The Asymptotic Distribution of Estimators with Overlapping Simulation Draws *

Tim Armstrong[a] A. Ronald Gallant[b] Han Hong[c] Huiyu Li[d]

First draft: September 2011
Current draft: December 2017

Abstract

We study the asymptotic distribution of simulation estimators, where the same set of draws are used for all observations under general conditions that do not require the function used in the simulation to be smooth. We consider two cases: estimators that solve a system of equations involving simulated moments and estimators that maximize a simulated likelihood. Many simulation estimators used in empirical work involve both overlapping simulation draws and nondifferentiable moment functions. Developing sampling theorems under these two conditions provides an important complement to the existing results in the literature on the asymptotics of simulation estimators.

Keywords: U-Process, Simulation estimators.

JEL Classification: C12, C15, C22, C52.

*Corresponding author is A. Ronald Gallant, The Pennsylvania State University, Department of Economics, 613 Kern Graduate Building, University Park PA 16802-3306, USA, aronaldg@gmail.com. We thank Donald Andrews, Bernard Salanié, and Joe Romano in particular, as well as seminar participants for insightful comments. We also thank a previous editor and two referees for constructive suggestions, and acknowledge support by the National Science Foundation (SES 1024504) and SIEPR.

[a] Yale University
[b] The Pennsylvania State University
[c] Stanford University
[d] Federal Reserve Bank of San Francisco

1 Introduction

Simulation estimation is popular in economics and is developed by Lerman and Manski (1981), McFadden (1989), Laroque and Salanie (1989), (1993), Duffie and Singleton (1993), Smith (1993), Gourieroux, Monfort, and Renault (1993), Gourieroux and Monfort (1996) and Gallant and Tauchen (1996) among others. Pakes and Pollard (1989) provided a general asymptotic approach for generalized method of simulated moment estimators, and verified the conditions in the general theory when a fixed number of independent simulations are used for each of the independent observations. In practice, however, researchers sometimes use the same set of simulation draws for all the observations in the dataset. Recent insightful papers by Lee and Song (2015) and Freyberger (2015) developed asymptotic results for a class of simulated maximum likelihood-like estimators and simulated method of moment estimators.

Independent simulation draws are doubly indexed, i.e., ω_{ir}, so that there are $n \times R$ simulations in total, where n is the number of observations and R is the number of simulations for each observation. Overlapping simulation draws are singly indexed, i.e., ω_r, so that there are R simulations in total. The same R simulations are used for each observation. The properties of simulation based estimators using overlapping and independent simulation draws are studied by Lee (1992) and Lee (1995) under the conditions that the simulated moment conditions are smooth and continuously differentiable functions of the parameters. This is, however, a strong assumption that is likely to be violated by many simulation estimators used in practice. We extend the above results to nonsmooth moment functions using empirical process and U process theories developed in a sequence of papers by Pollard (1984), Nolan and Pollard (1987, 1988) and Neumeyer (2004). In particular, the main insight relies on verifying the high level conditions in Pakes and Pollard (1989), Chen, Linton, and Van Keilegom (2003) and Ichimura and Lee (2010) by combining the results in Neumeyer (2004) with results from the empirical process literature (e.g. Andrews (1994)).

Even for the simulated method of moment estimator, the classical results in Pakes and Pollard (1989) and McFadden (1989) are for independent simulation draws. However, their results only apply to a finite number of independent simulations for each observation, since

the proof depends crucially on the fact that a finite sum of functions with limited complexity also has limited complexity. It is a challenging question with unclear answer how their analysis can be extended to a larger number of simulation draws. With overlapping simulation draws, this difficulty is resolved by appealing to empirical U-process theory.

A main application of maximum simulated likelihood estimators is multinomial probit discrete choice models and its various panel data versions (Newey and McFadden (1994)). Whether or not using overlapping simulations improves computational efficiency depends on the specific model. Generating the random numbers is cheap in terms of computational time easy but computing the moment conditions or the likelihood function is typically costly. To equate the order of computational effort, we will adopt the notation of letting R denote either the *total* number of overlapping simulations or the number of independent simulations *for each observation*. For a given R, Lee (1995), pointed out that the leading terms of the asymptotic expansion are smaller with independent draws than with overlapping draws. This suggests that independent draws are more desirable and leads to smaller confidence intervals whenever it is feasible.

There are still two reasons to consider overlapping draws, especially for simulated maximum likelihood estimators, based on theoretical and computational feasibility. Despite the theoretical advantage of the method of simulated moments, the method of simulated maximum likelihood is still appealing in empirical research, partly because it minimizes a well defined distance between the model and the data even when the model is misspecified. The asymptotic theory with independent draws in this case is difficult and to our knowledge has not been fully worked out in the literature. In particular, Pakes and Pollard (1989) only provided an analysis for simulated GMM, but did not provide an analysis for simulated MLE, which can be in fact far more involved.[1] Only the very recent insightful paper by Lee and Song (2015) studies an unbiased approximation to the simulated maximum likelihood, which still differs from most empirical implementation of simulated maximum likelihood methods using nonsmooth crude frequency simulators. Smoothing typically requires the

[1]While Pakes and Pollard (1989) refers to the method of simulated scores in likelihood based models, obtaining unbiased simulators for the score function is not immediate and may require independent simulations of the implied instrument functions, as highlighted in Train (2003). We are not aware of a previous definition for setting up the score function with nonsmooth likelihood simulators either.

choice of kernel and bandwidth parameters and introduces biases. For example, the Stern (1992) decomposition simulator, while smooth and unbiased, requires repeated calculations of eigenvalues and is computationally prohibitive. Kristensen and Salanié (2013) discuss bias reduction techniques for simulation estimators in smooth models.

When computing the simulated likelihood function is very difficult, overlapping simulations can be used to trade off computational feasibility with statistical accuracy. Using independent draws requires that R increases faster than $\sqrt{n}$, where n is the sample size in order that the estimator has an asymptotic normal distribution. With overlapping draws, the estimator will be asymptotically normal as long as R increases to infinity. A caveat, of course, is that when R is much smaller than n, the asymptotic distribution would mostly represent the simulation noise rather than the the sampling error, which reflects the cost in statistical accuracy as a result of more feasible computation.

In summary, previous work for method of simulated moments (MSM) or maximum simulated likelihood (MSL) may be classified according to whether: (a) the simulation draws are independent or overlapping; (b) the number of simulation draws per observation is R fixed or grows with the sample size; (c) the simulator is smooth in the parameter. This paper considers (a) overlapping simulation draws; (b) $R \to \infty$; (c) the simulator is nonsmooth, a combination which has not been addressed in the literature. Note that for both overlapping and independent draws, the same simulations are generally used for evaluating different parameter values θ.

2 Simulated Moments and Simulated Likelihood

We begin by formally defining the method of simulated moments and maximum simulated likelihood using overlapping simulation draws. These methods are defined in Lee (1992) and Lee (1995) in the context of multinomial discrete choice models. We use a more general notation to allow for both continuous and discrete dependent variables. Let $z_i = (y_i, x_i)$ be i.i.d. random variables in the observed sample for $i = 1, \ldots, n$, where the y_i are the dependent variables and the x_i are the covariates or regressors. Let $\omega_r, r = 1, \ldots, R$ be a set of simulation draws. We will use P to denote the distribution of z_i, and Q to denote the distribution of ω_r while P_n and Q_R are the empirical distributions of $z_i, i = 1, \ldots, n$ and

$\omega_r, r = 1, \ldots, R$, respectively. We are concerned about estimating an unknown parameter $\theta \in \Theta \subset \mathbb{R}^k$. The method of moment results are developed both for completeness and for expositional transition to the simulated maximum likelihood section.

The method of moments estimator is based on a set of moment conditions $g(z_i, \theta) \in \mathbb{R}^d$ such that $g(\theta) \equiv Pg(\cdot, \theta)$ is zero if and only if $\theta = \theta_0$ where θ_0 is construed as the true parameter value. The number of moment condition d is at least as large as the number of parameters k. In the above $Pg(\cdot, \theta) = \int g(z_i, \theta)\, dP(z_i)$ denotes expectation with respect to the true distribution of z_i. In models where the moment $g(z_i, \theta)$ can not be analytically evaluated, it can often be approximated using simulations. Given the simulation draws $\omega_r, r = 1, \ldots, R$, let $q(z_i, \omega_r, \theta)$ be a function such that it is an unbiased estimator of $g(z_i, \theta)$ for all z_i:

$$Qq(z, \cdot, \theta) \equiv \int q(z, \cdot, \theta)\, dQ(\omega) = g(z, \theta).$$

Then the unknown moment condition $g(z, \theta)$ can be estimated by

$$\hat{g}(z, \theta) = Q_R q(z, \omega_r, \theta) \equiv \frac{1}{R} \sum_{r=1}^{R} q(z, \omega_r, \theta),$$

which in turn is used to form an estimate of the population moment condition $g(\theta)$:

$$\hat{g}(\theta) = P_n \hat{g}(\cdot, \theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{g}(z_i, \theta) = \frac{1}{nR} \sum_{i=1}^{n} \sum_{r=1}^{R} q(z_i, \omega_r, \theta).$$

In the above, both $z_1, \ldots, z_n$ and $\omega_1, \ldots, \omega_R$ are iid and they are independent of each other. The method of simulated moments (MSM) estimator with overlapping simulated draws is defined with the usual quadratic norm as in Pakes and Pollard (1989)

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \|\hat{g}(\theta)\|^2_{W_n} \quad \text{where} \quad \|x\|^2_W = x' W x,$$

where both W_n and W are d dimensional weighting matrixes such that $W_n \xrightarrow{p} W$.

In the maximum simulated likelihood method, we reinterpret $g(z_i; \theta)$ as the likelihood function of θ at the observation z_i, and $\hat{g}(z_i; \theta)$ as the simulated likelihood function which is an unbiased estimator of $g(z_i; \theta)$. The MSL estimator is usually defined as, for i.i.d data,

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, P_n \log \hat{g}(\cdot; \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{g}(z_i; \theta).$$

While $g(z_i;\theta)$ is typically a smooth function of z_i and θ, $\hat{g}(z_i;\theta)$ oftentimes is not. The likelihood function $g\left(z;\theta\right)$ can be either the density for continuous data, or the probability mass function for discrete data. It can also be either the joint likelihood of the data, or the conditional likelihood $g\left(z;\theta\right) = g\left(y|x;\theta\right)$ when $z = (y, x)$.

