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: May 2013

Abstract

In this paper we study the asymptotic distribution of simulation estimators where the same set of simulation 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 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 compliment 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 StateUniversity, 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 acknowledge support by the National Science Foundation (SES 1024504) and SIEPR.

[a] Yale University
[b] The Pennsylvania State University
[c,d] Stanford University

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), and Gourieroux and Monfort (1996) among others. A general asymptotic approach is provided by Pakes and Pollard (1989) who consider generalized method of simulated moment estimators when a fixed number of independent simulations are used for each of the independent observations. A recent insightful paper by Lee and Song (2009) also developed results for a class of simulated maximum likelihood-like estimators. In practice, however, researchers sometimes use the same set of simulation draws for all the observations in the dataset.

The properties of simulation based estimators using overlapping 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 Chen, Linton, and Van Keilegom (2003) and Ichimura and Lee (2010) by combining the results in Neumeyer (2004) with empirical process theories (Pakes and Pollard (1989), Andrews (1994) and Newey and McFadden (1994)).

Whether using overlapping simulations for all observations presents an improvement in computational efficiency depends on the specific model. Generating the random numbers is easy but computing the moment condition or the likelihood function is typically difficult. If the observations y_i and x_i, $i = 1, \ldots, n$, where n is the sample size, are continuous and are different for each observation, then one might not save at all at computation if the same simulations are used for all observations because the total account of computations is still of the order of $n \times R$, where R is the total number of simulation draws. However, in practice, researchers often extrapolate the simulated moment conditions or the likelihood function over a range of the covariate variables (Gallant, Hong, and Khwaja (2011)). The

literature of dynamic discrete choice models, for example, also extrapolates the solution of the dynamic program over the space of state variables (e.g. Keane and Wolpin (1994)). Strictly, these methods that combine nonparametric smoothing with simulation estimation do not fall squarely into either the independent case in Pakes and Pollard (1989) or the overlapping simulation case studied here. Yet the results reported here together with the results for independent simulations in the literature should provide some guidance to how future work can be developed for the intermediate case of extrapolated simulation draws. Allowing for nonsmooth simulators may indeed improve computation time over smooth simulators. For example, the Stern (1992) decomposition simulator, while smooth and unbiased, requires repeated calculations of eigenvalues and is computationally prohibitive. Laffont, Ossard, and Vuong (1995) and Kristensen and Salanié (2010) develops bias reduction techniques for simulation estimators.

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 y_i are the dependent variables and x_i are the covariates or regressors. We are concerned about estimating an unknown parameter θ.

The method of moments estimator is based on a set of moment conditions $g(z_i, \theta)$ such that $g(\theta) \equiv Pg(z_i, \theta)$ is zero if and only if $\theta = \theta_0$ where θ_0 is construed as the true parameter value. In the above $Pg(z_i, \theta)$ denotes expectation with respect to the sample observation of z_i. In models where the moment $g(z_i, \theta)$ can not be analytically evaluated, it can often be approximated using simulations. Let $\omega_r,\ r = 1, \ldots, R$, be a set of simulation draws, and 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:

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

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

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

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

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

The method of simulated moments (MSM) estimator with overlapping simulated draws is defined with the usual quadratic norm as in Pakes and Pollard (1989)

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

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,

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

While $g(z_i;\theta)$ is typically a smooth function of z_i and θ, $\hat{g}(z_i;\theta)$ often times is not. In these situations it is difficult to obtain the exact optimum for both MSM and MSL, and these definitions will be relaxed below to only require that the MSM and MSL estimators obtain "near-optimum" of the respective objective functions.

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 $R >> \sqrt{n}$ 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 $R << n$, 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), and is strengthened by Lee and Song (2009) to $\sqrt{R}\log R/n \to \infty$ for nonsmooth maximum likelihood like estimators.

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):

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

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.

ASSUMPTION 1 The following three classes of functions

$$\begin{aligned} \mathcal{F} &= \{q(z_i, w_r, \theta),\ \theta \in \Theta\}, \\ \mathcal{QF} &= \{g(z_i, \theta),\ \theta \in \Theta\} \ = \ \{Qf,\ f \in \mathcal{F}\}, \\ \mathcal{PF} &= \{h(w_r, \theta),\ \theta \in \Theta\} \ = \ \{Pf,\ f \in \mathcal{F}\}, \end{aligned}$$

are Euclidean classes for the L_1 norm as defined in Pakes and Pollard (1989) whose envelope functions, denoted respectively by F, QF and PF, have at least two moments; cf. Lemma 25 (p. 27), Lemma 36 (p. 34), and Theorem 37 (p. 34) of Pollard (1984). The function $q(y_i, \omega_r, \theta)$ is mean square continuous at θ_0.

