \( F \), as commonly assumed in the simulation literature.

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

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

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

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

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

where \( V = \Omega^{-1} \) and

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

The rate condition \( \sqrt{n} \log(R)/R \to 0 \) is satisfied when \( R \) increases slightly faster than \( \sqrt{n} \). This condition is nearly close to the usual condition in the simulated likelihood literature. In particular, when \( \sqrt{n}/R \to 0 \), it is known that regardless of when one uses the simulation frequency of Lerman and Manski (1981) or the GHK simulator for \( p^*_{JR}(X_i, \theta) \) in (3), the estimator is consistent and asymptotically normal. (See Gourieroux and Monfort (1991)).

### 3.4 Computation of Standard Errors

As for standard errors, we suggest using a consistent estimator of the asymptotic covariance matrix in Theorem 3 based on large \( R \) asymptotics. It suffices for the computation of its standard error to evaluate the asymptotic covariance matrix formula once and for all, whereas the estimation of \( \hat{\theta} \) requires evaluation of the sample objective function for each optimization iteration. Therefore, using large \( R \) in computing the standard error does not increase the computational cost significantly.

One may consider an estimated version of \( \Omega \) in Theorem 3 by using the numerical derivatives. However, this asymptotic covariance matrix formula relies on information matrix equality which will not hold for finite simulation numbers. In this paper, we construct an estimator of covariance matrix that does not require information matrix equality. Although this alternative asymptotic covariance matrix formula is justified only when \( R \) goes to infinity, this choice may be a good alternative to the asymptotic covariance formula that requires information matrix equality.

Instead of computing numerical derivatives of the choice probability based on the simulated frequencies, we suggest using directly numerical derivatives of TSF: \( T_{R,j}(m_i(\theta)) \) in \( \theta \).
Define \( \varepsilon_n \) be a sequence such that \( \varepsilon_n \to 0 \) and \( \varepsilon_n/R \to \infty \). Then, we take \( \hat{g}_i \) to be the \( d \times 1 \) vector whose \( k \)-th entry is given by

\[
\frac{1}{\varepsilon_n} \sum_{j=1}^{J} D_{ij} \{ T_{R,j}(m_i(\hat{\theta} + \varepsilon_n e_k)) - T_{R,j}(m_i(\hat{\theta})) \},
\]

where \( e_k \) denotes the unit vector whose \( k \)-th entry is one and the other entries are zero. Using numerical derivatives of a discontinuous finite sample function to compute standard errors is not new here (e.g. see Sherman (1993)). Then we define

\[
S_R = \frac{1}{n} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i'.
\]

The standard error for the \( j \)-th parameter \( \theta_{0,j} \) in \( \theta_0 \) is taken to be the square-root of the \( j \)-th diagonal element of \( S_R^{-1} \). The covariance matrix estimator is consistent as \( n, R \to \infty \).

In the simulation studies, we compare the performance of asymptotic approximation using various different covariance matrix estimators.

4 Examples

4.1 Static Random Utility Models

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

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

is greater than zero. In this case, the decision rule \( \delta_j \) is defined by

\[
\delta_j(X_i, \eta_i; \theta) = 1 \{ \Delta_j(X_i, \eta_i; \theta) > 0 \}
\]

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

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

\[
\sup_{1 \leq j \leq J} \left| \mu_j(x, \theta) - \mu_j(x, \tilde{\theta}) \right| \leq C \| \theta - \tilde{\theta} \|.
\]
Furthermore, assume that the conditional density of $\eta_{ij} - \eta_{ik}$ given $X_i = x$ is bounded uniformly over $x$ in the support of $X$. Then the condition of Assumption 2(ii) holds.

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

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

Hence the $L_p$-continuity condition follows immediately when $B(\theta)$ and $\Gamma(\theta)$ are Lipschitz continuous in $\theta$ at $\theta_0$. When $X_i$ has an unbounded support, we may redefine $\tilde{u}(X_i, \eta_{ij}; \theta) = \Phi(A(\theta, \eta_{ij})'X_i)$ where $\Phi$ is a bounded strictly increasing function that is first order continuously differentiable with derivative $\phi$ such that $\sup_{x \in \mathbb{R}^{d_x}} \phi(x)||x|| < \infty$.

### 4.2 Simulated MLE with Cohort-Level Aggregate Data

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

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

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

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

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

16
and \(D(k) = (D_1(k), \ldots, D_J(k))'\). The variable \(D_j(k)\) indicates a proportion of agents in cohort \(k\) that have chosen the \(j\)-th choice. The econometrician observes only the cohort-level aggregate data \(\{D(k), X(k)\}_{k=1}^{K}\). The (infeasible) log-likelihood of the micro data after normalizing by \(n(k)\) is equal to

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

When the conditional distribution of the stochastic error \(\eta_{ij}(k)\) given \(X(k)\) is identical for each individual \(i\), the conditional probability \(P \{D_{ij}(k) = 1 | X(k), \theta\}\) is identical for all the individuals in the \(k\)-th cohort. This is the case when \(\{\eta_{ij}(k) : i = 1, \ldots, n(k), k = 1, \ldots, K\}\) is i.i.d. and independent of \(\{X(k) : k = 1, \ldots, K\}\). In this case, we can write the cohort-level likelihood as

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

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 \((\eta_{i1}(k), \ldots, \eta_{iJ}(k))\). Then, one draws \(R\) random sample from \(F\) to obtain \(\{\eta^*_r(k)\}_{r=1}^{R}\) where \(\eta^*_r(k) = (\eta^*_{r1}(k), \ldots, \eta^*_{rJ}(k))\). We define the simulated frequency

\[
m_{jr}(k, \theta) = \frac{1}{R} \sum_{r=1}^{R} \delta_j(X(k), \eta^*_{r,j}(k); \theta).
\]

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

\[
l_{K,R}^*(\theta; \{T^*_R\}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} D_j(k) T_{R,j}(m_{R}(k, \theta)).
\]

Note that

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

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.
5 Monte Carlo Studies

5.1 A Model of Schooling Choice

5.1.1 The Data Generating Process

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

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

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

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

where $E$ is the years of education taking values of 10, 12, 14, and 16 and $\varepsilon_1$ is normal, i.i.d., across individuals and periods with standard deviation of $\sigma_1$. Once an individual enters labor market and starts working, going back to school is not permitted. In each period $t$, the utility is given by $U_{1t}$ if the individual works, and $U_{2t}$ if he attends school. Also, we assume that the individual observes the labor income shock only after he enters the labor market and, therefore, the expected value of the wage only enters the utility function. This set-up yields the following two utilities corresponding to entering the labor market and attending school:

$$U_{1t} = E(w) = \exp \left( \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \frac{1}{2} \sigma_1^2 \right)$$

$$U_{2t} = \gamma_1 1\{\text{in high school}\} + \gamma_2 1\{\text{in two-year college}\} + \gamma_3 1\{\text{in four-year college}\} + \varepsilon_2,$$
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 $\gamma_1$ is the average utility of attending high school (we assume that there’s no tuition for attending high school), $\gamma_2$ the average utility of attending two year college including tuition cost, $\gamma_3$ the average utility of attending four year college including tuition cost, $\varepsilon_2$ is mean zero and normally distributed individual specific random effect on schooling utility which is independent across individuals, but is fixed over time for each individual. The standard deviation of $\varepsilon_2$ is denoted by $\sigma_2$.

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

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

where $\varepsilon_3$ is normally distributed with mean 0 and standard deviation $\sigma_3$ and does not change over time for each individual. The errors $\varepsilon_1, \varepsilon_2$, and $\varepsilon_3$ are independent.