In the following we will develop conditions under which both MSM and MSL are consistent as both $n \to \infty$ and $R \to \infty$. Under the conditions given below, they both converge at the rate of $\sqrt{m}$, where $m = \min(n, R)$ to a limiting normal distribution. These results are developed separately for MSM and MSL. For MSL, the condition that $R/\sqrt{n} \to \infty$ is required for asymptotic normality with independent simulation draws, e.g. Laroque and Salanie (1989) and Train (2003). With overlapping draws, asymptotic normality holds as long as both R and n converge to infinity. If $R/n \to 0$, then the convergence rate becomes $\sqrt{R}$ instead of $\sqrt{n}$. A simulation estimator with overlapping simulations can also be viewed as a profiled two step estimator to invoke the high level conditions in Chen, Linton, and Van Keilegom (2003). The derivations in the remaining sections are tantamount to verifying these high level conditions. For maximum likelihood with independent simulations, the bias reduction condition $\sqrt{R}/n \to \infty$ is derived in Laroque and Salanie (1989), (1993) and Gourieroux and Monfort (1996). For nonsmooth maximum likelihood like estimators, Lee and Song (2015) require the number of simulations to satisfy $\sqrt{R}\log R/n \to \infty$.

To summarize, the following assumption is maintained through the paper.

ASSUMPTION 1 Let $z_i = (y_i, x_i), i = 1, \ldots, n$ and $\omega_r, r = 1, \ldots, R$ be two independent sequences of i.i.d random variables with distributions P and Q respectively. The function $q(z_i, \omega_r, \theta)$ satisfies $Qq(z, \cdot, \theta) \equiv \int q(z, \omega, \theta)\, dQ(\omega) = g(z, \theta)$ for all z and all $\theta \in \Theta$.

3 Asymptotics of MSM with Overlapping Simulations

The MSM objective function takes the form of a two-sample U-process studied extensively in Neumeyer (2004):

$$\hat{g}(\theta) \equiv \frac{1}{nR} S_{nR}(\theta) \quad \text{where} \quad S_{nR}(\theta) \equiv \sum_{i=1}^{n} \sum_{r=1}^{R} q(z_i, \omega_r, \theta), \tag{1}$$

with kernel function $q(z_i, w_r, \theta)$ and its associated projections

$$g(z_i, \theta) = Qq(z_i, \cdot, \theta) \quad \text{and} \quad h(w_r, \theta) \equiv Pq(\cdot, w_r, \theta).$$

The following assumption restricts the complexity of the kernel function and its projections viewed as classes indexed by the parameter θ. Recall that a class of functions is called Euclidean if their graphs form a polynomial class of sets.

ASSUMPTION 2 For each $j = 1, \ldots, d$, the following three classes of functions

$$\begin{aligned} \mathcal{F}_j &= \{q_j(z_i, w_r, \theta),\ \theta \in \Theta\}, \\ \mathcal{QF}_j &= \{g_j(z_i, \theta),\ \theta \in \Theta\}, \\ \mathcal{PF}_j &= \{h_j(w_r, \theta),\ \theta \in \Theta\}, \end{aligned}$$

are Euclidean. Their envelope functions, denoted respectively by F_j, QF_j and PF_j, have at least two moments.

By Definition (2.7) in Pakes and Pollard (1989) (hereafter P&P), a class of functions $\mathcal{F}$ is called Euclidean for the envelope F if there exist positive constants A and V that do not depend on measures μ, such that if μ is a measure for which $\int F d\mu < \infty$, then for each $\epsilon > 0$, there are functions $f_1, \ldots, f_k$ in $\mathcal{F}$ such that (i) $k \leq A\epsilon^{-V}$; (ii) For each f in $\mathcal{F}$, there is an f_i with $\int |f - f_i| d\mu \leq \epsilon \int F d\mu$.

This assumption is satisfied by many known functions. A counter example is given on page 2252 of Andrews (1994). In the case of binary choice models, it is satisfied given common low level conditions on the random utility functions. For example, when the random utility function is linear with an addititive error term, $q(z_i, w_r, \theta)$ typically takes a form that resembles $1(z_i'\theta + w_r \geq 0)$, which is Euclidean by Lemma 18 in Pollard (1984). As another example, in random coefficient binary choice models, the conditional choice probability is typically the integral of a distribution function of a single index $\Lambda(x_i'\beta)$ over the distribution of the random coefficient β. Suppose β follows a normal distribution with mean $v_i'\theta_1$ and a variance matrix with Cholesky factor θ_2, then the choice probability is given by, for $\phi(\cdot; \mu, \Sigma)$ normal density function with mean μ and variance matrix Σ, $\int \Lambda(x_i'\beta)\, \phi(\beta; v_i'\theta_1, \theta_2'\theta_2)\, d\beta$. In

this model, for draws ω_r from the standard normal density, and for $z_i = (x_i, v_i)$, $q(z_i, w_r, \theta)$ takes a form that resembles

$$\Lambda\left(x_i'\left(v_i\theta_1 + \theta_2'\omega_r\right)\right) = \Lambda\left(x_i'v_i\theta_1 + \sum_{k=1}^{K} x_{ik}\theta_{2k}'\omega_r\right).$$

As long as $\Lambda(\cdot)$ is a monotone function, this function is Euclidean according to Lemma 2.6.18 in Van der Vaart and Wellner (1996).

Under assumption 2, the following lemma is analogous to Theorems 2.5, 2.7 and 2.9 of Neumeyer (2004). In the vector case, the notation of $||\cdot||$ (e.g. $||\tilde{S}_{nR}(\theta)||$) denotes Euclidean norms.

LEMMA 1 Under Assumptions 1 and 2 the following statements hold:

a. Define, for $g = (P \otimes Q)\, q$

$$\tilde{q}(z, \omega, \theta) = q(z, \omega, \theta) - g(z, \theta) - h(w, \theta) + g(\theta),$$

then

$$\sup_{\theta \in \Theta} ||\tilde{S}_{nR}(\theta)|| = O_p(\sqrt{nR}),$$

where

$$\tilde{S}_{nR}(\theta) \equiv \sum_{i=1}^{n}\sum_{r=1}^{R} \tilde{q}(z_i, \omega_r, \theta).$$

b. Define, for $m = n \wedge R$,

$$U_{nR}(\theta) \equiv \sqrt{m}\left(\frac{1}{nR}S_{nR}(\theta) - g(\theta)\right),$$

then

$$\sup_{d(\theta_1,\theta_2)=o(1)} ||U_{nR}(\theta_1) - U_{nR}(\theta_2)|| = o_p(1).$$

where $d(\theta_1, \theta_2)$ denotes the Euclidean distance $\sqrt{(\theta_1 - \theta_2)'(\theta_1 - \theta_2)}$.

c. Further,

$$\sup_{\theta \in \Theta} \left|\left|\frac{1}{nR}S_{nR}(\theta) - g(\theta)\right|\right| = o_p(1).$$

Lemma 1 will be applied in combination with Lemma 3 in the appendix, which restates a version of Theorem 7.2 of Newey and McFadden (1994) and Theorem 3.3 of Pakes and Pollard (1989) in a form that is suitable for our purpose.

Consistency, under the conditions stated in Lemma 1, is an immediate consequence of part (c) of Lemma 1 and Corollary 3.2 of Pakes and Pollard (1989). Asymptotic normality is an immediate consequence of Lemma 3.

THEOREM 1 Given Assumptions 1 and 2, $\hat{\theta} \xrightarrow{p} \theta_0$ under the following conditions: (a) $g(\theta) = 0$ if and only if $\theta = \theta_0$; (b) $W_n \xrightarrow{p} W$ for W positive definitive; (c) $g(\theta)$ is continuously differentiable at θ_0 with a full rank derivative matrix G; and (d)

$$\left\|\hat{g}(\hat{\theta})\right\|_{W_n} = \|\hat{g}(\theta_0)\|_{W_n} + o_p(1).$$

Furthermore, if $\left\|\hat{g}(\hat{\theta})\right\|_{W_n} = \|\hat{g}(\theta_0)\|_{W_n} + o_p(m^{-1/2})$, and if $R/n \to \kappa \in [0, \infty]$ as $n \to \infty$, $R \to \infty$, then under Assumptions 1 and 2,

$$\sqrt{m}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}). \qquad \blacksquare$$

with $\Sigma = (1 \wedge \kappa)\Sigma_g + (1 \wedge 1/\kappa)\Sigma_h$, where $\Sigma_g = \text{Var}(g(z_i, \theta_0))$ and $\Sigma_h = \text{Var}(h(\omega_r, \theta_0))$. $\blacksquare$.

The asymptotic distribution for independent draws simulators with a finite number of simulations R takes the form of

$$\sqrt{n}\hat{g}(\theta_0) \xrightarrow{d} N\left(0, \left(1 + \tfrac{1}{R}\Sigma_g\right)\right)$$

This typically dominates using overlapping draws even when the number of overlapping simulations diverges to infinity. In particular, note that R has to go to infinity with overlapping draws. In contrast, with independent draws, a finite R only incurs an efficiency loss of the order of $1/R$. Recall that with independent draws, R is used to denote the number of simulations *per observation*: the total number of simulations is $n \times R$ and still increases to infinity when n increases without bound.