This assumption is satisfied by most known functions, except for very large classes of functions such as the example in 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 $1\left(z_i'\theta + w_r \geq 0\right)$, 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 $\Lambda\left(x_i'\beta\right)$ over the disribution 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 $\phi\left(\cdot;\mu,\Sigma\right)$ normal density function with mean μ and variance matrix Σ, $\int \Lambda\left(x_i'\beta\right)\phi\left(\beta; v_i'\theta_1, \theta_2'\theta_2\right)d\beta$. 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 $\Lambda\left(\cdot\right)$ is a monotone function, this function is Euclidean according to Lemma 2.6.18 in Van der Vaart and Wellner (1996).

Under assumption 1, which implies that the class $\mathcal{F}$ and its projections $\mathcal{QF}$ and $\mathcal{PF}$ are VC-classes (see Neumeyer (2004), p. 79), the following lemma is analogous to Theorems 2.5, 2.7 and 2.9 of Neumeyer (2004).

LEMMA 1 Under Assumption 1 the following statements hold:

a. Define

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

then

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

where

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

b. Define

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

then

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

c. Further,

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

Proof 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 $U_{nR}(\theta)=\tilde{U}_{nR}(\theta)+\sqrt{m}(P_n-P)g(z_i,\theta)+\sqrt{m}(Q_R-Q)h(\omega_r,\theta)$, where we define $P_nf(z_i)\equiv\frac{1}{n}\sum_{i=1}^{n}f(z_i)$ and $Q_Rg(\omega_r)=\frac{1}{R}\sum_{r=1}^{R}g(\omega_r)$, it then only remains to verify the stochastic equicontinuity conditions for the two projection terms:

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

and

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

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

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

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

Lemma 1 will be applied in combination with the following restatement of a version of Theorem 7.2 of Newey and McFadden (1994) and Theorem 3.3 of Pakes and Pollard (1989).

THEOREM 1 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. $\sup_{d(\theta,\theta_0)=o(1)}\sqrt{m}\,\|\hat{g}(\theta)-g(\theta)-\hat{g}(\theta_0)\|_W=o_p(1)$.

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

Then the following result holds

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

Remark: The original Theorem 3.3 of Pakes and Pollard (1989) uses the Euclidean norm to define the GMM objective function, which amounts to using an identity weighting matrix $W_n \equiv I$. However, generalizing their proof arguments to a general random W_n is straightforward. First, note that their Theorem 3.3 is isophormic to using a fixed positive definite weighting matrix W to define the norm. This is because if one uses the square root A of W (such that $A'A = W$) to form a linear combination of the original moment conditions $\hat{g}(\theta)$, the moment condition in Theorem 3.3 can be reinterpreted as $A\hat{g}(\theta)$, and exactly the same arguments in the proof goes through, with the matrixes Γ and V in P&P being AG and $A\Sigma A'$.

Second, a close inspection of the proof of Theorem 3.3 in P&P shows that their Condition (i) is only used to the extent of requiring both

$$\|\hat{g}(\hat{\theta})\|_W \leq \|\hat{g}(\theta_0)\|_W + o_p(m^{-1/2}), \quad \text{and} \quad \|\hat{g}(\hat{\theta})\|_W \leq \|\hat{g}(\theta^*)\|_W + o_p(m^{-1/2}),$$

where θ^* is the minimizer of the quadratic approximation to the objective function $\|\hat{g}(\theta)\|_W$ defined in p. 1042 of P&P. These will follow from Condition [i] if:

$$\|\hat{g}(\hat{\theta})\|_{W_n} = \|\hat{g}(\hat{\theta})\|_W + o_p(m^{-1/2}), \quad \|\hat{g}(\theta_0)\|_{W_n} = \|\hat{g}(\theta_0)\|_W + o_p(m^{-1/2})$$

and

$$\|\hat{g}(\theta^*)\|_{W_n} = \|\hat{g}(\theta^*)\|_W + o_p(m^{-1/2}),$$

all of which follow in turn from combining Conditions [ii], [iv] and [v].

Consistency, under the conditions stated in Corollary 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 Theorem 1.

COROLLARY 1 Given Assumption 1, $\hat{\theta} \xrightarrow{p} \theta_0$ 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; and (c)

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

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 the conclusion of Theorem 1 holds under Assumption 1, with $\Sigma = (1 \wedge \kappa)\Sigma_g + (1 \wedge 1/\kappa)\Sigma_h$, where $\Sigma_g = \text{Var}(g(z_i, \theta_0))$ and $\Sigma_h = \text{Var}(h(\omega_r, \theta_0))$. ■.