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

$$U(E = 10) = \sum_{t=1}^{50} \beta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \times 10 + \frac{1}{2} \sigma_1^2}$$

$$U(E = 12) = \sum_{t=1}^{2} \beta^{t-1} (\gamma_1 + \varepsilon_2) + \sum_{t=3}^{50} \beta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \times 12 + \frac{1}{2} \sigma_1^2}$$

$$U(E = 14) = \sum_{t=1}^{2} \beta^{t-1} (\gamma_1 + \varepsilon_2) + \sum_{t=3}^{4} \beta^{t-1} (\gamma_2 + \varepsilon_2) + \sum_{t=5}^{50} \beta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 \times 14 + \frac{1}{2} \sigma_1^2}$$

$$U(E = 16) = \sum_{t=1}^{2} \beta^{t-1} (\gamma_1 + \varepsilon_2) + \sum_{t=3}^{6} \beta^{t-1} (\gamma_3 + \varepsilon_2) + \sum_{t=7}^{50} \beta^{t-1} e^{\alpha_0 + \alpha_1 X + \alpha_2 X_2 + \alpha_3 \times 16 + \frac{1}{2} \sigma_1^2}.$$

Given the model structure, we expect people with higher ability $X_1$ and $X_2$, higher discount factor $\beta$ and higher utility value of schooling $\varepsilon_2$ to attain a higher level of schooling.
We assume that the econometrician observes the ability measures \( X_1 \) and \( X_2 \), the schooling outcome, and characteristics \( X_3 \) that affect discount factor. Discount factor \( \beta \) and the utility value of schooling \( \varepsilon_2 \) are not observed. Given that there is no selection into the labor market, the parameters in the wage equation are simply estimated by a linear regression of wage on observable characteristics. Given the purpose of this simulation exercise, we assume that the parameters in the wage equation are known and focus only on the parameters in the schooling utility and the parameters in the discount factor. Hence the parameters of interest in this exercise is as follows:

- **Schooling utility parameters**: \( \gamma_1, \gamma_2, \gamma_3, \sigma_2 \)
- **Discount factor parameters**: \( \rho_0, \rho_1, \sigma_3 \)

For estimation, we consider our TSF-MLE, simulated MLE following Lerman and Manski (1981)’s proposal, McFadden (1989)’s smoothed SMLE, and the method of simulated moments of McFadden (1989). Our comparison is not exhaustive, but we believe that the simulated MLE of Lerman and Manski (1981) and the MSM of McFadden (1989) are the most common approach that an empirical researcher adopts in this type of models. Note that we cannot apply simulation methods that involve GHK simulators because the unobserved heterogeneity in discount factor is nonlinear in the latent process. The moment conditions that we consider for MSM are as follows:

\[
\sum_i (D_j - p_j(X_i; \theta)) \times 1 = 0, \ j = 2, 3, 4
\]

\[
\sum_i (D_j - p_j(X_i; \theta)) \times X_k = 0, \ j = 2, 3, 4, \ k = 1, 2, 3.
\]

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.

### 5.1.2 TSF-MLE, Lerman-Manski SMLE, and Smoothed SMLE

This section compares the performance of our estimator, the Lerman-Manski’s procedure, and smoothed SMLE. The Lerman-Manski procedure uses simulated frequencies to compute simulated choice probabilities. To prevent the zero-probability problem, we substituted \( 0.5/R \) for simulated probabilities that turned out to be zero. The second kind is a smoothed
Table 1: TSF-MLE and Lerman-Manski SMLEs: Log Likelihood

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

SMLE which is computed by using the following smoothed simulated choice probability:

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

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

Table 1 compares the overall simulation errors in terms of the log-likelihood evaluation of the simulation-based estimator using the true log-likelihood \( l_n(\theta) \). This number is bounded by \( l_n(\hat{\theta}_{MLE}) \) with \( \hat{\theta}_{MLE} \) denoting the MLE of \( \theta_0 \). As the number is higher, the simulation-based estimator suffers from a smaller overall simulation error. First, note that the performance
Table 2: TSF-MLE and Lerman-Manski SMLEs: MAE of Estimated Standard Deviation of Discount Factor \((\times 100), (\sigma_3 = 0.02, )\)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>(R = 10)</th>
<th>(R = 20)</th>
<th>(R = 50)</th>
<th>(R = 100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n = 100)</td>
<td>TSF-MLE</td>
<td>0.53</td>
<td>0.47</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.57</td>
<td>0.51</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.1))</td>
<td>0.66</td>
<td>0.54</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.01))</td>
<td>0.68</td>
<td>0.55</td>
<td>0.43</td>
<td>0.40</td>
</tr>
<tr>
<td>(n = 200)</td>
<td>TSF-MLE</td>
<td>0.39</td>
<td>0.35</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.44</td>
<td>0.36</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.1))</td>
<td>0.55</td>
<td>0.44</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.01))</td>
<td>0.52</td>
<td>0.42</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>(n = 500)</td>
<td>TSF-MLE</td>
<td>0.26</td>
<td>0.22</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.37</td>
<td>0.25</td>
<td>0.19</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.1))</td>
<td>0.55</td>
<td>0.47</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.01))</td>
<td>0.47</td>
<td>0.47</td>
<td>0.27</td>
<td>0.20</td>
</tr>
<tr>
<td>(n = 1000)</td>
<td>TSF-MLE</td>
<td>0.21</td>
<td>0.18</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.36</td>
<td>0.22</td>
<td>0.16</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.1))</td>
<td>0.52</td>
<td>0.54</td>
<td>0.32</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ((\lambda = 0.01))</td>
<td>0.49</td>
<td>0.53</td>
<td>0.31</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Table 3: TSF-MLE and Lerman-Manski SMLEs: MAE of the estimator of $\rho_1 = 0.02$ (×100).

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


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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td>TSF-MLE</td>
<td>968</td>
<td>856</td>
<td>740</td>
<td>667</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>989</td>
<td>875</td>
<td>747</td>
<td>655</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.1$)</td>
<td>1,196</td>
<td>897</td>
<td>746</td>
<td>659</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.01$)</td>
<td>1,173</td>
<td>877</td>
<td>728</td>
<td>641</td>
</tr>
<tr>
<td>$n = 200$</td>
<td>TSF-MLE</td>
<td>702</td>
<td>613</td>
<td>515</td>
<td>465</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>762</td>
<td>654</td>
<td>530</td>
<td>468</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.1$)</td>
<td>1,270</td>
<td>778</td>
<td>524</td>
<td>464</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.01$)</td>
<td>1,240</td>
<td>764</td>
<td>533</td>
<td>463</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>TSF-MLE</td>
<td>492</td>
<td>408</td>
<td>348</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>541</td>
<td>462</td>
<td>362</td>
<td>316</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.1$)</td>
<td>1,565</td>
<td>812</td>
<td>405</td>
<td>306</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.01$)</td>
<td>1,515</td>
<td>831</td>
<td>412</td>
<td>324</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>338</td>
<td>302</td>
<td>256</td>
<td>243</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>406</td>
<td>351</td>
<td>269</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.1$)</td>
<td>1,800</td>
<td>1,031</td>
<td>394</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>Smoothed SMLE ($\lambda = 0.01$)</td>
<td>1,782</td>
<td>1,008</td>
<td>397</td>
<td>243</td>
</tr>
</tbody>
</table>
of the Lerman-Manski is different from smoothed SMLEs. The simulation results show that the use of smoothing does not improve the performance, and sometimes, even worsen the quality of the estimator.