3.1 MSM Variance Estimation

Each component of the asymptotic variance can be estimated using sample analogs. A consistent estimate $\hat{G}$ of G, with individual elements G_j, can be formed by numerical differentiation, for e_j being a $d_\theta \times 1$ vector with 1 in the jth position and 0 otherwise, and δ_m a

step size parameter

$$\hat{G}_j \equiv \hat{G}_j\left(\hat{\theta}, \delta_m\right) = \frac{1}{2\delta_m}\left[\hat{g}(\hat{\theta} + e_j\delta_m) - \hat{g}(\hat{\theta} - e_j\delta_m)\right].$$

A sufficient, although likely not necessary, condition for $\hat{G}(\hat{\theta}) \xrightarrow{p} G(\theta_0)$ is that both $\delta_m \to 0$ and $\sqrt{m}\delta_m \to \infty$. Under these conditions, Lemma 1.b implies that $\hat{G}_j - G_j(\hat{\theta}) \xrightarrow{p} 0$, and $G_j(\hat{\theta}) \xrightarrow{p} G_j(\theta_0)$ as both $\delta_m \to 0$ and $\hat{\theta} \xrightarrow{p} \theta_0$. Σ can be consistently estimated by

$$\hat{\Sigma} = (1 \wedge R/n)\,\hat{\Sigma}_g + (1 \wedge n/R)\,\hat{\Sigma}_h,$$

where

$$\hat{\Sigma}_g = \frac{1}{n}\sum_{i=1}^{n} \hat{g}(z_i, \hat{\theta})\,\hat{g}'(z_i, \hat{\theta}) \quad \text{and} \quad \hat{\Sigma}_h = \frac{1}{R}\sum_{r=1}^{R} \hat{h}(\omega_r, \hat{\theta})\,\hat{h}'(\omega_r, \hat{\theta}).$$

In the above

$$\hat{h}(\omega, \theta) = \frac{1}{n}\sum_{i=1}^{n} q(z_i, \omega, \theta).$$

Resampling methods, such as bootstrap and subsampling, or MCMC, can also be used for inference.

4 Asymptotics of MSL with overlapping simulations

In this section we derive the asymptotic properties of maximum simulated likelihood estimators with overlapping simulations, which requires a different approach due to the nonlinearity of the log function. Recall that MSL is defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \hat{L}(\theta),$$

where

$$\hat{L}(\theta) = P_n \log Q_R\, q(\cdot, \cdot, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{R}\sum_{r=1}^{R} q(z_i, \omega_r, \theta) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{g}(z_i, \theta);$$

$\hat{L}(\theta)$ and $\hat{\theta}$ are implicitly indexed by $m = \min(n, R)$.

To begin with, the class of functions $q(z, \cdot, \theta)$ of ω indexed by both θ and z is required to be a VC-class, as defined in Van der Vaart and Wellner (1996) (pp 134, 141). Frequently $g(z, \theta)$ is a conditional likelihood in the form of $g(y \mid x, \theta)$ where $z = (y, x)$ includes both the dependent variable and the covariates. The "densities" $g(z_i; \theta)$ are broadly interpreted to

include also probability mass functions for discrete choice models or a mixture of probability density functions and probability mass functions for mixed discrete-continuous models.

ASSUMPTION 3 Both the class of functions indexed by both θ and z: $\mathcal{L} = \{ q(z, \cdot, \theta) : z \in Z,\ \theta \in \Theta\}$ and the class $\{g(\cdot, \theta), \theta \in \Theta\}$, have uniformly bounded envelopes.

The following boundedness assumption is restrictive, but is difficult to relax for nonsmooth simulators using empirical process theory. It is also assumed in Lee (1992, 1995).

ASSUMPTION 4 There is an $M < \infty$ such that $\sup_{z,\theta} \left| \frac{1}{g(z,\theta)} \right| < M$.

Let $L(\theta) = P \log g(\cdot; \theta)$. The VC property and boundedness assumption ensures uniform convergence.

LEMMA 2 Under Assumptions 1, 2, 3, and 4, $\hat{L}(\theta) - \hat{L}(\theta_0)$ converges to $L(\theta) - L(\theta_0)$ as $m \to \infty$ uniformly over Θ, which is assumed to be compact. Furthermore, if

1. $\hat{L}(\hat{\theta}) \geq \hat{L}(\theta_0) - o_p(1)$
2. For any $\delta > 0$, $\sup_{\|\theta - \theta_0\| \geq \delta} L(\theta) < L(\theta_0)$

then $\hat{\theta} - \theta_0 \xrightarrow{p} 0$.

Proof Consider the decomposition

$$\hat{L}(\theta) - L(\theta) - \hat{L}(\theta_0) + L(\theta_0) = A(\theta) + B(\theta)$$

where

$$\begin{aligned} A(\theta) &= (P_n - P)[\log g(\cdot, \theta) - \log g(\cdot, \theta_0)] \\ B(\theta) &= P_n[\log \hat{g}(\cdot, \theta) - \log \hat{g}(\cdot, \theta_0) - \log g(\cdot, \theta) - \log g(\cdot, \theta_0)]. \end{aligned} \tag{2}$$

First, by Theorem 19.13 of van der Vaart (1999), $A(\theta)$ converges uniformly to 0 in probability. By the monotonicity of log transformation and Lemma 2.6.18 (v) and (viii) in Van der Vaart and Wellner (1996), $\log \circ \mathcal{QF} - \log g(\cdot, \theta_0)$ is VC-subgraph.

Second, we show that $B(\theta)$ converges uniformly to 0 in probability as $R \to \infty$. By Taylor's theorem and Assumption 4,

$$\begin{aligned}
\sup_{\theta} |B(\theta)| &\leq 2 \sup_{z,\theta} |\log \hat{g}(z,\theta) - \log g(z,\theta)| \\
&= 2 \sup_{z,\theta} \left| \frac{\hat{g}(z,\theta) - g(z,\theta)}{g^*(z,\theta)} \right| \qquad \text{for } g^*(z,\theta) \in [g(z,\theta), \hat{g}(z,\theta)] \\
&\leq 2M \sup_{z,\theta} |\hat{g}(z,\theta) - g(z,\theta)|
\end{aligned}$$

Moreover, by Assumption 3 and Theorem 19.13 of van der Vaart (1999), as $R \to \infty$,

$$\sup_{z,\theta} |\hat{g}(z,\theta) - g(z,\theta)| \xrightarrow{p} 0.$$

Therefore, $B(\theta)$ converges uniformly to 0 as $R \to \infty$. The first part of the lemma then follows from the triangle inequality.

Consistency in the second part is a direct consequence of Theorem 2.1 in Newey and McFadden (1994) from uniform convergence when the true parameter is uniquely identified. □

In the remainder 2 of this section, we investigate the asymptotic normality of MSL, which requires that the limiting population likelihood is at least twice differentiable. First Lemma 4 in the appendix recalls a general result (see for example Sherman (1993) for optimization estimators and Chernozhukov and Hong (2003) for MCMC estimators, among others).

The following analysis consists of verifying the conditions in the general Lemma 4. The finite sample likelihood, without simulation, is required to satisfy the stochastic differentiability condition as required in the following high level assumption. It is typically satisfied when the true non-simulated log likelihood function is pointwise differentiable.

ASSUMPTION 5 The true log likelihood function $\log g(\cdot, \theta)$ satisfies the conditions in Lemma 3.2.19, p. 302, of Van der Vaart and Wellner (1996). In particular, the conditions in Lemma 3.2.19 of Van der Vaart and Wellner (1996) require the existence of $\dot{m}(\cdot)$ such that for some $\delta > 0$,

$$\left\{ \frac{\log g(\cdot,\theta) - \log g(\cdot,\theta_0) - \dot{m}(\cdot)'(\theta - \theta_0)}{||\theta - \theta_0||} : ||\theta - \theta_0|| < \delta \right\}$$

is P-Donsker, and

$$P\left(\log g(\cdot,\theta) - \log g(\cdot,\theta_0) - \dot{m}'(\theta - \theta_0)\right)^2 = o\left(||\theta - \theta_0||^2\right).$$

Under Assumption 5, for $D_0(\cdot) = \dot{m}(\cdot) - Em(\cdot)$, where $ED_0(\cdot)^2 < \infty$, such that for any $\delta_m \to 0$ we have

$$\sup_{\|\theta-\theta_0\|\leq\delta_m} \frac{n\text{Remainder}_n(\theta)}{1+n\|\theta-\theta_0\|^2} = o_p(1) \tag{3}$$

for

$$\text{Remainder}_n(\theta) \equiv (P_n - P)(\log g(\cdot,\theta) - \log g(\cdot,\theta_0)) - \hat{D}_0'(\theta-\theta_0),$$

where

$$\hat{D}_0 = \frac{1}{n}\sum_{i=1}^{n} D_0(z_i).$$

To account for the simulation error we need an intermediate step which is a modification of Theorem 1 of Sherman (1993). This intermediate step is given in Lemma 5 in the appendix.

The next assumption requires that the simulated likelihood is not only unbiased, but is also a proper likelihood function.

ASSUMPTION 6 For all simulation lengths R and all parameters θ, both $g(z_i;\theta)$ and $Q_R q(z_i,\cdot;\theta)$ are proper (possibly conditional) density functions.

We also need to regulate the amount of irregularity that can be allowed by the simulation function $q(z,\omega,\theta)$. The following assumption allows for $q(z,\omega,\theta)$ to be an indicator function, and is related to Theorem 2.37 of Pollard (1984). When $q(z,\omega,\theta)$ is Lipschitz in θ, a stronger condition holds where the right hand size of (2) can be replaced by $O(\delta_m^2)$. In the following, $Q\otimes P$ denotes independent expectations taken with respect to the two arguments in $f(\cdot,\cdot,\theta)$.

ASSUMPTION 7 Define $f(z,\omega,\theta) = q(z,\omega,\theta)/g(z,\theta) - q(z,\omega,\theta_0)/g(z,\theta_0)$, then (1) $Q\otimes P\left[\sup_{\|\theta-\theta_0\|=o(1)} f(\cdot,\cdot,\theta)^2\right] = o(1)$, (2) $\sup_{\|\theta-\theta_0\|\leq\delta_m, z\in Z} \text{Var}_\omega f(z,\omega,\theta) = O(\delta_m)$.