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

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

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 a step size parameter

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

A sufficient condition for $\hat{G}(\hat{\theta}) \xrightarrow{p} G(\theta_0)$ is that both $\delta \to 0$ and $\sqrt{m}\delta \to \infty$. 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 $\delta \to 0$ and $\hat{\theta} \xrightarrow{p} \theta_0$. (In practice one would not let δ fall below the value that is optimal for a machine's precision.) Hong, Mahajan, and Nekipelov (2009) shows that consistency holds under much weaker sufficient conditions. Extending these results to two sample U-statistics is beyond the scope of paper. 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. Note that in $\hat{\Sigma}$ above, 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 $1/R$.

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 Euclidean property is required of the class of functions $q(z, \cdot, \theta)$ of ω viewed as indexed by both θ and z. Frequently $g(z, \theta)$ is a conditional likelihood in the form of $g(y \,|\, x, \theta)$ 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 2 The class of functions indexed by both θ and z: $\mathcal{L} = \{\, q(z, \cdot, \theta) \,:\, z \in Z,\ \theta \in \Theta\}$ has polynomial degree of covering numbers and uniformly bounded envelope function L.

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

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

LEMMA 2 Under Assumptions 1, 2, and 3, $\hat{L}(\theta) - \hat{L}(\theta_0)$ converges to $L(\theta) - L(\theta_0)$ as $m \to \infty$ uniformly over Θ.

Proof Consider the decomposition

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

where

$$\begin{aligned} A &= (P_n - P)[\log g(z,\theta) - \log g(z,\theta_0)] \\ B &= P_n[\log \hat{g}(z,\theta) - \log \hat{g}(z,\theta_0) - \log g(z,\theta) - \log g(z,\theta_0)]. \end{aligned} \tag{1}$$

First, we show that A converges uniformly to 0 in probability. By Assumption 1, the monotonicity of log transformation and Lemma 2.6.18 (v) and (viii) in Van der Vaart and Wellner (1996), $\log \circ \mathcal{QF} - \log g(y,\theta_0)$ is VC-subgraph. Furthermore, by Assumptions 1, 3, and concavity of the log transformation, $\log QF - \log g(y,\theta_0)$ is an envelope function with bounded first moment for $\log \circ \mathcal{QF} - \log \circ g(y,\theta_0)$. Hence, by Lemma 19.13 and Lemma 19.15 of van der Vaart (1999), $\sup_{\theta\in\Theta} |A| \xrightarrow{p} 0$ as $n \to \infty$.

Second, we show that B converges uniformly to 0 in probability as $R \to \infty$. By Taylor's theorem and Assumption 3,

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

Moreover, by Assumption 2 and Lemma 19.13 and 19.15 of van der Vaart (1999), as $R \to \infty$,

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

Therefore, B converges uniformly to 0 as $R \to \infty$. The lemma then follows from the triangle inequality. □

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

COROLLARY 2 Under Assumptions 1, 2, and 3, 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 $\hat{\theta} - \theta_0 \xrightarrow{p} 0$.

As pointed out in Pollard (1984), the requirement that $\sup_{\theta\in\Theta} |\hat{L}(\theta) - L(\theta)| = o_p(1)$ implies, but can be weakened to

$$\limsup_{n\to\infty} P\left\{\sup_{\theta\in\Theta}\left[\hat{L}(\theta) - L(\theta)\right] \geq \epsilon\right\} = 0$$

for all $\epsilon > 0$.

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

THEOREM 2

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

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 $\delta \to 0$ and for $\hat{R}(\theta) = \hat{L}(\theta) - L(\theta) - \hat{L}(\theta_0) + L(\theta_0) - \hat{D}'(\theta - \theta_0)$,

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

(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 $\delta = o_p(1/r_m)$.)

The following analysis consists of verifying the conditions in the above general theorem. The finite sample likelihood, without simulation, is required to satisfy the stochastic differentiability condition as required in the high level assumption, but does not need to be pointwise differentiable.

ASSUMPTION 4 There exists a mean zero random variable $D_0(z_i)$ with finite variance such that for any $\delta \to 0$ we have

$$\lim_{n\to\infty} \sup_{\|\theta-\theta_0\|\leq\delta} \frac{nR_n(\theta)}{1+n\|\theta-\theta_0\|^2} = o_p(1)$$

for

$$R_n(\theta) \equiv (P_n - P)(\log g(z,\theta) - \log g(z,\theta_0)) - \hat{D}_0'(\hat{\theta} - \theta_0),$$

where

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

An primitive condition for this assumption is given in Lemma 3.2.19, p. 302, of Van der Vaart and Wellner (1996).

To account for the simulation error we need an intermediate step which is a modification of Theorem 1 of Sherman (1993).