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

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

5.1.3 TSF-MLE and MSM

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

Table 5 again compares the performance of MSM and TSF-MLE in terms of the log-likelihood evaluation. First, our estimator performs better than MSM that does not use optimal weighting matrix. This is true for most ranges of simulation numbers and sample sizes considered. Outperformance by our estimator becomes conspicuous in particular when the sample size is 100 and simulation number is 100. This appears to reflect the fact that our estimator becomes more like an MLE as the simulation number becomes large while MSM does not. When an optimal weighting matrix was used, the quality of MSM substantially improves, and sometimes outperforms our estimator, especially when the sample size is large and the simulation number is small. This may be due to the slower rate of convergence of our estimator than MSM estimators. However, it should be noted that the computation of MSM using the optimal weighting matrix involves the first step estimation of the parameters.
Table 5: TSF-MLE and MSM: Log Likelihood

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

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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100$</td>
<td>TSF-MLE</td>
<td>0.53</td>
<td>0.47</td>
<td>0.42</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>0.88</td>
<td>0.87</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>0.70</td>
<td>0.68</td>
<td>0.66</td>
<td>0.63</td>
</tr>
<tr>
<td>$n = 200$</td>
<td>TSF-MLE</td>
<td>0.39</td>
<td>0.35</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>0.71</td>
<td>0.44</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>0.45</td>
<td>0.42</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>$n = 500$</td>
<td>TSF-MLE</td>
<td>0.26</td>
<td>0.22</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>0.55</td>
<td>0.47</td>
<td>0.28</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>0.47</td>
<td>0.47</td>
<td>0.27</td>
<td>0.23</td>
</tr>
</tbody>
</table>
Table 7: TSF-MLE and MSM: MAE of the estimator of $\rho_1 = 0.02$ ($\times 100$).

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

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

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>TSF-MLE</td>
<td>968</td>
<td>856</td>
<td>740</td>
<td>667</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>836</td>
<td>807</td>
<td>790</td>
<td>785</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>946</td>
<td>913</td>
<td>894</td>
<td>879</td>
</tr>
<tr>
<td>200</td>
<td>TSF-MLE</td>
<td>702</td>
<td>613</td>
<td>515</td>
<td>465</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>549</td>
<td>532</td>
<td>531</td>
<td>526</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>570</td>
<td>549</td>
<td>545</td>
<td>544</td>
</tr>
<tr>
<td>500</td>
<td>TSF-MLE</td>
<td>492</td>
<td>408</td>
<td>348</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>MSM w/o Optimal Weighting Matrix</td>
<td>325</td>
<td>321</td>
<td>322</td>
<td>318</td>
</tr>
<tr>
<td></td>
<td>MSM w/ Optimal Weighting Matrix</td>
<td>346</td>
<td>336</td>
<td>330</td>
<td>330</td>
</tr>
</tbody>
</table>
Therefore, the direct comparison of these two estimators does not appear to be fair as one can increase the simulation number of TSF-MLE taking advantage of its fast computing time. In our simulations, the Lerman-Manski SMLE estimators and MSM that does not use an optimal weighting matrix took approximately the same computation time as our estimator. The MSM estimator that uses an optimal weighting matrix took about twice as long.

6 Conclusion

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

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

7 Appendix: Proofs of the Results

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

Proof of Lemma 1: The proof is algebraically straightforward. Indeed, we can write out $\lambda_k(p, p_0; T_R)$ as in (15) and check the condition (12). However, we offer a different proof which reveals the discovery process of the transforms $\{T_R^j\}$. First, we let $\{T_R^j\}$ take the following form:

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

for some map $T_R$. Observe that

$$\Lambda_R(p, p_0; \{T_R^j\}) = \sum_{j=1}^{J} p_j p_0 \sum_{m \in \mathbb{N}_{R,j}} T_R^j(m) \left( \begin{array}{c} R \\ m_1, \ldots, m_J \end{array} \right) p_1^{m_1} \cdots p_J^{m_J}.$$
Note that the derivative of $\Lambda_R$ with respect to $p_k$ at $p = p_0$ is

$$
\lambda_k(p_0, p_0; \{T_R^j\}) = \frac{\partial}{\partial p_k} \Lambda_R(p, p_0; \{T_R^j\})|_{p = p_0} = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,j}} T_R^j(m) \binom{R}{m_1, \ldots, m_J} m_k p_{j0}^m p_{20}^2 \cdots p_{k0}^k \cdots p_{j0}^{m_j}.
$$

Let $c_k(m_1, \ldots, m_J)$ be the coefficient of $p_{1}^{m_1} \cdots p_{J}^{m_J}$ in the expansion of $\lambda_k(p, p_0; \{T_R^j\})$ as above. For brevity, put $p = p_0$ so that we write $\lambda_1(p) \equiv \lambda_1(p, p; \{T_R^j\})$ as

$$
\lambda_1(p) = \sum_{j=1}^{J} \sum_{m \in \mathbb{N}_{R,j}} T_R^j(m) \binom{R}{m_1, \ldots, m_J} m_1 p_{10}^{m_1-1} p_{20}^{m_2} \cdots p_{j-1}^{m_{j-1}} p_j^{m_j+1} p_{j+1}^{m_{j+1}} \cdots p_J^{m_J}.
$$

Let us compute $c_1(m_1, \ldots, m_J)$. Then it suffices to show that $c_j(m_1, \ldots, m_J)$ is the same for all $j = 1, \ldots, J$, or, without loss of generality, that $c_1(m_1, \ldots, m_J) = c_2(m_1, \ldots, m_J)$.

First observe that $c_2(m_1, m_2, \ldots, m_J) = c_1(m_2, m_1, \ldots, m_J)$ by the form of $\{T_R^j\}$ in (14) and $\Lambda_R$. Hence it suffices to show that $c_1(m_1, m_2, \ldots, m_J) = c_1(m_2, m_1, \ldots, m_J)$.

To show this, first note that

$$
c_1(m_1, \ldots, m_J) = T_R^1(m) \binom{R}{m_1, \ldots, m_J} m_1 + U_R
$$

where

$$
U_R = \sum_{j=2}^{J} T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_{j-1}, m_{j+1}, \ldots, m_J) \times \binom{R}{m_1 + 1, m_2, \ldots, m_{j-1}, m_j - 1, m_{j+1}, \ldots, m_J} (m_1 + 1).
$$

The relation in (16) holds for any $(m_1, \ldots, m_J) \in \mathbb{N}_{R,J}$ and we can simply extend the domain of $T_R$ to negative numbers by taking $T_R(m_j; m_{-j}) = 0$ if $m_j < 0$. By noting

$$
\binom{R}{m_1 + 1, m_2, \ldots, m_{j-1}, m_j - 1, m_{j+1}, \ldots, m_J} (m_1 + 1) = \binom{R}{m_1, \ldots, m_J} m_j,
$$

we write the coefficient $c_1(m_1, \ldots, m_J)$ in (16) as

$$
\binom{R}{m_1, \ldots, m_J} \left[ m_1 T_R(m_1; m_{-1}) + \sum_{j=2}^{J} m_j T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_{j-1}, m_{j+1}, \ldots, m_J) \right] .
$$
Since the factor in front of the above bracket does not depend on "1", it suffices to show that

\[ c_1(m_1, m_2, \ldots, m_J) \]
\[ = m_1T_R(m_1; m_2, \ldots, m_J) + \sum_{j \neq 1} m_jT_R(m_j - 1; m_1 + 1, m_2, \ldots, m_J) \]
\[ = m_2T_R(m_2; m_1, \ldots, m_J) + \sum_{j \neq 2} m_jT_R(m_j - 1; m_2 + 1, m_1, \ldots, m_J) \]
\[ = c_1(m_2, m_1, m_3, \ldots, m_J). \]

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

\[ m_1 \left[ T_R(m_1; m_2, \ldots, m_J) - T_R(m_1 - 1; m_2 + 1, m_3, \ldots, m_J) \right] \]
\[ + \sum_{j=3}^{J} m_j \left[ T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_J) - T_R(m_j - 1; m_2 + 1, m_1, \ldots, m_J) \right] \]
\[ = m_2 \left[ T_R(m_2; m_1, m_3, \ldots, m_J) - T_R(m_2 - 1; m_1 + 1, m_3, \ldots, m_J) \right]. \]

Therefore, the proof is complete once we show that the above equality is satisfied by our choice of (26). One can check this equality immediately by considering each case: \( m_1 = m_2 = 0 \) and \( m_1, m_2 > 0 \) and finally \( m_1 = 0, m_2 > 0 \). However, here we take a different route, showing how the form of (26) was discovered. In the proof we generate sufficient conditions for the equality in (17). Then these sufficient conditions lead to the solution of (26).

