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: August 2015

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 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] 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), 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. A recent insightful paper by Lee and Song (2015) 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.

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, where all the same R total number of simulations are used for each observation. The properties of simulation based estimators using overlapping and independent simulation draws are studied by Lee (1992), Lee (1995) and Kristensen and Salanié (2010) 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 in 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 easy but computing the moment conditions or the likelihood function is typically difficult. To equate the order of computation 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) and Kristensen and Salanié (2010) 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. 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 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. Significant process is only made recently, in an important paper by Kristensen and Salanié (2010) who develop bias reduction techniques for simulation estimators.

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.

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. We are concerned about estimating an unknown parameter $\theta \in \Theta \subset \mathbb{R}^k$. As discussed in the introduction, independent draws are typically preferrable in the method of simulated moments. The method of moment results are developed both for completenes 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. In the above $Pg(\cdot, \theta) = \int g(z_i, \cdot)\, dP(z_i)$ 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. 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

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

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 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)

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

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,

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

While $g(z_i;\theta)$ is typically a smooth function of z_i and θ, $\hat{g}(z_i;\theta)$ 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. The likelihood function $g(z;\theta)$ 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 $g(z;\theta) = g(y|x;\theta)$ 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 $R >> \sqrt{n}$ 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 $\sqrt{R}/n \rightarrow \infty$ is derived in Laroque and Salanie (1989), (1993) and Gourieroux and Monfort (1996), and is strengthened by Lee and Song (2015) to $\sqrt{R}\log R/n \rightarrow \infty$ for nonsmooth maximum likelihood like estimators. 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):

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

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

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

are Euclidean, cf. Lemma 25 (p. 27), Lemma 36 (p. 34), and Theorem 37 (p. 34) of Pollard (1984). Their envelope functions, denoted respectively by F, QF and PF, 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 $1\left(z_i'\theta + w_r \geq 0\right)$, 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 2, which implies that the class $\mathcal{F}$ and its projections $\mathcal{QF}$ and $\mathcal{PF}$ are Euclidean (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 2 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

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

b. Define

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

then

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

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

c. Further,

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

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 $U_{nR}(\theta) = \tilde{U}_{nR}(\theta) + \sqrt{m}(P_n - P)g(\cdot,\theta) + \sqrt{m}(Q_R - Q)h(\cdot,\theta)$. 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 $q(z, \omega, \theta)$. In the vector case, the notation of $||\cdot||$ (e.g. $||\tilde{S}_{nR}(\theta)||$) denotes Euclidean norms. The stated results in the lemma then follow from the equivalence between the L^2 norm and the L^1 norm, since for $g \in R^d$, $||g|| \leq d \sum_{j=1}^{d} |g_j|$. □

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 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; 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 2, 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

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

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, although likely not necessary, 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 $\hat{\theta} \overset{p}{\to} \theta_0$. 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

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

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 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 $g(y \,|\, x, \theta)$ 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 The class of functions indexed by both θ and z: $\mathcal{L} = \{\, q(z, \cdot, \theta) \,:\, z \in Z, \theta \in \Theta\}$ and is VC with a uniformly bounded envelope function L. The classes $\{g\left(\cdot, \theta\right), \theta \in \Theta\}$ and $\{\log g\left(\cdot, \theta\right), \theta \in \Theta\}$ are also both VC with a uniformly bounded envelope.

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 2, 3, and 4, $\hat{L}\left(\theta\right) - \hat{L}\left(\theta_0\right)$ converges to $L\left(\theta\right) - L\left(\theta_0\right)$ as $m \to \infty$ uniformly over Θ.

Proof Consider the decomposition

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

where

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

First, by Theorem 19.13 of van der Vaart (1999), $A\left(\theta\right)$ 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 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 2, 3, and 4, 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) (pp 10), the requirement that $\sup_{\theta\in\Theta} |\hat{L}(\theta) - L(\theta)| = o_p(1)$ 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 $\hat{R}(\theta) = \hat{L}\left(\theta\right) - L\left(\theta\right) - \hat{L}\left(\theta_0\right) + L(\theta_0) - \hat{D}'\left(\theta - \theta_0\right)$,
$$\sup_{\|\theta-\theta_0\|\leq\delta} \frac{m\hat{R}(\theta)}{1 + m\left\|\theta - \theta_0\right\|^2} = o_p\left(1\right). \tag{2}$$

(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 following high level assumption. It is typically satisfied when the true non-simulated log likelihood function is pointwise differentiable.