THEOREM 3 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 $\kappa > 0$ such that $L(\theta) \leq L(\theta_0) - \kappa\|\theta\|^2$;
4. For every sequence of positive numbers $\{\delta_m\}$ that converges to zero, $\|\theta_m - \theta_0\| < \delta_m$ implies $\left|\hat{L}(\theta_m) - \hat{L}(\theta_0) - L(\theta_m) + L(\theta_0)\right| \leq O_p(\|\theta_m\|/b_m) + o_p(\|\theta_m\|^2) + O_p(1/c_m)$.

then

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

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

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

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

whence

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

Letting $\hat{W}$ denote an $O_p(1/\sqrt{d_m})$ random variable, the expression above implies that

$$\frac{1}{2}\kappa\|\hat{\theta}\|^2 - \hat{W}\|\hat{\theta}\| \leq O_p(1/d_m)$$

on an event that has probability one in the limit. Completing the square gives

$$\frac{1}{2}\kappa\left(\|\hat{\theta}\| - W/\kappa\right)^2 \leq O_p\left(\frac{1}{d_m}\right) + \frac{\hat{W}^2}{2\kappa} = O_p\left(\frac{1}{d_m}\right)$$

whence $\sqrt{d_m}\left\|\hat{\theta}\right\| \leq \sqrt{d_m}\hat{W} + O_p(1) = O_p(1)$. □

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

ASSUMPTION 5 For all simulation lengths R and all parameters θ, both $g\left(z_i;\theta\right)$ and $Q_n q\left(z_i,\cdot;\theta\right)$ are proper (possibly conditional) density functions.

We also need to regulate the amount of irregularity that can be allowed by the simulation function $q\left(z,\omega,\theta\right)$. In particular, it allows for $q\left(z,\omega,\theta\right)$ to be an indicator function.

ASSUMPTION 6 $\sup_{\|\theta-\theta_0\|\leq\delta, z\in Z} \mathrm{Var}_\omega\left(\frac{q(z,\omega,\theta)}{g(z,\theta)} - \frac{q(z,\omega,\theta_0)}{g(z,\theta)}\right) = O\left(\delta\right).$

ASSUMPTION 7 Define $\psi(\omega,\theta,\theta_0)=\int \frac{q(z,\omega,\theta)}{g(z,\theta)} g(z,\theta_0)\,dz$. There exists a mean zero random variable $D_1(\omega_r)$ with finite variance such that for $\hat{D}_1=\frac{1}{R}\sum_{r=1}^{R} D_1(\omega_r)$,

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

THEOREM 4 Under Assumptions 1, 2, 3, 4, 5, 6, and 7 and Conditions 1, 2, and 4 of Theorem 2, the conclusion of Theorem 2 holds with $\hat{D}=P_nD_0(z_i)+Q_RD_1(\omega_r)$ 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 Corollary 2. Consider again the decomposition given by Equation (1). Because of the linearity structure of Conditions (5) and (6) of Theorem 2, it suffices to verify them separately for the terms A and B.

It follows immediately from Assumption 4 that Conditions (5) and (6) of Theorem 2 hold for the first term A because $n\geq m$. Next we verify them for B.

Decompose B further into $B=B_1+B_2+B_3$, where

$$\begin{aligned}
B_1 &= P_n\left[\frac{1}{g_\theta}(\hat{g}_\theta-g_\theta)-\frac{1}{g_0}(\hat{g}_0-g_0)\right]\\
B_2 &= -\frac{1}{2}P_n\left[\frac{1}{g_\theta^2}(\hat{g}_\theta-g_\theta)^2-\frac{1}{g_0^2}(\hat{g}_0-g_0)^2\right]\\
B_3 &= \frac{1}{3}P_n\left[\frac{1}{\bar{g}_\theta^3}(\hat{g}_\theta-g_\theta)^3-\frac{1}{\bar{g}_0^3}(\hat{g}_0-g_0)^3\right].
\end{aligned}$$

In the above $\bar{g}_\theta$ and $\bar{g}_0$ are mean values, dependent on z_i, between $[g(z,\theta),\hat{g}(z,\theta)]$ and $[g(z,\theta_0),\hat{g}(z,\theta_0)]$ respectively. By Assumption 3,

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

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

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

This allows us to invoke Theorem 3, 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 $\|\hat{\theta}-\theta_0\| = o_p\left(1/\log R\right)$:

$$\sup_{\|\theta-\theta_0\|\ll(\log R)^{-1}} |B_2| = o_p\left(\frac{1}{R}\right). \tag{2}$$

To show (2), first note that

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

where

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

and

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

It follows again from Theorem 2.14.1 that

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

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

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

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

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

The second term goes to zero for the same reason as in Pollard. The first also goes to zero since $\log R - \frac{1}{\delta}\to-\infty$. Thus we have shown that $B_{22} = o_p\left(\frac{1}{\sqrt{R}}\right)$ uniformly in $\theta-\theta_0\leq\delta$ and $z\in Z$, and consequently (2) holds. By considering $n\gg R$, $n\ll R$ and $n\approx R$ separately, (2) also implies that for some $\alpha>0$:

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

It remains to investigate $B_1 = P_n Q_R f(z, \omega, \theta)$, which, using Assumption 5, can be written

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

where

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

$B_0 = (Q_R - Q)(\psi(\omega, \theta) - \psi(\omega, \theta_0))$, and $\psi(\omega, \theta) = \int \frac{q(z,\omega,\theta)}{g(z,\theta)} g(z, \theta_0)\, dz$ upon noting that the other projection term is 0 since $Q\frac{q(\cdot,z,\theta)}{g(z,\theta)} = 1$ identically. The proof of Theorem 2.5 (pp. 83) of Neumeyer (2004) shows that, since

$$Q \times P\left[\sup_{\|\theta-\theta_0\|=o(1)} f(z, \omega, \theta)^2\right] = o(1) \quad \Longrightarrow \quad \frac{1}{nR} S_{nR}\left(\tilde{f}_\theta\right) = o_p\left(\frac{1}{\sqrt{nR}}\right) = o_p\left(\frac{1}{m}\right).$$

Finally, B_0 is handled by Assumption 7. □

Assumption 7 can be further simplified when the true likelihood $g(z, \theta)$ is twice continuously differentiable (with bounded derivatives for simplicity). In this case

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

To see this, note that

$$\begin{aligned}
&(Q_R - Q)(\psi(\omega, \theta) - \psi(\omega, \theta_0)) \\
&\quad = \quad P\left[\frac{1}{g_\theta} - \frac{1}{g_0}\right](\hat{g}_0 - g_0) \\
&\quad + \quad P\frac{1}{g_0}(\hat{g}_\theta - g_\theta - \hat{g}_0 + g_0) \\
&\quad + \quad P\left(\frac{1}{g_\theta} - \frac{1}{g_0}\right)(\hat{g}_\theta - g_\theta - \hat{g}_0 + g_0).
\end{aligned}$$

The second line is zero because of assumption 5. The third line can be bounded by

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

using the same arguments that handle the B_{22} in the proof. Finally, the first line becomes

$$P\left[\frac{1}{g_\theta} - \frac{1}{g_0}\right](\hat{g}_0 - g_0) = (Q_R - Q) D_1(w_r)(\theta - \theta_0) + \tilde{R}(\theta),$$

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

4.1 MSL Variance Estimation

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

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

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

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

Let

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

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 $\hat{\Sigma} = \Sigma + o_p\left(1\right)$.

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

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

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

the above the objective function can be either GMM:

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

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 1, which leads to Theorem 1, 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 $\alpha > 0$:

$$\int |h|^{\alpha} |p_n(h) - p_{\infty}(h)| dh \xrightarrow{p} 0, \quad \text{where} \quad p_{\infty}(h) = \sqrt{\frac{|\det(J_0)|}{(2\pi)^{\dim\theta}}} \exp\left(-\frac{1}{2} h' J_0 h\right). \tag{3}$$

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 5 Under Assumption 1 for the GMM model, and under Assumptions 1 to 7, Conditions 1, 2, 4 of Theorem 2 for the MLE model, (3) 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 1, which leads to Theorem 1, 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 2 holds for $\delta = o(1)$, the original proof shows (3) over three areas of integration separately, $\{|h| \leq \sqrt{m}\delta\}$ and $\{|h| \geq \delta\sqrt{m}\}$. When Condition 6 in Theorem 2 only holds for $\delta = a_m = (\log m)^{-d}$, 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):

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

By arguments that handle the term B in the proof of Theorem 4, 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 written as

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

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 $W = asym\ \mathrm{Var}\left(\sqrt{m}\hat{g}(\theta_0)\right)$, 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 $W \neq asym\ \mathrm{Var}\left(\sqrt{m}\hat{g}(\theta_0)\right)$, 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 2.

6 Monte Carlo Simulations

In this section we report the results from a set of Monte Carlo simulations from a univariate Probit model to illustrate the finite sample properties of the asymptotic distributions derived in this paper. The true data generating process is specified to be:

$$y_i = \mathbf{1}\{\alpha_0 + \tilde{x}_i\beta_0 + \epsilon_i \geq 0\}, \quad \epsilon_i \perp\!\!\!\perp \tilde{x}_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0,1).$$

Define $z_i = (y_i, x_i)$ and $x_i = (1, \tilde{x}_i)$, $\theta := (\alpha, \beta)'$. In the earlier notation the likelihood function is $g(z_i, \theta) = \Phi(x_i'\theta_0)^{y_i}(1 - \Phi(x_i'\theta_0))^{1-y_i}$ and the true parameter θ_0 maximizes

$$L(\theta) = \mathbb{E}_z \log g(z_i, \theta) = \mathbb{E}_z[y_i \log \Phi(x_i'\theta) + (1 - y_i) \log(1 - \Phi(x_i'\theta))].$$