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 (17) is trivially satisfied.

**Case 1) **\( m_1, m_2 > 0 \). Then, the condition (17) is satisfied if

\[ m_1 \left[ T_R(m_1; m_2, \ldots, m_J) - T_R(m_1 - 1; m_2 + 1, m_3, \ldots, m_J) \right] = 1, \]

and

\[ T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_J) - T_R(m_j - 1; m_2 + 1, m_1, \ldots, m_J) = 0. \]

Restriction (19) implies that \( T_R(m_1; m_2, \ldots, m_J) \) depends on \((m_2, \ldots, m_J)\) only through \( \nu(m_2, \ldots, m_J) \), the number of non-zero elements from the non-choices \( \{m_2, \ldots, m_J\} \). To see this, choose \((m'_2, \ldots, m'_J)\) such that \( \nu(m'_2, \ldots, m'_J) = \nu(m_2, \ldots, m_J) \). Then, we can show that

\[ T_R(m_1; m_2, \ldots, m_J) = T_R(m_1; m'_2, \ldots, m'_J), \]

by repeating the process in (19) with adding and subtracting by 1 between two non-zero members from \( \{m_2, \ldots, m_J\} \).

Therefore we write

\[ T_R(m_1; m_2, \ldots, m_J) = T_R(m_1, \nu(m_2, \ldots, m_J)), \]

where \( \nu \) denotes the number of non-zero elements in the non-choice set. Using the observation in (19), (18) can be re-written as

\[ m_1 \left[ T_R(m_1; \nu(m_2, \ldots, m_J)) - T_R(m_1 - 1; \nu(m_2 + 1, m_3, \ldots, m_J)) \right] = 1, \]
and note that \( \nu(m_2, \ldots, m_J) = \nu(m_2 + 1, m_3, \ldots, m_J) \). Hence we extract one condition for \( T_R \) that leads to (20):

\[
T_R(m, \nu) - T_R(m - 1, \nu) = \frac{1}{m} \quad \text{for all possible } m. \tag{21}
\]

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

\[
m_1 [T_R(m_1; m_2, \ldots, m_J) - T_R(m_1 - 1; m_2 + 1, m_3, \ldots, m_J)]
= R[T_R(R; m_2 = 0, \ldots, m_J = 0) - T_R(R - 1; m_2 + 1, m_3 = 0, \ldots, m_J = 0)] = 0
\]

or

\[
T_R(R, 0) = T_R(R - 1, 1). \tag{23}
\]

If on the other hand \( m_3 > 0 \), we have from (12)

\[
m_1 [T_R(m_1; m_2, m_3, \ldots, m_J) - T_R(m_1 - 1; m_2 + 1, m_3, \ldots, m_J)]
+ \sum_{j=3}^{J} m_j [T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_J) - T_R(m_j - 1; m_2 + 1, m_1, \ldots, m_J)]
= 0.
\]

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

\[
m_1 [T_R(m_1; m_2, m_3, \ldots, m_J) - T_R(m_1 - 1; m_2, m_3 + 1, \ldots, m_J)]
= m_1 [T_R(m_1 - 1; m_2 + 1, m_3, \ldots, m_J) - T_R(m_1 - 1; m_2, m_3 + 1, \ldots, m_J)]
+ \sum_{j=3}^{J} m_j [T_R(m_j - 1; m_2 + 1, m_1, \ldots, m_J) - T_R(m_j - 1; m_1 + 1, m_2, \ldots, m_J)]
\]

Note that the left hand side in (24) is 1 by (20) and the difference in the number of non-zero elements in \( T_R \) for each difference term on the right-hand side is exactly 1. For example, \( \nu(m_2 + 1, m_3, \ldots, m_J) = \nu(m_2, m_3 + 1, \ldots, m_J) + 1 \) and \( \nu(m_2 + 1, m_1, \ldots, m_J) = \nu(m_1 + 1, m_2, \ldots, m_J) + 1 \). Therefore, if

\[
T_R(m, \nu) - T_R(m, \nu - 1) = c
\]

for some \( c \) independent of \( m \) and \( \nu \), (24) is satisfied. In this case, (24) becomes

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

Therefore, we extract a condition for (24):

\[
T_R(m, \nu) - T_R(m, \nu - 1) = \frac{1}{R} \tag{25}
\]

for all \( m \) and \( \nu \). To summarize, conditions (21), (23), and (25) are sufficient for (17).

Now, if we define

\[
T_R(m_j, m_{-j}) = - \sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\nu(m_{-j})}{R}, \tag{26}
\]
this choice of $T_R$ satisfies conditions (21), (23), and (25), and hence the equation (17) follows, completing
the proof. On the other hand, it is also worth noting that the conditions (21), (23), and (25) for $T_R$ also
lead to the form of (26) up to an affine transform. This is the way the transform $T_R$ is determined. ■

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

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

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

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

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

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

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

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

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

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

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

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

$$B_R (m) = p_{10} \left( \frac{R}{m_1, m_2, m_3} \right) \left[ 1 - \frac{m_3}{m_1} \right] + p_{20} \left( \frac{R - 1}{m_1 - 1, m_2, m_3} \right) \left[ \frac{1}{m_1} - \frac{1}{m_3} \right]
+ p_{30} \left( \frac{R}{m_1 - 1, m_2, m_3 + 1} \right) \left[ -1 + \frac{m_3 + 1}{m_1} \right] .$$

32
Now, write the summand in $\lambda_1 - \lambda_3$:

$$B_R(m)p_{1}^{m_{1}-1}p_{2}^{m_{2}}p_{3}^{m_{3}} = \left\{ \begin{array}{l} p_{10} \left( \begin{array}{l} R \\ m_{1}, m_{2}, m_{3} \end{array} \right) p_{1}^{m_{1}}p_{2}^{m_{2}}p_{3}^{m_{3}} - p_{10} \left( \begin{array}{l} R - 1 \\ m_{1} - 1, m_{2}, 0 \end{array} \right) p_{1}^{m_{1}-1}p_{2}^{m_{2}}I(m_1 > 0) \\ + p_{20} \left( \begin{array}{l} R - 1 \\ 0, m_{2}, m_{3} \end{array} \right) p_{2}^{m_{2}}p_{3}^{m_{3}} - p_{20} \left( \begin{array}{l} R - 1 \\ m_{1} - 1, m_{2}, 0 \end{array} \right) p_{1}^{m_{1}-1}p_{2}^{m_{2}}I(m_1 > 0) \\ - \frac{p_{30}}{p_{3}} \left( \begin{array}{l} R \\ m_{1} - 1, m_{2}, m_{3} + 1 \end{array} \right) p_{1}^{m_{1}-1}p_{2}^{m_{2}}p_{3}^{m_{3}+1}I(m_1 > 0) + p_{30} \left( \begin{array}{l} R - 1 \\ 0, m_{2}, m_{3} \end{array} \right) p_{2}^{m_{2}}p_{3}^{m_{3}}. \end{array} \right\}$$