ASSUMPTION 5 There exists a mean zero random variable $D_0\left(z_i\right)$ with finite variance such that for any $\delta \to 0$ we have
$$\sup_{\|\theta-\theta_0\|\leq\delta} \frac{nR_n\left(\theta\right)}{1 + n\|\theta - \theta_0\|^2} = o_p\left(1\right) \tag{3}$$
for
$$R_n\left(\theta\right) \equiv \left(P_n - P\right)\left(\log g(\cdot, \theta) - \log g(\cdot, \theta_0)\right) - \hat{D}_0'(\hat{\theta} - \theta_0),$$
where
$$\hat{D}_0 = \frac{1}{n}\sum_{i=1}^{n} D_0\left(z_i\right).$$

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 $\bar{\kappa} > 0$ such that $L(\theta) \leq L(\theta_0) - \bar{\kappa}\|\theta\|^2$;

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

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

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

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

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

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

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

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 6 For all simulation lengths R and all parameters θ, both $g\left(z_i;\theta\right)$ and $Q_R q\left(z_i,\cdot;\theta\right)$ 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 7 Define $f\left(z,\omega,\theta\right) = q\left(z,\omega,\theta\right)/g\left(z,\theta\right) - q\left(z,\omega,\theta_0\right)/g\left(z,\theta_0\right)$, then (1) $Q\times P\left[\sup_{\|\theta-\theta_0\|=o(1)} f\left(\cdot,\cdot,\theta\right)^2\right] = o\left(1\right)$, (2) $\sup_{\|\theta-\theta_0\|\leq\delta, z\in Z} \mathrm{Var}_\omega f\left(z,\omega,\theta\right) = O\left(\delta\right)$.

ASSUMPTION 8 Define $\psi\left(\omega,\theta\right) = \int \frac{q(z,\omega,\theta)}{g(z,\theta)} f\left(z\right)dz$, where $f\left(z\right)$ is the joint density or probability mass function of the data. There exists a random vector $D_1\left(\omega_r\right)$ with finite variance such that for $\hat{D}_1 = \frac{1}{R}\sum_{r=1}^R D_1\left(\omega_r\right) - QD_1\left(\omega_r\right)$,

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

Remark When $g\left(z;\theta\right)$ represents the joint likelihood of the data, $f\left(z\right) = g\left(z;\theta_0\right)$. When $g\left(z;\theta\right) = g\left(y|x;\theta\right)$ represents a conditional likelihood, $f\left(z\right) = g\left(z;\theta_0\right)f\left(x\right)$ where $f\left(x\right)$ is the marginal density or probability mass function of the conditioning variables, in which case $\psi\left(\omega,\theta\right) = \int\int \frac{q(z,\omega,\theta)}{g(z,\theta)} g\left(y|x;\theta_0\right)dy f\left(x\right)dx$, with the understanding that integrals become summations in the case of discrete data. Assumption 8 can be further simplified when the true likelihood $g\left(z,\theta\right)$ is twice continuously differentiable (with bounded derivatives for simplicity). In this case

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

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

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

$$+ \quad P\left(\frac{1}{g\left(\cdot,\theta\right)} - \frac{1}{g\left(\cdot,\theta_0\right)}\right)\left(\hat{g}\left(\cdot,\theta\right) - g\left(\cdot,\theta\right) - \hat{g}\left(\cdot,\theta_0\right) + g\left(\cdot,\theta_0\right)\right).$$

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

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

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

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

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

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

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

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 $A\left(\theta\right)$ and $B\left(\theta\right)$.