ASSUMPTION 8 Define $\psi(\omega,\theta) = \int \frac{q(z,\omega,\theta)}{g(z,\theta)} f(z)\,dz$, where $f(z)$ is the joint density or probability mass function of the data. The function $\psi(\omega,\theta)$ also satisfies the conditions in Assumption 5: there exists $D_1(\cdot)$ with $Var(D_1(\omega_r)) < \infty$, such that for some $\delta > 0$

$$\left\{\frac{\psi(\cdot,\theta) - \psi(\cdot,\theta_0) - D_1(\cdot)'(\theta-\theta_0)}{||\theta-\theta_0||} : ||\theta-\theta_0|| < \delta\right\}$$

is Q-Donsker, and

$$Q(\psi(\cdot,\theta) - \psi(\cdot,\theta_0) - D_1(\theta-\theta_0))^2 = o\left(||\theta-\theta_0||^2\right).$$

Remark 1 Under Assumption 8, for $\hat{D}_1 = \frac{1}{R}\sum_{r=1}^{R} D_1(\omega_r) - QD_1(\omega_r)$,

$$\sup_{\|\theta-\theta_0\|=o\left((\log R)^{-1}\right)} \frac{R(Q_R - Q)(\psi(\cdot,\theta) - \psi(\cdot,\theta_0)) - R\hat{D}_1'(\theta-\theta_0)}{1 + R\|\theta-\theta^0\|^2} = o_p(1)$$

Remark 2 When $g(z;\theta)$ represents the joint likelihood of the data, $f(z) = g(z;\theta_0)$. When $g(z;\theta) = g(y|x;\theta)$ represents a conditional likelihood, $f(z) = g(z;\theta_0) f(x)$ where $f(x)$ is the marginal density or probability mass function of the conditioning variables, in which case $\psi(\omega,\theta) = \int\int \frac{q(z,\omega,\theta)}{g(z,\theta)} g(y|x;\theta_0)\, dy f(x)\, dx$, with the understanding that integrals become summations in the case of discrete data.

Remark 3 Assumption 8 can be further simplified when the true likelihood $g(z,\theta)$ is twice continuously differentiable (with bounded derivatives for simplicity). To verify this statement, note that in this case

$$D_1(\omega_r) = -\int \frac{q(\omega_r, z, \theta_0)}{g^2(z;\theta_0)} \frac{\partial}{\partial\theta} g(z;\theta_0) f(z)\, dz. \tag{4}$$

Equation (4) applies when $g(z;\theta)$ is the joint likelihood of the data. When $g(z;\theta)$ is a conditional likelihood $g(z;\theta) = g(y|x;\theta)$, $D_1(\omega_r) = -\int \frac{q(\omega_r,z,\theta_0)}{g(z;\theta_0)} \frac{\partial}{\partial\theta} g(z;\theta_0) f(x)\, dz$. To see (4), note that

$$\begin{aligned}
&(Q_R - Q)(\psi(\cdot,\theta) - \psi(\cdot,\theta_0)) \\
&\quad = P\left[\frac{1}{g(\cdot,\theta)} - \frac{1}{g(\cdot,\theta_0)}\right](\hat{g}(\cdot,\theta_0) - g(\cdot,\theta_0)) \\
&\quad + P\frac{1}{g(\cdot,\theta_0)}(\hat{g}(\cdot,\theta) - g(\cdot,\theta) - \hat{g}(\cdot,\theta_0) + g(\cdot,\theta_0)) \\
&\quad + P\left(\frac{1}{g(\cdot,\theta)} - \frac{1}{g(\cdot,\theta_0)}\right)(\hat{g}(\cdot,\theta) - g(\cdot,\theta) - \hat{g}(\cdot,\theta_0) + g(\cdot,\theta_0)).
\end{aligned}$$

The second line is zero because of assumption 6. The third line can be bounded by

$$M\|\theta-\theta_0\| \sup_{\|\theta-\theta_0\|=o\left((\log R)^{-1}\right), z\in Z} |(Q_R - Q)(q(\cdot,z,\theta) - q(\omega_r,z,\theta_0))| = o_p\left(\frac{1}{\sqrt{R}}\right)\|\theta-\theta_0\|,$$

using the same arguments that handle the $B_{22}(\theta,z)$ in the proof. Finally, the first line becomes

$$P\left[\frac{1}{g(\cdot,\theta)} - \frac{1}{g(\cdot,\theta_0)}\right](\hat{g}(\cdot,\theta_0) - g(\cdot,\theta_0)) = (Q_R - Q) D_1(\cdot)(\theta-\theta_0) + \text{Remainder}(\theta),$$

where $\|\text{Remainder}(\theta)\| \le o_p(\|\theta-\theta_0\|)\,|\sup_{z\in Z}(Q_R - Q) q(\cdot,z,\theta_0)| = o_p\left(\frac{\|\theta-\theta_0\|}{\sqrt{R}}\right)$.

THEOREM 2 Under Assumptions 1 2, 3, 4, 5, 6, 7, and 8 and Conditions 1, 2, 3 and 4 of Theorem 4, the conclusion of Theorem 4 holds with $\hat{D} = P_n D_0(\cdot) + Q_R D_1(\cdot)$ and

$$\Sigma = (1 \wedge \kappa) \operatorname{Var}(D_0(z_i)) + (1 \wedge 1/\kappa) \operatorname{Var}(D_1(\omega_r)).$$

Proof Consistency is given in Lemma 2. Consider again the decomposition given by Equation (2). Because of the linearity structure of Conditions (5) and (6) of Lemma 4, it suffices to verify them separately for the terms $A(\theta)$ and $B(\theta)$.

It follows immediately from Assumption 5 that Conditions (5) and (6) of Lemma 4 hold for the first term $A(\theta)$ because $n \geq m$, since (3) is increasing in n:

$$\sup_{||\theta-\theta_0|| \leq \delta_m} \frac{m\left(A(\theta) - \hat{D}_0'(\theta-\theta_0)\right)}{1 + m||\theta-\theta_0||^2} \leq \sup_{||\theta-\theta_0|| \leq \delta_m} \frac{\hat{R}_0(\theta)}{1/n + ||\theta-\theta_0||^2} = o_P(1), \tag{5}$$

for $\hat{R}_0(\theta) = A(\theta) - \hat{D}_0'(\theta-\theta_0)$. Next we verify these conditions for the $B(\theta)$ term.

Decompose B further into $B(\theta) = B_1(\theta) + B_2(\theta) + B_3(\theta)$, where

$$\begin{aligned}
B_1(\theta) &= P_n\left[\frac{1}{g(\cdot,\theta)}\left(\hat{g}(\cdot,\theta) - g(\cdot,\theta)\right) - \frac{1}{g(\cdot,\theta)}\left(\hat{g}(\cdot,\theta_0) - g(\cdot,\theta_0)\right)\right] \\
B_2(\theta) &= -\frac{1}{2}P_n\left[\frac{1}{g(\cdot,\theta)^2}\left(\hat{g}(\cdot,\theta) - g(\cdot,\theta)\right)^2 - \frac{1}{g(\cdot,\theta_0)^2}\left(\hat{g}(\cdot,\theta_0) - g(\cdot,\theta_0)\right)^2\right] \\
B_3(\theta) &= \frac{1}{3}P_n\left[\frac{1}{\bar{g}(\cdot,\theta)^3}\left(\hat{g}(\cdot,\theta) - g(\cdot,\theta)\right)^3 - \frac{1}{\bar{g}(\cdot,\theta_0)^3}\left(\hat{g}(\cdot,\theta_0) - g(\cdot,\theta_0)\right)^3\right].
\end{aligned}$$

In the above $\bar{g}(z,\theta)$ and $\bar{g}(z,\theta_0)$ are mean values, dependent on z, between $[g(z,\theta), \hat{g}(z,\theta)]$ and $[g(z,\theta_0), \hat{g}(z,\theta_0)]$ respectively. By Assumption 4,

$$\sup_{\theta\in\Theta} |B_3(\theta)| \leq \frac{2}{3}M^3 |\sup_{\theta\in\Theta, z\in Z} \left(\hat{g}(z,\theta) - g(z,\theta)\right)|^3 \leq O_p\left(\frac{1}{R\sqrt{R}}\right),$$

where the last inequality follows from $\sup_{\theta\in\Theta, z\in Z} |\hat{g}(z,\theta) - g(z,\theta)| = O_p\left(\frac{1}{\sqrt{R}}\right)$ due, e.g., to Theorem 2.14.1 of Van der Vaart and Wellner (1996). By Theorem 2.14.1 it also holds that

$$\sup_{\theta\in\Theta} |B_1(\theta)| = O_p\left(\frac{1}{\sqrt{R}}\right) \quad \text{and} \quad \sup_{\theta\in\Theta} |B_2(\theta)| = O_p\left(\frac{1}{R}\right).$$

This allows us to invoke Theorem 5, with $d_m = \sqrt{m}$, to claim that

$$\|\hat{\theta} - \theta_0\| = O_p\left(m^{-1/4}\right).$$

Next we bound the second term by, up to a constant, within $\|\hat{\theta}-\theta_0\| = o_p\left(1/\log R\right)$:

$$\sup_{\|\theta-\theta_0\|\ll(\log R)^{-1}} |B_2(\theta)| = o_p\left(\frac{1}{R}\right). \tag{6}$$

To show (6), first note that

$$\sup_{\|\theta-\theta_0\|\ll(\log R)^{-1}} |B_2(\theta)| \leq \sup_{\|\theta-\theta_0\|\ll(\log R)^{-1}, z\in Z} B_{21}(\theta,z) \times B_{22}(\theta,z)$$

where

$$B_{21}(\theta,z) = \left| (Q_R - Q)\left(\frac{q(z,\cdot,\theta)}{g(z,\theta)} + \frac{q(z,\cdot,\theta_0)}{g(z,\theta_0)}\right)\right|$$

and

$$B_{22}(\theta,z) = \left| (Q_R - Q)\left(\frac{q(z,\cdot,\theta)}{g(z,\theta)} - \frac{q(z,\cdot,\theta_0)}{g(z,\theta_0)}\right)\right|.$$

It follows again from Theorem 2.14.1 of Van der Vaart and Wellner (1996) that

$$\sup_{\|\theta-\theta_0\|\ll(\log R)^{-1}, z\in Z} |B_{21}(\theta,z)| = O_p\left(\frac{1}{\sqrt{R}}\right).$$

Next we consider $B_{22}(\theta,z)$ in light of arguments similar to Theorem 2.37 in Pollard (1984), for which it follows that for $\delta_m = o\left((\log R)^{-1}\right)$, for

$$f(z,\omega,\theta) = q(z,\omega,\theta)/g(z,\theta) - q(z,\omega,\theta_0)/g(z,\theta_0)$$

where $\|\theta-\theta_0\| \leq \delta_m$, and for $\epsilon_R = \epsilon/\sqrt{R}$: $\mathrm{Var}\left(Q_R f(z,\cdot,\theta)\right)/\epsilon_R^2 \to 0$ for each $\epsilon > 0$. Therefore the symmetrization inequalities (30) in p. 31 of Pollard (1984) apply and subsequently, for $\mathcal{F}_R = \{f(z,\omega,\theta), z\in Z, \|\theta-\theta_0\| \leq \delta_m\}$,