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

$$\frac{p_{10}}{p_{1}} \left[ 1 - \sum_{m \in \mathbb{N}_{R,3}} \left( \begin{array}{l} R \\ m_{1}, m_{2}, m_{3} \end{array} \right) p_{2}^{m_{2}}p_{3}^{m_{3}} \right] = - \left( p_{10} (p_{1} + p_{2})^{R-1} + p_{20} (p_{2} + p_{3})^{R-1} - p_{20} (p_{1} + p_{2})^{R-1} \right)$$

or

$$\frac{p_{30}}{p_{3}} \left[ 1 - \sum_{m \in \mathbb{N}_{R,3}} \left( \begin{array}{l} R \\ m_{1} - 1, m_{2}, 0 \end{array} \right) p_{1}^{m_{1}-1}p_{2}^{m_{2}} \right] + p_{30} (p_{2} + p_{3})^{R-1}$$

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,

$$\frac{p_{10}}{p_{1}} \left[ 1 - (1 - p_{1})^{R} \right] + (1 - p_{10}) (1 - p_{1})^{R-1} - \frac{p_{30}}{p_{3}} \left[ 1 - (1 - p_{3})^{R} \right] = \frac{p_{10}}{p_{1}} \left[ (1 - p_{10}) (1 - p_{1})^{R-1} - \frac{p_{30}}{p_{3}} (1 - p_{3})^{R-1} \right].$$

Therefore, $\partial (\lambda_1 - \lambda_3) / \partial p_{1}$ is equal to

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

and by symmetry, $\partial (\lambda_3 - \lambda_1) / \partial p_{3}$ is equal to

$$\lambda_{33} - \lambda_{31} = - \frac{p_{30}}{p_{3}} + \frac{p_{30}}{p_{3}} (1 - p_{3})^{R-1} - \left( 1 - \frac{p_{30}}{p_{3}} \right) (R - 1) (1 - p_{3})^{R-2}.$$

We also obtain that $\partial (\lambda_1 - \lambda_3) / \partial p_{2} = \lambda_{12} - \lambda_{32} = 0$. Likewise, from

$$\lambda_{2} - \lambda_{3} = \frac{p_{20}}{p_{2}} - \left( 1 - \frac{p_{20}}{p_{2}} \right)(1 - p_{2})^{R-1} - \frac{p_{30}}{p_{3}} - \left( 1 - \frac{p_{30}}{p_{3}} \right)(1 - p_{3})^{R-1},$$
we obtain

\[ \lambda_{22} - \lambda_{32} = -\frac{p_{20}}{p_2^2} + \frac{p_{20}}{p_2} (1 - p_2)^{R-1} - \left(1 - \frac{p_{20}}{p_2}\right) (R - 1) (1 - p_2)^{R-2}, \]

\[ \lambda_{33} - \lambda_{32} = -\frac{p_{30}}{p_3^2} + \frac{p_{30}}{p_3} (1 - p_3)^{R-1} - \left(1 - \frac{p_{30}}{p_3}\right) (R - 1) (1 - p_3)^{R-2}, \]

\[ \lambda_{21} - \lambda_{31} = 0. \]

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

\[
\begin{pmatrix}
\lambda_{11} - \lambda_{31} - (\lambda_{13} - \lambda_{33}) & \lambda_{21} - \lambda_{31} - (\lambda_{23} - \lambda_{33}) \\
\lambda_{12} - \lambda_{32} - (\lambda_{13} - \lambda_{33}) & \lambda_{22} - \lambda_{32} - (\lambda_{23} - \lambda_{33})
\end{pmatrix}
\]

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

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

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

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

First, when \( p_{10} = 1 \),

\[ \lambda_{11} - \lambda_{31} = -\frac{1}{p_1^2} + \frac{1}{p_1^2} (1 - p_1)^{R-1} + \frac{1}{p_1} (R - 1) (1 - p_1)^{R-1} \]

\[ = -\frac{1}{p_1^2} + \frac{1}{p_1} [1 + p_1 (R - 1)] (1 - p_1)^{R-1}. \]

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

\[ \lambda_{11} - \lambda_{31} \leq -\frac{1}{p_1^2} + \frac{1}{p_1^2} (1 + p_1)^{R-1} (1 - p_1)^{R-1} \]

\[ \leq -\frac{1}{p_1^2} + \frac{1}{p_1^2} (1 - p_1^2)^{R-1} < 0. \]

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

Consider the case \( J > 3 \). First, we get

\[ \lambda_j - \lambda_k = \frac{p_{j0}}{p_{j}} + \left(1 - \frac{p_{j0}}{p_{j}}\right) (1 - p_j)^{R-1} - \frac{p_{k0}}{p_{k}} - \left(1 - \frac{p_{k0}}{p_{k}}\right) (1 - p_k)^{R-1}. \]
for all $j, k = 1, 2, \ldots, J$. Then it suffices to check the negative definiteness of the matrix

\[
\begin{pmatrix}
\lambda_{11} - \lambda_{j1} & 0 & \cdots & 0 \\
0 & \lambda_{22} - \lambda_{j2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{J-1,j-1} - \lambda_{j,j-1}
\end{pmatrix} - (\lambda_{11} - \lambda_{j1}) \begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1
\end{pmatrix}.
\]

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 $j = 1, 2, \ldots, J - 1$. This can be proved exactly in the same way as before. ■

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

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

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

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

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

\[
E \left[ \sup_{\theta \in B(\tilde{\theta}, \varepsilon)} \left| g_{jk}(\theta) - g_{jk}(\tilde{\theta}) \right|^2 |X_i \right] \leq P \left\{ \mu^D_{ijk}(\tilde{\theta}) - h_{ijk}(\varepsilon) \leq \eta_{ik} - \eta_{ij} \leq \mu^D_{ijk}(\theta) + h_{ijk}(\varepsilon) |X_i \right\}
\]

\[
= F_{ikj}(\mu^D_{ijk}(\tilde{\theta}) + h_{ijk}(\varepsilon), X_i) - F_{ikj}(\mu^D_{ijk}(\tilde{\theta}) - h_{ijk}(\varepsilon), X_i)
\]

\[
\leq C \left( \sup_{x \mu} f_{ikj}(\mu|x) \right) \sup_{\theta \in B(\tilde{\theta}, \varepsilon)} ||\theta - \tilde{\theta}|| \leq C \varepsilon,
\]

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

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

\[
\sup_{\theta \in \Theta} \left| l^*_n, R(\theta) - l_R(\theta) \right| \rightarrow_p 0 \text{ as } n \rightarrow \infty,
\]

where $l_R(\theta) = E l^*_n, R(\theta)$. Since $\Theta$ is compact and for each $\theta \in \Theta$ we have

\[
l^*_n, R(\theta) \rightarrow_p l_R(\theta),
\]

by the Law of Large Numbers, it suffices to show the following stochastic equicontinuity condition, i.e., for
any $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that for each $\tilde{\theta} \in \Theta$,

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

Define $T_R$ as in (26). Recall that we can write $l_{n,R}(\tilde{\theta})$ as

$$
l_{n,R}(\tilde{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_R(m_{ij}^*(\theta), m_{-ij}^*(\theta))
+ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \nu(m_{-ij}^*(\theta)).
$$

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

$$
P \left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| l_{n,R}^*(\theta) - l_{n,R}^*(\tilde{\theta}) \right| > \eta \right\}
\leq J P \left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \sup_{1 \leq j \leq J} \left| m_{j,R}(X_i, \eta_i^*; \theta) - m_{j,R}(X_i, \eta_i^*; \tilde{\theta}) \right| \geq \eta/(CJ) \right\}.
$$

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

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

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

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

$$
|E l_{n,R}^*(\tilde{\theta}) - E l_{n,R}^*(\theta_0)| \leq C ||\theta - \theta_0||^2,
$$

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

$$
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_j(X_i, \eta_i; \theta_0) h_R(p_{j,R}(X_i, \theta), \nu(m_{-ij}^*(\theta))).
$$
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,

$$E \left[ \sup_{\theta : ||\theta - \theta_0|| \leq \delta} \left| h_R(p^*_R(X_i, \theta), \nu(m^*_{-ij}(\theta))) - h_R(p^*_R(X_i, \theta_0), \nu(m^*_{-ij}(\theta_0))) \right| \right]^2$$