It follows immediately from Assumption 5 that Conditions (5) and (6) of Theorem 2 hold for the first term $A\left(\theta\right)$ because $n \geq m$, since (3) is increasing in n:

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

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

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

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

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,

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

where the last inequality follows from $\sup_{\theta\in\Theta,z\in Z} |\hat{g}(z,\theta) - g(z,\theta)| = O_p\left(\frac{1}{\sqrt{R}}\right)$ 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 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 $\|\hat{\theta} - \theta_0\| = o_p(1/\log R)$:

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

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 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 $\delta = o\left((\log R)^{-1}\right)$, for

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

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

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

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 $B_{22}\left(\theta,z\right) = o_p\left(\frac{1}{\sqrt{R}}\right)$ uniformly in $\theta-\theta_0 \leq \delta$ 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 $\alpha > 0$:

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

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

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

where

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

$B_0\left(\theta\right) = \left(Q_R - Q\right)\left(\psi\left(\cdot,\theta\right) - \psi\left(\omega,\theta_0\right)\right)$, and $\psi\left(\omega,\theta\right) = \int \frac{q(z,\omega,\theta)}{g(z,\theta)} f\left(z\right)dz$. Upon noting that $Q\frac{q(\cdot,z,\theta)}{g(z,\theta)} = 1$ identically, $Qf\left(z,\cdot,\theta\right) = 0$ and $PQf\left(z,\cdot,\theta\right) = 0$. The proof of Theorem 2.5 (pp. 83) of Neumeyer (2004) shows that, since by assumption 7,

$$
Q \times P\left[\sup_{\|\theta-\theta_0\|=o(1)} f\left(\cdot,\cdot,\theta\right)^2\right] = o\left(1\right)
$$

the envelope expectation $Q \otimes P\left(F\right)^2$ converges to zero. Hence,

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

Finally, B_0 is handled by Assumption 8. So that, since $B\left(\theta\right) = B_1\left(\theta\right) - B_0\left(\theta\right) + B_2\left(\theta\right) + B_3\left(\theta\right) + B_0\left(\theta\right)$, and each of $B_1\left(\theta\right) - B_0\left(\theta\right)$, $B_2\left(\theta\right)$ and $B_3\left(\theta\right)$ is $o_P\left(\frac{1}{R}\right)$ uniformly in $||\theta-\theta_0|| \leq$

δ, for any $\delta \to 0$, we have, for $\hat{R}_1(\theta) = B(\theta) - \hat{D}_1'(\theta - \theta_0)$,

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

This together with (5) implies that condition 6 of Theorem 2 is satisfied with $\hat{D} = \hat{D}_0 + \hat{D}_1$, since we can bound (2) by

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

Finally to verify condition 5 in Theorem 2, 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 $\sqrt{m}\hat{D} \overset{d}{\to} N(0,\Sigma)$ 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

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

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

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

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

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

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 $\hat{H} = H + o_p(1)$ and $\hat{\Sigma}_h = \Sigma_h + o_p(1)$, $\hat{\Sigma}_g = \Sigma_g + o_p(1)$. Hence $\hat{\Sigma} = \Sigma + o_p(1)$ 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

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

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 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 $\alpha > 0$:

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

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 5 Under Assumption 2 for the GMM model or Assumptions 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 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 (7) 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):

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

By arguments that handle the term B in the proof of Theorem 4, in this region,

$$\omega(h) \equiv m\hat{L}\left(\hat{\theta}+h/\sqrt{m}\right)-m\hat{L}\left(\hat{\theta}\right) = -\frac{1}{2}\left(1+o_p(1)\right)h'J_0h + mO_p\left(\frac{1}{\sqrt{m}}\right).$$

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

$$\int_{\sqrt{m}a_m \leq |h| \leq \sqrt{m}\delta} \exp\left(\omega(h)\right)dh = \exp\left(O_p\left(\sqrt{m}\right)\right)\int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{2}\left(1+o_p(1)\right)h'J_0h\right)dh.$$

The tail of the normal distribution can be estimated by w.p. → 1:

$$\begin{aligned}&\textstyle\int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{2}\left(1+o_p(1)\right)h'J_0h\right)dh\\ &\leq \textstyle\int_{\sqrt{m}a_m \leq |h|} \exp\left(-\frac{1}{4}h'J_0h\right)dh \leq C\left(\sqrt{m}a_m\right)^{-1}\exp\left(-ma_m^2\right),\end{aligned}$$