The likelihood function is simulated by $\hat{g}(z_i, \theta) = \frac{1}{R}\sum_{r=1}^{R} q(w_r, z_i, \theta)$, where

$$q(w_r, z_i, \theta) = \mathbf{1}\{x_i'\theta + w_r \geq 0\}^{y_i}(1 - \mathbf{1}\{x_i'\theta + w_r \geq 0\})^{1-y_i},$$

and $w_r \overset{\text{iid}}{\sim} N(0, 1)$. Note that $\mathbb{E}_w \hat{g}(z_i, \theta) = g(z_i, \theta)$. The simulated maximum likelihood estimator maximizes

$$\hat{L}_{nR}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{g}(z_i, \theta)$$

and is computed using the simulated annealing routine in Matlab's global optimization toolbox. The starting value for optimization is taken to be the OLS estimates. The numerical results are not sensitive to the choice of the starting values when the temperature parameter in the simulated annealing routine is reduced sufficiently slowly. Obviously this simple example can be estimated by the probit command in Stata. The goal of this section is to illustrate the finite properties of the simulated maximum likelihood estimator when we are agnostic about the normal distribution function and density function.

We compute an estimate of the asymptotic variances using the empirical analog of Theorem 2:

$$\sqrt{m}\left(\hat{\theta}_{SMLE} - \theta\right) \overset{A}{\sim} N\left(0, \hat{H}^{-1}\left((1 \wedge \kappa)\hat{\Sigma}_0 + (1 \wedge 1/\kappa)\hat{\Sigma}_1\right)\hat{H}^{-1}\right).$$

In the above $\kappa = R/n$, $m = \min(R, n)$.

While analytical derivatives can be easily computed in this example, in practice, the analytical derivatives of the likelihood function is usually unknown. In our baseline results, we are agnostic about the analytical derivatives and estimate the asymptotic variance using numerical differentiation:

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_0(z_i)\hat{D}_0(z_i)', \quad \hat{H} = -\hat{\Sigma}_0,$$

$$\hat{\Sigma}_1 = \frac{1}{R}\sum_{r=1}^{R} \hat{D}_1(w_r)\hat{D}_1(w_r)', \quad \hat{D}_1(w_r) := -\frac{1}{n}\sum_{i=n}^{N} \frac{q(w_r, z_i, \theta_{SMLE})}{\hat{g}(z_i, \theta_{SMLE})}\hat{D}_0(z_i).$$

In the above, for

$$\hat{g}(z_i, \theta) \;=\; (1 - \hat{\Phi}_R(-x_i'\theta))^{y_i} (\hat{\Phi}_R(-x_i'\theta))^{1-y_i}$$

and an estimate of the derivative of $\hat{g}(z_i, \theta^{SMLE})$, denoted as $\nabla \hat{g}(z_i, \theta^{SMLE})$, we define:

$$\hat{D}_0(z_i) := \frac{1}{\hat{g}(z_i, \theta^{SMLE})} \nabla \hat{g}(z_i, \theta^{SMLE}).$$

We use the index structure of $g\left(z_i, \theta\right)$ and numerical differentiation to obtain $\nabla \hat{g}(z_i, \theta^{SMLE})$. For this purpose we note that

$$\begin{aligned}
\frac{\partial g(z_i, \theta)}{\partial \theta} &= \left(-\frac{\partial}{\partial \theta} \Phi_R(-x_i'\theta) \right)^{y_i} \left(\frac{\partial}{\partial \theta} \Phi_R(-x_i'\theta) \right)^{1-y_i} \\
&= x_i \left[\left(\frac{\partial}{\partial w} \Phi_R(w) \Big|_{w=-x_i'\theta} \right)^{y_i} \left(-\frac{\partial}{\partial w} \Phi_R(w) \Big|_{w=-x_i'\theta} \right)^{1-y_i} \right] \\
&= (-1)^{1-y_i} x_i \, \frac{\partial}{\partial w} \Phi_R(w) \Big|_{w=-x_i'\theta}.
\end{aligned}$$

Therefore we use:

$$\nabla \hat{g}(z_i, \theta^{SMLE}) = (-1)^{1-y_i} x_i \, \nabla \hat{\Phi}_R(w) \Big|_{w=-x_i'\theta},$$

where we use a first order two-sided formula to define:

$$\begin{aligned}
\nabla \hat{\Phi}_R(w) = \frac{1}{R} \sum_{r=1}^{R} \frac{\mathbf{1}\{w_r \leq w + \epsilon\} - \mathbf{1}\{w_r \leq w - \epsilon\}}{2\epsilon} &= \frac{1}{R} \sum_{r=1}^{R} \frac{1}{2\epsilon} \mathbf{1} \left\{ \frac{|w_r - w|}{\epsilon} \leq 1 \right\} \\
&= \frac{1}{2\epsilon} \frac{\#\{w - \epsilon \leq w_r \leq w + \epsilon\}}{R},
\end{aligned}$$