(28)

$$\leq \text{C}{\text{R}} \mathbb{P} \left\{ \sup_{\theta : ||\theta - \theta_0|| \leq \delta} \sup_{1 \leq r \leq R} |\delta_j(X_i, \eta^*_r; \theta) - \delta_j(X_i, \eta^*_r; \theta_0)| \geq 1 \right\}$$

$$\leq \text{C}{\text{R}} E \left[ \sup_{\theta : ||\theta - \theta_0|| \leq \delta} \sup_{1 \leq r \leq R} \left| \delta_j(X_i, \eta^*_r; \theta) - \delta_j(X_i, \eta^*_r; \theta_0) \right|^2 \right] \leq C\delta,$$

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 $G = \{\gamma(\cdot, \cdot, \cdot; \theta) : \theta \in \Theta\}$. From the proof of Theorem 3.1 in Chen, Linton, and van Keilegom (2003), the result of (28) gives us

$$\int_0^1 \sqrt{1 + \log N_{||\cdot||}(G_{\delta}) ||G||_2} \mathbb{P}(d\delta) \leq \int_0^1 \sqrt{1 + \log N_{||\cdot||}(G_{\delta}) ||G||_2} \mathbb{P}(d\delta) < \infty,$$

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

$$E \left[ \sup_{\theta \in B(\theta_0, \delta)} \sqrt{\frac{n}{2}} |l^*_n(\theta) - l^*_n(\theta_0) - \mathbb{E}l^*_n(\theta) + \mathbb{E}l^*_n(\theta_0)| \right]$$

$$\leq C \int_0^1 \sqrt{1 + \log N_{||\cdot||}(G_{\delta}) ||G||_2} \mathbb{P}(d\delta) ||G_{\delta}||_2 \leq C||G_{\delta}||_2,$$

where $G_{\delta}$ indicates the envelope of $G_{\delta}$. The second inequality follows from the fact that

$$N_{||\cdot||}(G_{\delta}) ||G||_2 \leq N_{||\cdot||}(G_{\delta}) ||G - G_{\delta}||_2 \leq C||G_{\delta}||_2.$$

By the result of (28), we can take $G_{\delta}$ such that $||G_{\delta}||_2 \leq C\delta^{1/2}$, and deduce that the continuity modulus of $l^*_n(\theta)$ in $\theta$ turns out to be $O(\delta^{1/2})$. Now, following Kim and Pollard (1990) (e.g. see van der Vaart and Wellner (1996), p.323.), the rate of convergence $r_n$ for $\hat{\theta}$ satisfies $r_n^{2-1/2} \leq \sqrt{n}$. Hence $r_n \sim n^{1/3}$, yielding the result of the theorem.

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

$$P \left\{ \sup_{\theta \in B(\theta_0, \delta)} |p^*_ij(\theta) - p_i(\theta)| > \varepsilon_p/2 \big| X_i = x \right\}$$

(29)

$$\leq \text{2} \varepsilon_p E \left[ \sup_{\theta \in B(\theta_0, \delta)} |p^*_ij(\theta) - p_i(\theta)| \big| X_i = x \right]$$

$$\leq \text{2} \varepsilon_p \sqrt{\frac{C}{R}} \int_0^1 \sqrt{1 + \log N_{||\cdot||}(\varepsilon/C)^2, \mathbb{P}(d\delta) \mathbb{P}(d\delta)}$$

$$= O(R^{-1/2}).$$

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

$$P \left\{ \inf_{\theta \in B(\theta_0, \delta)} p^*_ij(\theta) > \varepsilon_p/2 \big| X_i \right\}$$

$$\geq P \left\{ \inf_{\theta \in B(\theta_0, \delta)} p_i(\theta) > \varepsilon_p/2 + \sup_{\theta \in B(\theta_0, \delta)} |p^*_ij(\theta) - p_i(\theta)| \big| X_i \right\}$$

$$\geq P \left\{ \varepsilon_p/2 > \sup_{\theta \in B(\theta_0, \delta)} |p^*_ij(\theta) - p_i(\theta)| \big| X_i \right\} \rightarrow 1$$

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

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

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

where

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

Combining the results, we can establish the \( \sqrt{n} \)-rate of convergence as we demonstrate now. Suppose we have shown Claims 1-2. Take a sequence \( r_n = n^{1/2} \) and partition \( \Theta \) 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

\[
P\left\{ r_n ||\hat{\theta} - \theta_0|| > 2^M \right\} \leq \sum_{j \geq M} \sum_{2^j \leq r_n} P\left\{ \sup_{\theta \in R_{j,n}} l_n, R(\theta) - l_n, R(\theta_0) \geq \eta \right\} + P\left\{ 2||\hat{\theta} - \theta_0|| \geq \eta \right\}.
\]

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

\[
l_n, R(\theta) - l_n, R(\theta_0) \leq \frac{C2^{2j-2} \log(R)}{r_n^2}
\]

by Claim 1. By using Claim 2, the sum of probabilities on the right-hand side in (30) is bounded by

\[
\sum_{j \geq M} \sum_{2^j \leq r_n} P\left\{ \sup_{\theta \in R_{j,n}} l_n, R(\theta) - l_n, R(\theta_0) \leq E l_n, R(\theta) + E l_n, R(\theta_0) \geq \frac{C2^{2j-2} \log(R)}{r_n^2} \right\}
\]

\[
\leq C \sum_{j \geq M} \frac{C \sqrt{n}}{2^{2j-2} \log(R)} \mu_R \left( \frac{2^{j-1}}{\sqrt{n}} \right)
\]

\[
\leq C n^{1/4} \sqrt{\frac{\log n}{\log R}} \sum_{j \geq M} 2^{-3j/2} + \frac{C}{\sqrt{\log R}} \sum_{j \geq M} 2^{-j+1} + \frac{C \sqrt{n}}{2^{2j-2} \log(R)} \sum_{j \geq M} 2^{-2j+2} \rightarrow 0,
\]

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

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

\[
T^*_ij(\theta) = T_{R}(m^*_ij(\theta), m^*_ij(\theta)),
\]

\[
\Delta ij(\theta) = T^*_ij(\theta) - T^*_ij(\theta_0), \text{ and } V(X_i) = \sum_{j=1}^{J} D_{ij} \frac{\partial}{\partial \theta} p_j(X_i, \theta_0) / p_j(X_i, \theta_0).
\]

We first show the following.
Claim 3:

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} [D_{ij} \Delta_{ij}(\theta) - E(D_{ij} \Delta_{ij}(\theta))] \]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ \frac{D_{ij}(p_j(X_i, \theta) - p_j(X_i, \theta_0))}{p_j(X_i, \theta_0)} - E \left[ \frac{D_{ij}(p_j(X_i, \theta) - p_j(X_i, \theta_0))}{p_j(X_i, \theta_0)} \right] \right\} + o_P(1),
\]

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

Now, note that by the mean-value theorem,

\[
\sum_{j=1}^{J} E(D_{ij} \Delta_{ij}(\theta)) = \sum_{j=1}^{J} \psi_j(\theta_0)'(\theta - \theta_0) + \sum_{j=1}^{J} (\theta - \theta_0) \Omega_j(\theta_0)'(\theta - \theta_0)
\]

where \( \theta_* \) lies on the line segment between \( \theta \) and \( \theta_0 \) and \( \psi_j(\theta) = \frac{\partial}{\partial \theta} E(D_{ij} \Delta_{ij}(\theta)) \) and \( \Omega_j(\theta) = \frac{\partial^2}{\partial \theta \partial \theta} E(D_{ij} \Delta_{ij}(\theta)) \).