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

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

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 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 $g(z_i,\theta) = \Phi(x_i'\theta)^{y_i}(1-\Phi(x_i'\theta))^{1-y_i}$, where $\Phi(u) = \int^u \frac{1}{\sqrt{2\pi}} e^{-v^2/2}dv$, and the true parameter θ_0 maximizes

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

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

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

and $w_r \overset{\text{iid}}{\sim} N(0,1)$. Note that $Q\hat{g}(z_i,\theta) = Qq(z_i,\cdot,\theta) = g(z_i,\theta)$. These functions fall into the class of multinomial discrete choice models studied in Pakes and Pollard (1989) and hence satisfy Assumptions 2 and 3. Assumption 6 is immediate. Assumptions 4 and 7 are satisfied because of the indicator function in $q(z_i, w_r, \theta)$ when θ and z have bounded support. Furthermore, since $g(z_i,\theta)$ is differentiable, Assumptions 5 and 8 can be directly verified with $D_0(z_i) = \frac{y_i - \Phi(x_i'\theta_0)}{\Phi(x_i'\theta_0)(1-\Phi(x_i'\theta_0))}\phi(x_i'\theta_0)x_i$ and

$$D_1(\omega_r) = \int \frac{\mathbf{1}\{x_i'\theta_0 + w_r \geq 0\} - \Phi(x_i'\theta_0)}{\Phi(x_i'\theta_0)(1-\Phi(x_i'\theta_0))}\phi(x_i'\theta_0)x_i f(x_i)\,dx_i.$$

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:

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

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

In the above, for an estimate of the derivative of $\hat{g}(z_i, \hat{\theta}^{SMLE})$ with respect to θ, denoted as $\nabla\hat{g}(z_i, \hat{\theta}^{SMLE})$, we define:

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

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

$$\frac{\partial g(z_i, \theta)}{\partial \theta} = \left(-\frac{\partial}{\partial \theta}\Phi(-x_i'\theta)\right)^{y_i}\left(\frac{\partial}{\partial \theta}\Phi(-x_i'\theta)\right)^{1-y_i}$$

$$= \quad x_i \left[\left(\frac{\partial}{\partial w} \Phi(w) \Big|_{w=-x_i'\theta} \right)^{y_i} \left(-\frac{\partial}{\partial w} \Phi(w) \Big|_{w=-x_i'\theta} \right)^{1-y_i} \right] = (-1)^{1-y_i} x_i \; \phi\left(x_i'\theta\right)$$

Therefore we use:

$$\nabla \hat{g}(z_i, \hat{\theta}^{SMLE}) = (-1)^{1-y_i} x_i \; \hat{\phi}\left(x_i'\hat{\theta}\right)$$

where to be agnostic about knowledge of $\phi\left(\cdot\right)$, we use a first order two-sided formula to define:

$$\hat{\phi}\left(w\right) \equiv \frac{1}{R} \sum_{r=1}^{R} \frac{\mathbf{1}\{w_r \leq w + \epsilon\} - \mathbf{1}\{w_r \leq w - \epsilon\}}{2\epsilon} \quad = \quad \frac{1}{2\epsilon} \frac{\#\{w - \epsilon \leq w_r \leq w + \epsilon\}}{R},$$

where ϵ is a step size parameter. In the simulation, we experiment with a range of the step size parameter, $\epsilon = R^{-\alpha}$, where α ranges in $\left[2 \; \frac{3}{2} \; 1 \; \frac{3}{4} \; \frac{1}{2} \; \frac{1}{3} \; \frac{1}{4} \; \frac{1}{8} \; \frac{1}{10} \; \frac{1}{15}\right]$. It turns out that the larger step sizes produces more accurate coverage for a larger range of n and R. Therefore we use $\alpha = 1/15$ in the results reported in the following tables.