where ϵ is a step size parameter. Hong, Mahajan, and Nekipelov (2009) showed that a wide range of step size parameter allows for consistent estimation of the analytical derivatives. In the simulation, we experiment with a range of the step size parameter, $\epsilon = R^{-\alpha}$, where α ranges in $\left[\frac{1}{2} \; \frac{1}{3} \; \frac{1}{4} \; \frac{1}{8} \; \frac{1}{10} \; \frac{1}{15}\right]$. It turns out that the largest step size produces the smallest mean square error between the numerical variance and the analytic variance for small and moderate ranges of n and R. Therefore we use $\alpha = 1/15$ in the results reported in the following tables. Empirically, choosing an optimal step size for numerical gradient calculation can be difficult and depends on knowledge of the underlying function to be simulated. Without this knowledge, we recommend using a rule of thumb of the form $Cn^{-\alpha}$ for $\alpha < 1/2$ for the

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

"n \ R/n"	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.9346	0.9420	0.9448	0.9430	0.9468	0.9518	0.9550	0.9580	0.9582	0.9598
	0.9950	0.9862	0.9804	0.9792	0.9716	0.9708	0.9704	0.9718	0.9722	0.9732
100	0.9298	0.9366	0.9416	0.9412	0.9432	0.945	0.9478	0.9442	0.9502	0.9514
	0.9846	0.9634	0.9618	0.9608	0.9592	0.9578	0.9616	0.9608	0.9604	0.9592
200	0.9396	0.9466	0.9430	0.9450	0.9496	0.9462	0.9468	0.9484	0.9476	0.9504
	0.969	0.9538	0.9548	0.956	0.9528	0.9556	0.9542	0.9546	0.9536	0.954
400	0.9470	0.9492	0.9472	0.9468	0.9482	0.9478	0.9492	0.9502	0.9512	0.9484
	0.9532	0.9498	0.9544	0.954	0.9536	0.9558	0.9546	0.9546	0.957	0.9574
800	0.9482	0.9486	0.9410	0.9510	0.9476	0.9486	0.9462	0.9480	0.9428	0.9466
	0.9538	0.9494	0.949	0.9516	0.9532	0.949	0.9522	0.953	0.9492	0.9516

The total number of Monte Carlo simulations is 5000.

step size choice, where we have chosen $\alpha = 1/15$ in this simulation example. We conjecture that it is possible, but beyond the scope of this paper, to use cross-validation methods for choosing the numerical differentiation step size parameter.

For comparison, we also provide the empirical coverages when the analytical derivatives is used to compute $\hat{H}$ and $\hat{\Sigma}_0$. Here

$$\hat{H} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\phi\left(x_i'\hat{\theta}\right)^2}{\Phi\left(x_i'\theta\right)\left(1-\Phi\left(x_i'\theta\right)\right)} x_i x_i', \qquad \hat{\Sigma}_0 = -\hat{H},$$

and

$$\hat{\Sigma}_1 = \frac{1}{R}\sum_{r=1}^{R} \hat{D}_1(w_r)\hat{D}_1(w_r)' \qquad \hat{D}_1\left(w_r\right) = -\frac{1}{n}\sum_{i=1}^{n} \frac{q\left(w_r, z_i, \hat{\theta}\right)}{\hat{g}\left(z_i, \hat{\theta}\right)} \frac{\left(y_i - \Phi\left(x_i'\hat{\theta}\right)\right)\phi\left(x_i'\theta\right)x_i}{\Phi\left(x_i'\hat{\theta}\right)\left(1-\Phi\left(x_i'\theta\right)\right)}.$$

Table 1 reports the empirical coverage of the 95% confidential interval constructed from the estimate of the asymptotic distribution using numerical derivatives, over 5000 Monte Carlo simulations. 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

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

"n \ R/n"	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.6624	0.7736	0.8240	0.8486	0.8992	0.9348	0.9468	0.9534	0.9566	0.9592
	0.9676	0.9582	0.9590	0.9606	0.9628	0.9668	0.9692	0.9718	0.9716	0.9728
100	0.6122	0.7528	0.8192	0.8368	0.8844	0.9224	0.9362	0.9402	0.9478	0.9502
	0.9276	0.9366	0.9472	0.9476	0.9548	0.9556	0.9596	0.9602	0.9602	0.959
200	0.5948	0.7548	0.8178	0.8412	0.8914	0.9258	0.9354	0.9434	0.9456	0.9498
	0.9256	0.935	0.9448	0.9472	0.948	0.953	0.9536	0.9544	0.9536	0.954
400	0.5710	0.7382	0.8168	0.8314	0.8870	0.9260	0.9354	0.9454	0.9492	0.948
	0.9206	0.9334	0.9476	0.9476	0.9504	0.9546	0.9544	0.9542	0.957	0.9572
800	0.5800	0.7346	0.8022	0.8268	0.8848	0.9246	0.9346	0.9428	0.9408	0.9456
	0.929	0.9376	0.9436	0.9462	0.9502	0.9482	0.9508	0.9526	0.9492	0.9514

The total number of Monte Carlo simulations is 5000.

correspond to the intercept and the slope coefficient, respectively. The results show that 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 $R/n = 5$.