However,

\[
\sum_{j=1}^{J} \psi_j(\theta_0) = \frac{\partial}{\partial \theta} \sum_{j=1}^{J} E(D_{ij} \Delta_{ij}(\theta)) \bigg|_{\theta=-\theta_0} = \frac{\partial}{\partial \theta} \sum_{j=1}^{J} E(D_{ij}(T_{ij}(\theta) - T_{ij}(\theta_0))) \bigg|_{\theta=-\theta_0}
\]

\[
= \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \left( \Lambda_R(p_{ij}(\theta), p_{ij}(\theta_0)) - \Lambda_R(p_{ij}(\theta_0), p_{ij}(\theta_0)) \right) \bigg|_{\theta=-\theta_0} = 0
\]

by the identification result in Theorem 1. Hence we have

\[
\sum_{j=1}^{J} E(D_{ij} \Delta_{ij}(\theta)) = \sum_{j=1}^{J} (\theta - \theta_0) \Omega_j(\theta_0)'(\theta - \theta_0).
\]

Therefore, using the fact that \( E V(X_i) = 0 \),

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

uniformly over \( \theta \in B(\theta_0, M n^{-1/2}) \), where \( Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(X_i) \). Now, we follow similar steps in the proof of Theorem 3.2.16 in van der Vaart and Wellner (1996). Combined with Claim 3, the result of (32) yields

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta_0)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \Delta_{ij}(\hat{\theta}) = \frac{1}{2} (\hat{\theta} - \theta_0)'\Omega(\hat{\theta} - \theta_0) + \frac{1}{\sqrt{n}} (\hat{\theta} - \theta_0)'Z_n + o_P(n^{-1}).
\]
Similarly

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^{*}(\theta_{0} - n^{-1/2} \Omega^{-1} Z_{n}) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^{*}(\theta_{0}) \tag{34}
\]

\[
= - \frac{1}{2n} Z_{n}^{\prime} \Omega^{-1} Z_{n} + o_{p}(n^{-1}).
\]

By the definition of \( \hat{\theta} \), the left-hand side of (33) is larger than the left-hand side of (34). We subtract the second equation from the first equation to obtain

\[
\frac{1}{2}(\hat{\theta} - \theta_{0} + n^{-1/2} \Omega^{-1} Z_{n}) \Omega(\hat{\theta} - \theta_{0} + n^{-1/2} \Omega^{-1} Z_{n}) \geq -o_{p}(n^{-1}).
\]

Since \( \Omega \) is negative definite, we conclude that

\[
\sqrt{n}(\hat{\theta} - \theta_{0}) = \Omega^{-1} Z_{n} + o_{p}(1).
\]

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

**Proof of Claim 1:** First, observe that \( l_{R}(\theta) - l_{R}(\theta_{0}) \) is less than or equal to

\[
\frac{1}{2} \left[ \sup_{\theta \in B(\theta_{0}, \delta)} (\theta - \theta_{0})^{\prime} \frac{\partial^{2} E}{\partial \theta \partial \theta'} \left[ \sum_{j=1}^{J} D_{ij} T_{ij}(\theta) \right] (\theta - \theta_{0}) \right].
\]

Note that

\[
\frac{\partial^{2} E[D_{ij} T_{ij}(\theta)]}{\partial \theta \partial \theta'} = -\frac{1}{R} \sum_{m=0}^{R-1} \frac{\partial^{2} E}{\partial \theta \partial \theta'} \left( D_{ij} 1 \{ R - m > \sum_{m=0}^{R-1} \delta_{j}(X_{i}, \eta_{i,m}; \theta) \} \right) \tag{35}
\]

We consider the first term. We can write it as

\[
-\frac{1}{R} \sum_{m=0}^{R-1} \frac{\partial^{2} E}{\partial \theta \partial \theta'} \left( D_{ij} 1 \{ R - m > \sum_{m=0}^{R-1} \delta_{j}(X_{i}, \eta_{i,m}; \theta) \} \right) \tag{35}
\]

\[
= -\frac{p_{j}(X_{i}, \theta_{0})}{R} \sum_{m=0}^{R-1} F_{R,p_{j}(X_{i}, \theta)}^{(1)}(R - m - 1) \frac{\partial^{2} p_{j}(X_{i}, \theta)}{\partial \theta \partial \theta'}
\]

\[
+ \frac{p_{j}(X_{i}, \theta_{0})}{R} \sum_{m=0}^{R-1} F_{R,p_{j}(X_{i}, \theta)}^{(2)}(R - m - 1) \frac{\partial p_{j}(X_{i}, \theta)}{\partial \theta} \frac{\partial p_{j}(X_{i}, \theta)}{\partial \theta'}
\]

where \( F_{R,p_{j}(X_{i}, \theta)}^{(1)}(\cdot) \) and \( F_{R,p_{j}(X_{i}, \theta)}^{(2)}(\cdot) \) are the first order and the second order derivatives of the binomial distribution function with parameter \( (R, p_{j}(X_{i}, \theta)) \). By Assumption 1(ii), we have \( p(X_{i}, \theta) \in B(p(X_{i}, \theta_{0}), C(X_{i}) \delta) \) for all \( \theta \in B(\theta_{0}, \delta) \) where \( C(X_{i}) \) is square integrable and does not depend on \( \theta \in B(\theta_{0}, \delta) \). By taking \( \delta \) small, we have eventually \( 1 \geq p_{j}(X_{i}, \theta) > \varepsilon > 0 \) for the constant \( \varepsilon_{p} \) in Assumption 2(iii). For this \( p(X_{i}, \theta) \), the derivatives \( F_{R,p_{j}(X_{i}, \theta)}^{(1)}(\cdot) \) and \( F_{R,p_{j}(X_{i}, \theta)}^{(2)}(\cdot) \) are bounded uniformly over \( \theta \in B(\theta_{0}, \delta) \) with
As argued before, we can take large probability. We can bound the Euclidean norm of the first term in (35) by
\[
\sup_{\theta \in B(\theta_0, \delta)} \sqrt{\mathbb{E} \left[ \left\| \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta'} \right\|^2 \right]} + C \times \frac{1}{R} \sum_{m=0}^{R-1} \frac{1}{1 - m/R}.
\]
And note that
\[
\frac{1}{R} \sum_{m=0}^{R-1} \frac{1}{1 - m/R} \leq C \int_0^{1-1/R} \frac{1}{1 - u} \, du + O(R^{-1}) = \log(R) + O(R^{-1}).
\]
Hence the first term on the right-hand side of (35) is \(O(\log(R))\).

Now we consider the second term in (35). Note that
\[
\frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{E} [\nu(m_{-ij}^*(\theta))] = \sum_{k=1, k \neq j}^J \frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{E} \left[ \mathbb{P} \left\{ \sum_{m=1}^{R} \delta_k(X_i, \eta_{i,m}; \theta) > 0 \right\} \mathbb{P} \left\{ \sum_{m=1}^{R} \delta_k(X_i, \eta_{i,m}; \theta) > 0 \right\} \right]
\]
\[
= \sum_{k=1, k \neq j}^J \frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{E} \left[ \left\{ 1 - F_{R,p_k(X_i, \theta)}(0) \right\} \right]
\]
\[
= \sum_{k=1, k \neq j}^J \mathbb{E} \left[ - F_{R,p_k(X_i, \theta)}^{(2)}(0) \frac{\partial p_j(X_i, \theta)}{\partial \theta} \frac{\partial p_j(X_i, \theta)}{\partial \theta} - F_{R,p_k(X_i, \theta)}^{(1)}(0) \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta'} \right].
\]
As argued before, we can take \(\delta\) small so that for all \(\theta \in B(\theta_0, \delta), 1 > p(X_i, \theta) > \varepsilon > 0\) for some \(\varepsilon\). And this leads to the fact that \(F_{R,p_j(X_i, \theta)}^{(s)}(0)\) and \(F_{R,p_j(X_i, \theta)}^{(s)}(0), s = 1, 2,\) are bounded uniformly over \(\theta \in B(\theta_0, \delta)\).
Therefore, the Euclidean norm of the second term in (35) is again bounded by
\[
\sup_{\theta \in B(\theta_0, \delta)} \sqrt{\mathbb{E} \left[ \left\| \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta'} \right\|^2 \right]} + C \times \frac{1}{R}.
\]
Hence we conclude \(\sup_{\theta \in B(\theta_0, \delta)} [l_R(\theta) - l_R(\theta_0)] \leq C \log(R) \delta^2\).