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}_R}\left|(Q_R-Q)f\right| > 8\frac{\epsilon}{\sqrt{R}}\right) & \\ \leq\ & 4P\left(\sup_{f\in\mathcal{F}_R}|Q_R^0 f| > 2\frac{\epsilon}{\sqrt{R}}\right) \\ \leq\ & 8A\epsilon^{-W}R^{W/2}\exp\left(-\frac{1}{128}\epsilon^2\delta_m^{-1}\right) + P\left(\sup_{f\in\mathcal{F}_R} Q_R f^2 > 64\delta_m\right). \end{aligned}$$

The second term goes to zero for the same reason as in Pollard. The first also goes to zero since $\log R - \frac{1}{\delta_m} \to -\infty$. Thus we have shown that $B_{22}(\theta,z) = o_p\left(\frac{1}{\sqrt{R}}\right)$ uniformly in $\theta-\theta_0 \leq \delta_m$ and $z\in Z$, and consequently (6) holds. By considering $n \gg R$, $n \ll R$ and $n \approx R$ separately, (6) also implies that for some $\alpha > 0$:

$$\sup_{\|\theta-\theta_0\|\ll m^{-\alpha}} |B_2(\theta)| = o_p\left(\frac{1}{m}\right).$$

It remains to investigate $B_1(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R} f(z_i, \omega_r, \theta)$, which, using Assumption 6, can be written

$$B_1(\theta) = \frac{1}{nR} S_{nR}\left(\tilde{f}(\cdot,\cdot,\theta)\right) + B_0(\theta),$$

where, for S_{nR} defined in (1),

$$\tilde{f}(z,\omega,\theta) = f(z,\omega,\theta) - Qf(z,\cdot,\theta) - Pf(\cdot,\omega,\theta) + P\otimes Qf(\cdot,\cdot,\theta),$$

$B_0(\theta) = (Q_R - Q)(\psi(\cdot,\theta) - \psi(\omega,\theta_0))$, and $\psi(\omega,\theta) = \int \frac{q(z,\omega,\theta)}{g(z,\theta)} f(z)\, dz$. Upon noting that $Q\frac{q(z,\cdot,\theta)}{g(z,\theta)} = 1$ identically, $Qf(z,\cdot,\theta) = 0$ and $P\otimes Qf(z,\cdot,\theta) = 0$. Consider $S_{nR}\left(\tilde{f}(\cdot,\cdot,\theta)\right)$ as a U-process indexed by the class $\tilde{\mathcal{F}}_{n,R} = \{\tilde{f}(\cdot,\cdot,\theta), \theta\in\Theta, ||\theta-\theta_0|| \le r_{n,R}\}$, where $r_{n,R} = o(1)$, with its envelope denoted as $F_{n,R}$. It then follows immediately from the proof of Theorem 2.5 (pp. 83) of Neumeyer (2004) that

$$\frac{1}{nR} S_{nR}(f(\cdot,\cdot,\theta)) = o_p\left(\frac{1}{\sqrt{nR}}\right) = o_p\left(\frac{1}{m}\right).$$

Finally, B_0 is handled by Assumption 8. So that, since $B(\theta) = B_1(\theta) - B_0(\theta) + B_2(\theta) + B_3(\theta) + B_0(\theta)$, and each of $B_1(\theta) - B_0(\theta)$, $B_2(\theta)$ and $B_3(\theta)$ is $o_P\left(\frac{1}{R}\right)$ uniformly in $||\theta-\theta_0|| \le \delta_m$, for any $\delta_m \to 0$, we have, for $\hat{R}_1(\theta) = B(\theta) - \hat{D}_1'(\theta-\theta_0)$,

$$\sup_{||\theta-\theta_0||\le\delta_m} \frac{R\hat{R}_1(\theta)}{1+R||\theta-\theta_0||^2} = \sup_{||\theta-\theta_0||\le\delta_m} \frac{RB_0(\theta) - R\hat{D}_1'(\theta-\theta_0)}{1+R||\theta-\theta_0||^2} + o_P(1) = o_P(1).$$

This together with (5) implies that condition 6 of Theorem 4 is satisfied with $\hat{D} = \hat{D}_0 + \hat{D}_1$, since we can bound (8) by

$$(8) = \sup_{||\theta-\theta_0||\le\delta_m} \frac{\hat{R}_0(\theta) + \hat{R}_1(\theta)}{1/m + ||\theta-\theta_0||^2} \le \sup_{||\theta-\theta_0||\le\delta_m} \frac{\hat{R}_0(\theta)}{1/n + ||\theta-\theta_0||^2} + \sup_{||\theta-\theta_0||\le\delta_m} \frac{\hat{R}_1(\theta)}{1/R + ||\theta-\theta_0||^2}.$$

Finally to verify condition 5 in Theorem 4, write

$$\sqrt{m}\hat{D} = \sqrt{\left(1\wedge\frac{R}{n}\right)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_0(z_i) + \sqrt{\left(\frac{n}{R}\wedge 1\right)}\frac{1}{\sqrt{R}}\sum_{r=1}^{R} D_1(\omega_r).$$

That $\sqrt{m}\hat{D} \xrightarrow{d} N(0,\Sigma)$ follows from $1\wedge\frac{R}{n} \to 1\wedge\kappa$, $\frac{n}{R}\wedge 1 \to \frac{1}{\kappa}\wedge 1$, the continuous mapping Theorem, Slutsky's Lemma, and CLTs applied to $\sqrt{n}\hat{D}_0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_0(z_i)$ and $\sqrt{R}\hat{D}_1 = \frac{1}{\sqrt{R}}\sum_{r=1}^{R} D_1(\omega_r)$. □

4.1 MSL Variance Estimation

A consistent estimate of the asymptotic variance can be formed by sample analogs. In general, each of

$$\hat{H} = P_n \frac{\partial^2}{\partial\theta\partial\theta'} \log Q_R q\left(\cdot,\cdot,\hat{\theta}\right), \quad \hat{D}_0\left(z_i\right) = \frac{\partial}{\partial\theta} \log \hat{g}\left(z_i,\hat{\theta}\right) \quad \text{and} \quad \hat{D}_1\left(\omega_r\right) = \frac{\partial}{\partial\theta} P_n \frac{q\left(\omega_r,\cdot,\hat{\theta}\right)}{\hat{g}\left(\cdot,\hat{\theta}\right)}$$

can not be computed analytically, and has to be replaced by numerical estimates:

$$\begin{aligned}
\hat{H}_{ij} &= \frac{1}{4\epsilon^2}\Big(P_n \log Q_R q\left(\cdot,\cdot,\hat{\theta}+e_i\epsilon+e_j\epsilon\right) - P_n \log Q_R q\left(\cdot,\cdot,\hat{\theta}-e_i\epsilon+e_j\epsilon\right) \\
&\quad -P_n \log Q_R q\left(\cdot,\cdot,\hat{\theta}+e_i\epsilon-e_j\epsilon\right) + P_n \log Q_R q\left(\cdot,\cdot,\hat{\theta}-e_i\epsilon-e_j\epsilon\right)\Big), \\
\hat{D}_{0j}\left(z_i\right) &= \frac{1}{2h}\left(\log \hat{g}\left(z_i,\hat{\theta}+e_j h\right) - \log \hat{g}\left(z_i,\hat{\theta}-e_j h\right)\right), \\
\hat{D}_{1j}\left(w_r\right) &= \frac{1}{2h}\left(P_n \frac{q\left(\omega_r,\cdot,\hat{\theta}+e_j h\right)}{\hat{g}\left(\cdot,\hat{\theta}+e_j h\right)} - P_n \frac{q\left(\omega_r,\cdot,\hat{\theta}-e_j h\right)}{\hat{g}\left(\cdot,\hat{\theta}-e_j h\right)}\right).
\end{aligned}$$

Let, for $\hat{\kappa} = R/n$,

$$\hat{\Sigma}_h = P_n \hat{D}_0\left(\cdot\right)\hat{D}_0\left(\cdot\right)' \qquad \hat{\Sigma}_g = Q_R \hat{D}_1\left(\cdot\right)\hat{D}_1\left(\cdot\right) \qquad \hat{\Sigma} = \left(1\wedge\hat{\kappa}\right)\hat{\Sigma}_h + \left(1\wedge 1/\hat{\kappa}\right)\hat{\Sigma}_g.$$

Under the given assumptions, if $\epsilon \to 0$, $h \to 0$, $\sqrt{n}h \to \infty$ and $n^{\frac{1}{4}}\epsilon \to \infty$, then $\hat{H} = H + o_p\left(1\right)$ and $\hat{\Sigma}_h = \Sigma_h + o_p\left(1\right)$, $\hat{\Sigma}_g = \Sigma_g + o_p\left(1\right)$. Hence $\hat{\Sigma} = \Sigma + o_p\left(1\right)$ by the continuous mapping theorem.

5 MCMC

Simulated objective functions that are nonsmooth can be difficult to optimize by numerical methods. An alternative to optimizing the objective function is to run it through a MCMC routine, as in Chernozhukov and Hong (2003). Under the assumptions given in the previous sections, the MCMC Laplace estimators can also be shown to be consistent and asymptotically normal. The Laplace estimator is defined as

$$\tilde{\theta} = \text{argmin}_{\theta\in\Theta} \int \rho\left(\sqrt{m}\left(u-\theta\right)\right) \exp\left(m\hat{L}\left(u\right)\right) \pi\left(u\right) du.$$

In the above $\rho(\cdot)$ is a convex symmetric loss function such that $\rho(h) \leq 1 + |h|^p$ for some $p \geq 1$, and $\pi(\cdot)$ is a continuous density function with compact support and postive at θ_0. In the above the objective function can be either GMM:

$$\hat{L}(\theta) = \frac{1}{2} P_n Q_R q(\cdot, \cdot, \theta)' W_n P_n Q_R q(\cdot, \cdot, \theta),$$

or the log likelihood function $\hat{L}(\theta) = \sum_{i=1}^{n} \log \hat{g}(z_i, \theta)$.