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 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 intercept and the slope coefficient, respectively. The results show that the asymptotic distribution accurately represents the finite sample distribution when $m = \min\left(R, n\right)$ 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,

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	100
50	0.9334	0.9428	0.9436	0.9458	0.9476	0.9508	0.9488	0.9526	0.9524	0.9546
	0.9946	0.9844	0.979	0.9736	0.9712	0.9672	0.9638	0.9656	0.9676	0.9682
100	0.9254	0.9266	0.9342	0.9326	0.9356	0.9342	0.9378	0.939	0.939	0.9408
	0.983	0.9628	0.9564	0.956	0.9536	0.9538	0.9524	0.9536	0.9514	0.9558
200	0.9342	0.9388	0.9378	0.9382	0.938	0.937	0.9408	0.9366	0.9424	0.943
	0.9672	0.9528	0.9462	0.9456	0.9442	0.9472	0.9466	0.9462	0.9474	0.9488
400	0.9448	0.9412	0.9388	0.9398	0.9358	0.9388	0.936	0.9362	0.9346	0.9378
	0.9512	0.9442	0.9516	0.9472	0.9498	0.9482	0.9514	0.9528	0.9512	0.9536
800	0.9438	0.9442	0.9384	0.9386	0.9458	0.9412	0.9454	0.9408	0.9422	0.9434
	0.9468	0.9462	0.9394	0.9442	0.948	0.9518	0.9512	0.9492	0.9538	0.9508

The total number of Monte Carlo repetitions is 5000.

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	100
50	0.662	0.7822	0.8344	0.8504	0.8954	0.9298	0.939	0.9474	0.9508	0.9538
	0.9614	0.9578	0.9566	0.9562	0.9626	0.964	0.9622	0.9652	0.9672	0.968
100	0.594	0.742	0.799	0.8236	0.878	0.9136	0.9282	0.9352	0.9362	0.9402
	0.9246	0.9348	0.9414	0.9406	0.9456	0.951	0.9518	0.9534	0.951	0.9554
200	0.5892	0.737	0.8016	0.8232	0.8784	0.9154	0.9292	0.931	0.9384	0.9426
	0.9172	0.931	0.9338	0.9334	0.9378	0.9452	0.9454	0.9454	0.9474	0.9486
400	0.5744	0.727	0.794	0.8218	0.8734	0.9156	0.9224	0.9292	0.932	0.9362
	0.9134	0.9308	0.9412	0.9394	0.9472	0.947	0.9508	0.9524	0.9512	0.9536
800	0.575	0.7302	0.8022	0.8182	0.8828	0.9192	0.9314	0.9362	0.9394	0.942
	0.9242	0.9362	0.9328	0.9376	0.9442	0.9498	0.9504	0.949	0.9538	0.9508

The total number of Monte Carlo repetitions is 5000.

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

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.9956	0.953	0.9542	0.9574	0.9646	0.9662	0.9604	0.9542	0.9518	0.9526
	0.9944	0.9824	0.976	0.9724	0.9694	0.9582	0.9586	0.9558	0.9574	0.9582
100	0.94	0.943	0.9522	0.9572	0.9576	0.9436	0.9402	0.9422	0.9402	0.9408
	0.9856	0.9722	0.9628	0.962	0.9534	0.952	0.9506	0.952	0.9528	0.954
200	0.9326	0.9472	0.9572	0.951	0.9464	0.9398	0.9428	0.9378	0.9418	0.9422
	0.9742	0.9568	0.9446	0.9456	0.9392	0.9434	0.9412	0.9418	0.9432	0.9436
400	0.9356	0.9512	0.954	0.9494	0.9346	0.941	0.9368	0.9374	0.9352	0.9376
	0.9638	0.9488	0.9514	0.9498	0.9482	0.9504	0.9492	0.9514	0.9526	0.9526
800	0.9416	0.948	0.943	0.943	0.9458	0.9416	0.9452	0.9422	0.9418	0.9434
	0.9542	0.9462	0.9402	0.943	0.9468	0.9496	0.9502	0.95	0.95	0.9504

The total number of Monte Carlo repetitions is 5000.

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.