**Proof of Claim 2**: First, observe that for \(p, p_0 \in (\varepsilon_p/2, 1),\)
\[
\frac{1}{R} \sum_{m=0}^{R-1} \frac{1\{1 - m/R > p\} - 1\{1 - m/R > p_0\}}{1 - (m/R)} \leq \int_{1-p_0}^{1-p} \frac{1}{1 - u} \, du + O \left(R^{-1}\right)
\]
\[
= \log(p) - \log(p_0) + O \left(R^{-1}\right) = \frac{p - p_0}{p_0} + O(|p - p_0|^2 + R^{-1}).
\]
Let \(\hat{T}_{ij}^*(\theta) = \log(p_{ij}^*(\theta)) - \log(p_{ij}^*(\theta_0)).\) Since \(\inf_{\theta \in B(\theta_0, \delta)} p_{ij}^*(\theta) > \varepsilon_p/2,\) we deduce that
\[
\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^J \left( D_{ij} T_{ij}^*(\theta) - D_{ij} T_{ij}^*(\theta_0) \right) - \mathbb{E} \left[ D_{ij} T_{ij}^*(\theta) - D_{ij} T_{ij}^*(\theta_0) \right] \right]
\]
\[
= \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^J \left( D_{ij} \hat{T}_{ij}^*(\theta) - D_{ij} \hat{T}_{ij}^*(\theta_0) \right) - \mathbb{E} \left[ D_{ij} \hat{T}_{ij}^*(\theta) - D_{ij} \hat{T}_{ij}^*(\theta_0) \right] \right] + O(\sqrt{n} R^{-1}).
\]
We focus on the last expectation. Observe that

\[
\sqrt{\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| D_{ij} \tilde{T}_{ij}(\theta) - D_{ij} \tilde{T}_{ij}(\theta_0) \right|^2 \right]} \leq C \sqrt{\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_{ij}(\theta) - p^*_{ij}(\theta_0) \right|^2 \right]}
\]

\[
\leq C \sqrt{\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_{ij}(\theta) - p_{ij}(\theta) - \{p^*_{ij}(\theta_0) - p_{ij}(\theta_0)\} \right|^2 \right]}
\]

\[
+ C \sqrt{\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p_{ij}(\theta) - p_{ij}(\theta_0) \right|^2 \right]}.
\]

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

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

As for the leading expectation above, we proceed similarly as in (29):

\[
\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_{ij}(\theta) - p_{ij}(\theta) - \{p^*_{ij}(\theta_0) - p_{ij}(\theta_0)\} \right| | X_i = x \right]
\]

\[
\leq \frac{C}{\sqrt{R}} \int_0^{C \delta^{1/2}} \sqrt{1 + \log N_\varepsilon((\varepsilon/C)^2, \Theta, \| \cdot \|)} \, d\varepsilon \leq C \delta^{1/2} \sqrt{-\log \delta} / \sqrt{R},
\]

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

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

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

\[
\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| \frac{1}{\sqrt{R}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left( D_{ij} \tilde{T}_{ij}(\theta) - D_{ij} \tilde{T}_{ij}(\theta_0) - \mathbb{E} \left[ D_{ij} \tilde{T}_{ij}(\theta) - D_{ij} \tilde{T}_{ij}(\theta_0) \right] \right) \right| \right]
\]

\[
\leq C \int_0^{C \delta^{1/2} \sqrt{-\log \delta} / \sqrt{R} + C \delta} \sqrt{1 - C \log \{\varepsilon \sqrt{R}/\sqrt{-\log \delta}\}} \, d\varepsilon
\]

\[
+ C \int_0^{C \delta^{1/2} \sqrt{-\log \delta} / \sqrt{R} + C \delta} \sqrt{1 - C \log \varepsilon} \, d\varepsilon
\]

\[
\leq C \left\{ \delta^{1/2} \sqrt{-\log \delta} / \sqrt{R} + \delta \right\} \times \left\{ \sqrt{\log(R)} + \sqrt{\log(-\log \delta)} \right\}.
\]

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

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

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

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

42
we write (using (37)),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ D_{ij} \Delta_{ij}(\theta) - \mathbb{E} \left( D_{ij} \Delta_{ij}(\theta) \right) \right]$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ D_{ij} \left[ p_{ij}(\theta) - p_{ij}(\theta_0) \right] / p_{ij}(\theta_0) \right\} - \mathbb{E} \left[ D_{ij} \left[ p_{ij}(\theta) - p_{ij}(\theta_0) \right] / p_{ij}(\theta_0) \right] \right\}$$

$$+ D_n(\theta) + O_P(\delta_n^2 (- \log \delta_n)) + O_P(\sqrt{n} R^{-1})$$

where

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

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

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

$$\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| g_R(Y_i, \theta) - g_R(Y_i, \theta_0) \right|^2 \right]$$

$$= \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \left( p_{ij}(\theta) - p_{ij}(\theta_0) \right) / \left( p_{ij}(\theta) p_{ij}(\theta_0) \right) \right|^2 \right]$$

$$\leq \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \left( p_{ij}(\theta) - p_{ij}(\theta_0) \right) (p_i (\cdot) - p_{ij}(\theta_0)) / p_{ij}(\theta) (p_{ij}(\theta) p_{ij}(\theta_0)) \right|^2 \right]$$

$$+ \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \left( p_{ij}(\theta) - p_{ij}(\theta_0) - (p_{ij}(\theta) - p_{ij}(\theta_0)) \right) p_{ij}(\theta) / p_{ij}(\theta) (p_{ij}(\theta) p_{ij}(\theta_0)) \right|^2 \right].$$

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

$$\left( \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{1}{R} \sum_{r=1}^{R} d_{ij}^*(\theta) - \mathbb{E} \left[ d_{ij}^*(\theta) \right] \right| \right| X_i \right)^2 \right)^{1/2}$$

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

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

$$N_{\| \cdot \|_2}(\varepsilon, G_R(\delta_n), \| \cdot \|_2) \leq N_{\| \cdot \|_2}(C \epsilon^{R^{1/2}} / \sqrt{- \log \delta_n}, \Theta, \| \cdot \|_2) \leq C(\epsilon^{-1} R^{-1/2} \sqrt{- \log \delta_n})^{2d}.$$

Now, we write

$$\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} | D_n(\theta) | \right] = \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ D_{ij} g_R(Y_i, \theta) - \mathbb{E} \left( D_{ij} g_R(Y_i, \theta) \right) \right] \right| \right]$$

$$\leq \int_0^{CR^{-1/2} \delta_n^{1/2} \sqrt{- \log \delta_n}} \sqrt{1 + \log N_{\| \cdot \|_2}(\varepsilon, G_R(\delta_n), \| \cdot \|_2)} d\varepsilon \to 0,$$
because \( R^{-1/2} \delta^{1/2} \sqrt{-\log \delta_n} \leq R^{-1/2} \delta_n^{1/2} \sqrt{\log(R)} \rightarrow 0 \). Therefore, sup_{\theta \in B(\theta_0, \delta_n)} |D_n(\theta)| = o_P(1). We conclude that

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

We obtain the wanted result. ■

References