Table 3 and 4 report the counterparts of Table 1 and 2 when analytical derivatives are used instead to compute the asymptotic variances. In Table 3, we see that using analytical derivatives do not necessarity give a more accurate coverage than using numerical derivatives. The results in Table 4 is similar to that in Table 2: ignoring variances due to simulation when R is smaller than n can lead to significant errors in the confidence interval.

Table 3: Empirical coverage frequency for 95% confidence interval, analytical derivatives

"n \ R/n"	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.9962	0.9554	0.9570	0.9570	0.9646	0.9650	0.9616	0.9590	0.9568	0.9554
	0.9958	0.9790	0.9736	0.9738	0.9722	0.9602	0.9594	0.9592	0.9602	0.9600
100	0.9408	0.948	0.9584	0.9564	0.9604	0.9526	0.9512	0.9484	0.951	0.9506
	0.9842	0.9680	0.9648	0.9644	0.9548	0.9568	0.9584	0.9566	0.9592	0.9584
200	0.9334	0.9516	0.9570	0.9574	0.9602	0.9506	0.9468	0.9506	0.9494	0.9494
	0.976	0.964	0.9556	0.955	0.95	0.9524	0.9506	0.953	0.9516	0.9524
400	0.9406	0.9510	0.9570	0.9562	0.9492	0.9496	0.9492	0.9514	0.9506	0.9496
	0.9646	0.9516	0.9538	0.9534	0.9516	0.9520	0.9534	0.9522	0.9562	0.9560
800	0.9416	0.9458	0.9456	0.9560	0.9504	0.9500	0.9464	0.9478	0.9432	0.9476
	0.957	0.9482	0.9462	0.9518	0.9502	0.9488	0.9532	0.9536	0.9516	0.9518

The total number of Monte Carlo simulations is 5000.

Table 4: False empirical coverage frequency for 95% confidence interval, analytical derivatives

"n \ R/n"	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.6708	0.7892	0.8376	0.8580	0.9016	0.9354	0.9456	0.9502	0.9536	0.9536
	0.9194	0.9314	0.9362	0.9412	0.9496	0.9516	0.9552	0.9574	0.9598	0.9598
100	0.6204	0.7588	0.8222	0.8404	0.8926	0.9274	0.9418	0.9424	0.9486	0.949
	0.9114	0.9252	0.937	0.938	0.9418	0.9536	0.9562	0.956	0.959	0.9584
200	0.6032	0.7592	0.8190	0.8426	0.8944	0.9272	0.9358	0.942	0.9462	0.9486
	0.913	0.9278	0.9346	0.9412	0.944	0.951	0.9496	0.9522	0.9512	0.9522
400	0.5728	0.7382	0.8164	0.8312	0.8892	0.9260	0.9390	0.946	0.9482	0.949
	0.9146	0.9362	0.9456	0.9442	0.9476	0.951	0.9528	0.952	0.9562	0.956
800	0.5816	0.7356	0.8056	0.8260	0.8850	0.9252	0.9344	0.9426	0.9418	0.9468
	0.9222	0.9344	0.9392	0.946	0.9484	0.9476	0.953	0.9534	0.9514	0.9518

The total number of Monte Carlo simulations is 5000.

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.

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.

Gallant, A., H. Hong, and A. Khwaja (2011): "Bayesian estimation of a dynamic game with endogenous, partially observed, serially correlated state," Working Paper, Duke University, Stanford University, Yale University.

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

Hong, H., A. Mahajan, and D. Nekipelov (2009): "Statistical Properties of Numerical Derivatives," working paper, UC Berkeley and Stanford University.

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

Keane, M., and K. Wolpin (1994): "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence," *The Review of Economics and Statistics*, pp. 648–672.

Kristensen, D., and B. Salanié (2010): "Higher order improvements for approximate estimators," *CAM Working Papers*.

Laffont, J., H. Ossard, and Q. Vuong (1995): "Econometrics of first-price auctions," *Econometrica: Journal of the Econometric Society*, pp. 953–980.

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 (2009): "Simulated MLE for Discrete Choices using Transformed Simulated Frequencies1," *Manuscript, Department of Economics, University of Pennsylvania*.

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.

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.