The asymptotic distribution of the posterior distribution and $\tilde{\theta}$ follows immediately from Assumption 2, which leads to Theorem 3, and Chernozhukov and Hong (2003). Define $h = \sqrt{m}\left(\theta - \hat{\theta}\right)$, and consider the posterior distribution on the localized parameter space:

$$p_n(h) = \frac{\pi\left(\hat{\theta} + \frac{h}{\sqrt{m}}\right) \exp\left(m\hat{L}\left(\hat{\theta} + h/\sqrt{m}\right) - m\hat{L}\left(\hat{\theta}\right)\right)}{C_m}$$

where

$$C_m = \int_{\hat{\theta}+h/\sqrt{m}\in\Theta} \pi\left(\hat{\theta} + \frac{h}{\sqrt{m}}\right) \exp\left(m\hat{L}\left(\hat{\theta} + h/\sqrt{m}\right) - m\hat{L}\left(\hat{\theta}\right)\right) dh.$$

Desirable properties of the MCMC method include the following, for any $\alpha > 0$:

$$\int |h|^{\alpha} |p_n(h) - p_{\infty}(h)| dh \xrightarrow{p} 0, \quad \text{where} \quad p_{\infty}(h) = \sqrt{\frac{|\det(J_0)|}{(2\pi)^{\dim\theta}}} \exp\left(-\frac{1}{2} h' J_0 h\right). \tag{7}$$

In the above $J_0 = G'WG$ for the GMM model and $J_0 = -\frac{\partial^2}{\partial\theta\partial\theta'} L(\theta_0)$ for the likelihood model.

THEOREM 3 Under Assumptions 1 and 2 for the GMM model or Assumptions 1, 2, and 8, Conditions 1, 2, 4 of Theorem 2 for the MLE model, (7) holds. Consequently, $\sqrt{m}\left(\tilde{\theta} - \hat{\theta}\right) \xrightarrow{p} 0$, and the variance of $p_{n,R}(h)$ converges to J_0^{-1} in probability.

Proof For the GMM model, the stated results follow immediately from Assumption 2, which leads to Theorem 3, and Chernozhukov and Hong (2003) (CH). The MLE case is also almost identical to CH but requires a small modification. When Condition (6) in Theorem 4 holds for $\delta_m = o(1)$, the original proof shows (7) over three areas of integration separately, $\{|h| \leq \sqrt{m}\delta_m\}$ and $\{|h| \geq \delta_m\sqrt{m}\}$. When Condition 6 in Theorem 4 only holds for $\delta_m = a_m = (\log m)^{-d}$, we need to consider separately, for a fixed δ, $\{|h| \leq \sqrt{m}a_m\}$, $\{\sqrt{m}a_m \leq |h| \leq \sqrt{m}\delta\}$ and $\{|h| \geq \delta\sqrt{m}\}$. The arguments for the first and third regions

$\{|h| \leq \sqrt{m}a_m\}$ and $\{|h| \geq \delta\sqrt{m}\}$ are identical to the ones in CH. Hence we only need to show that (since the prior density is assumed bounded around θ_0):

$$\int_{\sqrt{m}a_m \leq |h| \leq \sqrt{m}\delta} \pi\left(\hat{\theta} + \frac{h}{\sqrt{m}}\right) \exp\left(m\hat{L}\left(\hat{\theta} + h/\sqrt{m}\right) - m\hat{L}\left(\hat{\theta}\right)\right) dh \xrightarrow{p} 0.$$

By arguments that handle the term B in the proof of Theorem 2, in this region,

$$\omega(h) \equiv m\hat{L}\left(\hat{\theta} + h/\sqrt{m}\right) - m\hat{L}\left(\hat{\theta}\right) = -\frac{1}{2}\left(1 + o_p(1)\right) h' J_0 h + mO_p\left(\frac{1}{\sqrt{m}}\right).$$

Hence the left hand side integral can be bounded by, up to a finite constant

$$\int_{\sqrt{m}a_m \leq |h| \leq \sqrt{m}\delta} \exp(\omega(h))\, dh = \exp\left(O_p\left(\sqrt{m}\right)\right) \int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{2}\left(1 + o_p(1)\right) h' J_0 h\right) dh.$$

The tail of the normal distribution can be estimated by w.p. $\to 1$:

$$\begin{aligned} &\textstyle\int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{2}\left(1 + o_p(1)\right) h' J_0 h\right) dh \\ &\leq \textstyle\int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{4} h' J_0 h\right) dh \leq C\left(\sqrt{m}a_m\right)^{-1} \exp\left(-m a_m^2\right), \end{aligned}$$

for $a_m >> m^{-\alpha}$ for any $\alpha > 0$, hence for some $\alpha > 0$.

$$\int_{\sqrt{m}a_m \leq |h| \leq \sqrt{m}\delta} \exp(\omega(h))\, dh \leq C \exp\left(O_p\left(\sqrt{m}\right)\right) \left(m^{\frac{1}{2}-\alpha}\right)^{-1} \exp\left(-m^{1-2\alpha}\right) = o_p(1).$$

The rest of the proof is identical to CH. □

The MCMC method can always be used to obtain consistent and asymptotically normal parameter estimates. For the GMM model with W being the asymptotic variance of $\sqrt{m}\hat{g}(\theta_0)$, or for the likelihood model where $n >> R$, the posterior distribution from the MCMC can also be used to obtain valid asymptotic confidence intervals for θ_0.

For the GMM model where $W \neq$ asym Var$(\sqrt{m}\hat{g}(\theta_0))$, or the likelihood model where $R >> n$, $R \sim n$, the posterior distribution does not resemble the asymptotic distribution of $\hat{\theta}$ or $\tilde{\theta}$. However, in this case the variance of the posterior distribution can still be used to estimate the inverse of the Hessian term $(G'WG)^{-1}$ or $H(\theta_0)$ in Condition (4) of Theorem 4.

6 Monte Carlo Simulations

In this section, we presented Monte Carlo results for a multinomial probit model. Let there be M alternatives and N individuals. Suppose alternative $m = M$ is the base alternative.

x_m is the value of K attributes for each alternative. It includes a constant. ϵ_i is the taste shock for individual i. Relative utility from choosing alternative $m = 1, 2, \cdots, M-1$ for individual i is:

$$y*_{im} = (\beta + \Gamma\epsilon_i)'(x_m - x_M)$$
$$\Gamma\epsilon_i \overset{\text{iid}}{\sim} MVN_K(0, \Omega)$$

So Γ is the lower triangular Cholesky decomposition of Ω: $\Omega = \Gamma\Gamma'$ and $\epsilon_i \overset{\text{iid}}{\sim} MVN_K(0, I)$.

If all $y*_{im} < 0$ for all m, then individual i chooses the base alternative M. Otherwise the alternative with the largest positive utility is chosen. Let y_{im}, $m = 1, 2, \ldots, M-1$ be 1 if alternative m is chosen and 0 otherwise. That is,

$$y_{im} = \mathbf{1}\{y*_{im} \geq 0, y^*_{im} = \max\{y*_{in}\}_{n=1}^{M-1}\}, \quad m = 1, 2, \ldots, M-1, i = 1, 2, \ldots, N$$

For our Monte Carlo simulation, we set $M = 3$, $k = 2$, $x_m - X_M = (1, \tilde{x}_m)$ where $\tilde{x}_m \overset{\text{iid}}{\sim} N(0, 1)$ for $m = 1, 2, \cdots, M-1$. $\beta = \beta_0 = [1, 1]'$. $\omega = [1, 0; 0, 0.5]$. We create $J = 1000$ samples of data from this data generating process. We assume Σ is known and estimate β.

Define $z_i = \{x_{im}, y_{im}\}_m$. The probability of observing z_i conditional on $\{x_{im}\}_m$ is

$$g(z_i, \beta, \Omega) \quad = \quad \mathbb{P}[\{y_{im}\}_m | \{x_{im}\}_m, \beta, \Omega]$$

The probability operator $\mathbb{P}$ comes from $MVN(0, \Omega)$, which can not be evaluated analytically. The true parameter β_0 maximizes the true likelihood

$$L(\beta) = \mathbb{E}_\epsilon \log g(z_i, \beta, \Omega)$$

The maximum likelihood estimator, $\hat{\beta}$, maximizes the sample likelihood

$$L_N(\beta) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} \log g(z_i, \beta, \Omega)$$

Let $w_{ri} = (w_{ri1}, w_{ri2}, \ldots, w_{ri,M-1})'$ denote the rth draw of taste shocks from MVN(0,Σ). We take R draws for each individual. The simulated likelihood for an observation is the share of draws with the observed choice:

$$\hat{g}(z_i, \beta) = \frac{1}{R}\sum_{r=1}^{R} q(w_{ri}, z_i, \beta) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{y(w_{ri}, z_i, \beta) = y_i\}$$

For overlapping draws, we use the same R draws for all i. For independent simulated maximum likelihood, we use different draws across is. The simulated maximum likelihood estimator is the β that maximizes the simulated likelihood

$$\hat{L}_{NR}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\log \hat{g}(z_i,\beta)$$

We find the solution to this maximization problem using simulated annealing with a initial guess of $\beta = [0.5,\ 0.5]$. The same algorithm is used for overlapping draws and independent draws.

We ran 1000 MC for combinations of[2]

$$n \in [50, 100, 200, 400, 800], \quad R/n \in [0.2, 0.5, 0.8, 1, 2, 5, 10, 20, 50].$$

For each MC trial, we derived asymptotic distribution to approximate the 95% confidence interval and check if the true parameter falls inside the computed interval. We calculated the asymptotic distribution using the numerical integration formula in subsections 4.1. We need to choose a stepsize for numerical integration. We tried step sizes $\epsilon = R^{-\alpha}, \alpha = \{2, 3/2, 1, 1/2, 1/3, 1/4, 1/8, 1/10, 1/15\}$. In the following we report results for the smallest stepsize $\alpha = -2$, which turned out to generate better coverage for both overlapping and independent draws[3].

We compare the accuracy of the confidence interval constructed using overlapping draws SMLE and the asymptotic distribution derived in our paper with

1. overlapping draws SMLE with asymptotic distribution that does not correct for simulation bias and variance inflation

2. independent draws SMLE with asymptotic distribution that does not correct for simulation bias and variance inflation

Table 1 reports the empirical coverage of the 95% confidence interval constructed from the estimate of the asymptotic distribution using numerical derivatives, over 1000 Monte Carlo

[2]For one subset of the simulations, we also looked at $R/n = 100$ and found that it is similar to the $R/n = 50$ results. Hence we set the largest R/n to 50.

[3]We found that overlapping draws with adjustment for simulation noise performs better than the other two methods for all other stepsizes.

Table 1: Empirical coverage frequency for 95% confidence interval, numerical derivatives

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	3	24	49	59	83	93	94	95	95
	3	23	49	59	84	92	93	93	95
100	9	53	77	82	91	93	94	95	94
	10	55	77	82	88	93	94	94	94
200	34	78	87	90	92	93	95	95	95
	34	78	87	88	92	93	94	94	95
400	66	89	91	91	93	95	95	95	95
	66	88	90	91	92	93	94	95	94
800	83	91	92	93	93	95	95	95	96
	83	91	91	92	92	94	94	94	94

The total number of Monte Carlo repetitions is 1000.

repetitions. The column dimension corresponds to the sample size n and the row dimension corresponds to the ratio between R and n. The two rows for each sample size correspond to the coverage for the first and second elements of β, respectively. Better accuracy means being closer to 95%. The asymptotic distribution accurately represents the finite sample distribution when $m = \min(R, n)$ is not too small.

Table 2 reports the false empirical coverage of the 95% confidence interval when the simulation noise is ignored in the asymptotic distribution of the estimator. As expected, when R/n is large, in particular above 10, the improvement from accounting for Σ_1 in the asymptotic distribution is very small. When R/n is very small, the size distortion from ignoring Σ_1 is very sizable. The size distortion is quite visible when R/n is as big as 2, and still visible even when $R/n = 5$.

Overlapping draws are applicable in situations in which independent draws are not computationally practicable, or with nonsmooth moment conditions where the theoretical validity of independent draws is more difficult and beyond the scope of the current paper. In spite of the lack of a theoretical proof, in tables 3 we report the counterpart with independent draws. The independent draws method does not perform well relative to overlapping draws when either N or R/N is small.

Table 2: False empirical coverage frequency for 95% confidence interval, numerical derivatives

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	2	20	42	51	78	90	91	93	94
50	2	18	42	53	79	89	91	92	95
100	7	45	64	71	85	90	91	93	94
100	7	44	64	70	83	90	92	93	93
200	24	65	74	80	87	91	94	94	94
200	23	62	75	77	86	89	93	93	94
400	44	73	80	83	89	93	94	94	95
400	43	69	76	79	85	89	92	94	94
800	58	78	84	86	90	93	94	94	95
800	52	71	78	83	87	91	92	93	94

The total number of Monte Carlo repetitions is 1000.

Table 3: Independent draws empirical coverage frequency for 95% confidence interval, numerical derivatives

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	1	8	30	41	77	90	91	93	94
50	1	7	29	41	79	89	94	94	95
100	2	27	63	73	87	91	92	94	94
100	1	26	65	73	87	93	93	94	95
200	4	62	83	86	92	93	94	94	95
200	3	62	81	86	90	92	93	95	95
400	33	85	88	90	92	94	94	94	94
400	28	83	87	88	91	93	92	93	94
800	71	89	91	91	94	94	94	94	95
800	70	87	91	90	92	93	93	93	94

The total number of Monte Carlo repetitions is 1000.

Table 4: Mean bias as % of the true parameter values for overlapping draws

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	0	0	1	0	0	0	0	0	0
50	-1	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0
100	0	0	0	0	0	0	0	0	0
200	0	0	0	0	0	0	0	0	0
200	0	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0
800	0	0	0	0	0	0	0	0	0
800	0	0	0	0	0	0	0	0	0

The total number of Monte Carlo repetitions is 1000.

We also report the mean bias and root mean square error for overlapping draws in tables 4 to 6 and for independent draws in tables 5 to 7. Overlapping draws have almost no bias. The largest bias is 1% of the true parameter value. Compared to independent draws in 5, overlapping draws tend to have smaller bias for very small n and R/n. Comparing tables 6 and 7, independent and overlapping draws have similar RMSE.

Finally, Figure 6 graphically illustrates the difference in the coverage probabilities between overlapping and independent draws. It plots the difference between the absolute deviation of independent draws coverage from 95% and the absolute deviation of overlapping draws coverage from 95%. A positive value means independent draws performs worse than overlapping draws.

7 Conclusion

We provide an asymptotic theory for simulated GMM and simulated MLE for nonsmooth simulated objective function. The total number of simulations, R, has to increase without bound but can be much smaller than the total number of observations. In this case, the error in the parameter estimates is dominated by the simulation errors. This is a necessary cost of inference when the simulation model is very intensive to compute.

Table 5: Mean bias as % of the true parameter values for independent draws

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	-1	0	0	0	0	0	0	0	0
50	-2	0	0	0	0	0	0	0	0
100	-1	0	0	0	0	0	0	0	0
100	-1	0	0	0	0	0	0	0	0
200	0	0	0	0	0	0	0	0	0
200	-1	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0
800	0	0	0	0	0	0	0	0	0
800	0	0	0	0	0	0	0	0	0

The total number of Monte Carlo repetitions is 1000.

Table 6: RMSE as % of the true parameter values for overlapping draws

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	14	10	9	8	8	8	7	7	7
50	10	8	7	7	7	6	6	5	5
100	9	6	6	6	6	5	5	5	5
100	8	5	4	4	3	3	3	3	3
200	5	5	5	4	4	4	3	4	3
200	5	4	3	3	3	3	2	2	2
400	4	3	3	3	2	2	2	2	2
400	3	3	2	2	2	2	2	2	2
800	3	2	2	2	2	1	2	1	2
800	2	2	1	1	1	1	1	1	1

The total number of Monte Carlo repetitions is 1000.

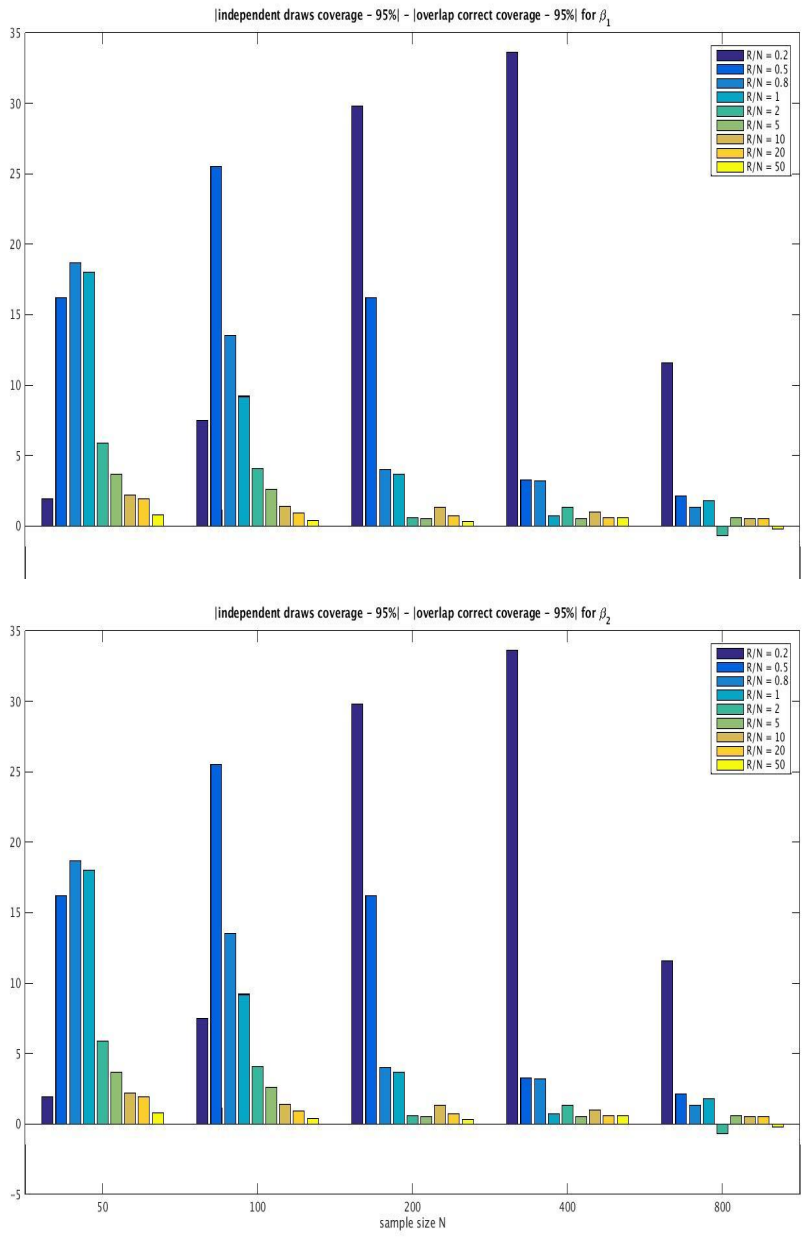

Figure 1: Difference in coverage probabilities between independent and overlapping draws

Table 7: RMSE as % of the true parameter values for independent draws

n,κ	0.2	0.5	0.8	1	2	5	10	20	50
50	10	8	7	8	8	6	7	7	8
50	11	6	6	6	7	6	6	6	6
100	7	6	6	5	5	5	5	5	5
100	7	3	3	3	4	3	3	3	3
200	4	4	4	3	4	3	3	4	3
200	4	2	2	2	3	2	2	2	2
400	3	2	2	2	2	2	2	2	2
400	2	2	2	2	2	2	1	2	2
800	2	2	2	2	1	2	2	1	1
800	1	1	1	1	1	1	1	1	1

The total number of Monte Carlo repetitions is 1000.

References

Andrews, D. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2248–2292. North Holland.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., and H. Hong (2003): "A MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115(2), 293–346.

Duffie, D., and K. J. Singleton (1993): "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica*, 61(4), pp. 929–952.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185(1), 162–181.

Gallant, R., and G. Tauchen (1996): "Which Moments to Match," *Econometric Theory*, 12, 363–390.

Gourieroux, C., and A. Monfort (1996): *Simulation-based econometric methods*. Oxford University Press, USA.

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect inference," *Journal of applied econometrics*, 8(S1).

Ichimura, H., and S. Lee (2010): "Characterization of the asymptotic distribution of semiparametric M-estimators," *Journal of Econometrics*, 159(2), 252–266.

Kristensen, D., and B. Salanié (2013): "Higher order properties of approximate estimators," *CAM Working Papers*.

Laroque, G., and B. Salanie (1989): "Estimation of multi-market fix-price models: An application of pseudo maximum likelihood methods," *Econometrica: Journal of the Econometric Society*, pp. 831–860.

Laroque, G., and B. Salanié (1993): "Simulation-based estimation of models with lagged latent variables," *Journal of Applied Econometrics*, 8(S1), S119–S133.

Lee, D., and K. Song (2015): "Simulated MLE for Discrete Choices using Transformed Simulated Frequencies," *Journal of Econometrics*, 187, 131–153.

Lee, L. (1992): "On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models," *Econometric Theory*, 8, 518–552.

——— (1995): "Asymptotic bias in simulated maximum likelihood estimation of discrete choice models," *Econometric Theory*, 11, 437–483.

Lerman, S., and C. Manski (1981): "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski, and D. McFadden. MIT Press.

McFadden, D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*.

Neumeyer, N. (2004): "A central limit theorem for two-sample U-processes," *Statistics & Probability Letters*, 67, 73–85.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2113–2241. North Holland.

Nolan, D., and D. Pollard (1987): "U-processes:rates of convergence," *The Annals of Statistics*, pp. 780–799.

——— (1988): "Functional limit theorems for U-processes," *The Annals of Probability*, pp. 1291–1298.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Pollard, D. (1984): *Convergence of Stochastic Processes*. Springer Verlag.

Sherman, R. P. (1993): "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61, 123–137.

Smith, A. A. (1993): "Estimating nonlinear time-series models using simulated vector autoregressions," *Journal of Applied Econometrics*, 8(S1).

Stern, S. (1992): "A method for smoothing simulated moments of discrete probabilities in multinomial probit models," *Econometrica*, 60(4), 943–952.

Train, K. (2003): *Discrete choice methods with simulation*. Cambridge Univ Pr.

van der Vaart, A. (1999): *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence and empirical processes.* Springer-Verlag, New York.

A Technical Addendum

Proof of Lemma 1

Proof Consider first the case when the moment condition $q(z, \omega, \theta)$ is univariate, so that $d = 1$. The first statement (a) follows from Theorem 2.5 in Neumeyer (2004). The proof of part (b) resembles Theorem 2.7 in Neumeyer (2004) but does not require $n/(n+R) \to \kappa \in (0,1)$. First define $\tilde{U}_{nR}(\theta) = \frac{\sqrt{m}}{nR}\tilde{S}_{nR}(\theta)$. It follows from part (a) that

$$\sup_{\theta \in \Theta} ||\tilde{U}_{nR}(\theta)|| = O_p\left(\sqrt{\frac{m}{nR}}\right) = o_p(1).$$

Since the following equality follows: $U_{nR}(\theta) = \tilde{U}_{nR}(\theta) + \sqrt{m}(P_n - P)g(\cdot, \theta) + \sqrt{m}(Q_R - Q)h(\cdot, \theta)$, it then only remains to verify the stochastic equicontinuity conditions for the two projection terms:

$$\sup_{d(\theta_1, \theta_2) = o(1)} ||\sqrt{m}(P_n - P)(g(\cdot, \theta_1) - g(\cdot, \theta_2))|| = o_p(1),$$

and

$$\sup_{d(\theta_1, \theta_2) = o(1)} ||\sqrt{m}(Q_R - Q)(h(\cdot, \theta_1) - h(\cdot, \theta_2))|| = o_p(1).$$

This in turn follows from $m \leq n, R$ and the equicontinuity lemma of Pollard (1984), p. 150. Part (c) mimicks Theorem 2.9 in Neumeyer (2004), noting that

$$\frac{1}{nR}S_{nR}(\theta) - g(\theta) = \frac{1}{nR}\tilde{S}_{nR}(\theta) + (P_n - P)g(\cdot, \theta) + (Q_R - Q)h(\cdot, \theta),$$

and invoking part (a) and Theorem 24 of Pollard (1984), p. 25.

When the moment conditions $q(z, \omega, \theta)$ are multivariate, so that $d > 1$, the above arguments apply to each univariate element of the vector moment condition $q(z, \omega, \theta)$. □

LEMMA 3 Let $\hat{\theta} \xrightarrow{p} \theta_0$, where $g(\theta) = 0$ if and only if $\theta = \theta_0$, which is an interior point of the compact Θ. If

i. $\|\hat{g}(\hat{\theta})\|_{W_n} \leq \inf_\theta \|\hat{g}(\theta)\|_{W_n} + o_p(m^{-1/2})$.

ii. $W_n = W + o_p(1)$ where W is positive definite.

iii. $g(\theta)$ is continuously differentiable at θ_0 with a full rank derivative matrix G.

iv. $\sup_{d(\theta,\theta_0)=o(1)} \sqrt{m}\,\|\hat{g}(\theta) - g(\theta) - \hat{g}(\theta_0)\| = o_p(1)$, for $m \equiv n \wedge R$.

v. $\sqrt{m}\,\hat{g}(\theta_0) \xrightarrow{d} N(0, \Sigma)$.

Then the following result holds

$$\sqrt{m}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}). \qquad \blacksquare$$

In particular, Lemma 1.b delivers condition [iv]. Condition [v] is implied by Lemma 1.a because

$$\begin{aligned}
\sqrt{m}\hat{g}(\theta_0) &= \tilde{U}_{nR}(\theta_0) + \sqrt{m}(P_n - P)g(\cdot,\theta_0) + \sqrt{m}(Q_R - Q)h(\cdot,\theta_0) \\
&= \sqrt{m}(P_n - P)g(\cdot,\theta) + \sqrt{m}(Q_R - Q)h(\cdot,\theta_0) + o_p(1) \\
&\xrightarrow{d} N(0, (1 \wedge \kappa)\Sigma_g + (1 \wedge 1/\kappa)\Sigma_h).
\end{aligned}$$

Recall a general result (see for example Theorem 3.2.16 of Van der Vaart and Wellner (1996)), which for completeness is restated as the following lemma.

LEMMA 4

$$\sqrt{m}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1})$$

under the following conditions:

1. $\hat{L}(\hat{\theta}) \geq \sup_{\theta\in\Theta} \hat{L}(\theta) - o_p(\frac{1}{m})$;

2. $\hat{\theta} \xrightarrow{p} \theta_0$;

3. θ_0 is an interior point of Θ;

4. $L(\theta)$ is twice continuously differentiable in an open neighborhood of θ_0 with positive definite Hessian $H(\theta)$;

5. There exists $\hat{D}$ such that $\sqrt{m}\hat{D} \xrightarrow{d} N(0, \Sigma)$; and such that

6. For any $\delta_m \to 0$ and for $\hat{R}(\theta) = \hat{L}(\theta) - L(\theta) - \hat{L}(\theta_0) + L(\theta_0) - \hat{D}'(\theta - \theta_0)$,

$$\sup_{\|\theta-\theta_0\|\leq\delta_m} \frac{m\hat{R}(\theta)}{1 + m\|\theta - \theta_0\|^2} = o_p(1). \tag{8}$$

(If $\hat{\theta}$ is known to be r_m consistent, i.e., $\hat{\theta} - \theta_0 = o_p(1/r_m)$ for $r_m \to \infty$, then Condition 6 only has to hold for $\delta_m = o_p(1/r_m)$.)

LEMMA 5 Let $\{a_m\}$, $\{b_m\}$, and $\{c_m\}$ be sequences of positive numbers that tend to infinity. Suppose

1. $\hat{L}(\hat{\theta}) \geq \hat{L}(\theta_0) - O_p(a_m^{-1})$;

2. $\hat{\theta} \xrightarrow{p} \theta_0$;

3. In a neighborhood of θ_0 there is a $\bar{\kappa} > 0$ such that $L(\theta) \leq L(\theta_0) - \bar{\kappa}\|\theta - \theta_0\|^2$;

4. For every sequence of positive numbers $\{\delta_m\}$ that converges to zero, $\|\theta_m - \theta_0\| < \delta_m$ implies $\left|\hat{L}(\theta_m) - \hat{L}(\theta_0) - L(\theta_m) + L(\theta_0)\right| \leq O_p(\|\theta_m - \theta_0\|/b_m) + o_p(\|\theta_m - \theta_0\|^2) + O_p(1/c_m)$.

then

$$\left\| \hat{\theta} \right\| = O_p\left(\frac{1}{\sqrt{d_m}}\right),$$

where $d_m = \min(a_m, b_m^2, c_m)$.

Proof The proof is a modification of Sherman (1993). Condition 2 implies that there is a sequence of positive numbers $\{\delta_m\}$ that converges to zero slowly enough that $P(\|\hat{\theta} - \theta_0\| \leq \delta_m) \to 1$. When $\|\hat{\theta} - \theta_0\| \leq \delta_m$ we have from Conditions 1 and 2 that

$$\bar{\kappa}\|\hat{\theta}\|^2 - O_p(1/a_m) \leq \hat{L}(\hat{\theta}) - \hat{L}(\theta_0) - L(\hat{\theta}) + L(\theta_0) \leq O_p\Big(\|\hat{\theta}\|/b_m\Big) + o_p\Big(\|\hat{\theta}\|^2\Big) + O_p(1/c_m)$$

whence

$$[\bar{\kappa} + o_p(1)]\,\|\hat{\theta}\|^2 \leq O_p(1/a_m) + O_p\Big(\|\hat{\theta}\|/b_m\Big) + O_p(1/c_m) \leq O_p(1/d_m) + O_p\Big(\|\hat{\theta}\|/\sqrt{d_m}\Big).$$

The remaining arguments then follow exactly from 2. □