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 step size choice, where we have chosen $\alpha = 1/15$ in this simulation example. The optimal choice of C and α is an open challenging theoretical question in this context which is beyond what

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

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.6676	0.7916	0.842	0.8606	0.9002	0.9314	0.9424	0.9474	0.9504	0.9508
	0.9186	0.9316	0.9394	0.9364	0.9472	0.9506	0.9546	0.9534	0.9566	0.958
100	0.6056	0.754	0.8064	0.8274	0.88	0.9166	0.9298	0.937	0.938	0.9392
	0.91	0.922	0.9322	0.932	0.9418	0.9488	0.9488	0.9508	0.9528	0.954
200	0.5948	0.7442	0.8052	0.8258	0.8804	0.9156	0.9288	0.9318	0.9396	0.9412
	0.9114	0.9224	0.9244	0.9272	0.9322	0.941	0.9404	0.9414	0.9428	0.9436
400	0.5736	0.7236	0.795	0.8234	0.8758	0.916	0.9232	0.93	0.9324	0.9364
	0.9132	0.9294	0.94	0.9368	0.9446	0.9498	0.9482	0.9512	0.9526	0.9524
800	0.5724	0.7308	0.8028	0.8226	0.8826	0.919	0.9316	0.9364	0.9394	0.942
	0.9198	0.9338	0.9338	0.936	0.9428	0.9486	0.9494	0.949	0.9498	0.9504

The total number of Monte Carlo repetitions is 5000.

we are able to obtain in this paper. We can also adopt Silverman's rule of thumb or use cross-validation methods for data driven automated bandwidth selection. However, these methods are designed for nonparametric curve fitting. They are not known to be optimal in semiparametric problems such as variance estimation that we consider. We conducted a small experiment where we looked for the stepsize that minimizes the difference between the analytical and asymptotic variance in each Monte Carlo trial across the above values of α, and find substantial variation in the best step size in this sense, which is tabulated in Table 7. However, minimizing the difference between the analytical and asymptotic variance does not seem to translate into coverage accuracy. Comparing Table 1 (for $\alpha = 1/15$) with Table 8 (for $\alpha = 1/2$) shows that the larger step size produces more accurate confidence intervals than the smaller step size in general. We leave it for future research for a theoretical guidance on the properties of the step sizes.

Overlapping draws are applicable in situations in which independent draws are not computationally practicable, or with nonsmooth moment conditions where the theoretical va-

Table 5: False empirical coverage frequency for 95% confidence interval, analytical, independent draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.7816	0.9182	0.9208	0.922	0.9332	0.942	0.9506	0.9506	0.9502	0.948
	0.912	0.9266	0.9384	0.94	0.9434	0.9506	0.9514	0.9558	0.957	0.9598
100	0.8838	0.908	0.8992	0.9068	0.9208	0.9366	0.9354	0.9358	0.9392	0.9426
	0.9102	0.9184	0.9288	0.9254	0.9378	0.945	0.951	0.9518	0.952	0.9542
200	0.8622	0.8966	0.9108	0.914	0.9198	0.9352	0.937	0.9414	0.9426	0.941
	0.8912	0.9152	0.918	0.9266	0.9356	0.9378	0.9422	0.9448	0.9454	0.9458
400	0.8764	0.9056	0.9188	0.9214	0.929	0.934	0.9338	0.9372	0.9386	0.9378
	0.9068	0.9256	0.9366	0.936	0.9426	0.9486	0.9486	0.9528	0.952	0.9528
800	0.9056	0.9228	0.9288	0.932	0.9352	0.9406	0.9422	0.9454	0.9448	0.9442
	0.9116	0.9332	0.9378	0.94	0.9428	0.9488	0.9476	0.9498	0.9496	0.9488

The total number of Monte Carlo repetitions is 5000.

Table 6: False empirical coverage frequency for 95% confidence interval, numerical, independent draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.2892	0.7306	0.8178	0.8348	0.8944	0.9202	0.935	0.94	0.9462	0.9488
	0.417	0.749	0.853	0.8786	0.9252	0.947	0.9524	0.9614	0.9602	0.9652
100	0.5322	0.8216	0.855	0.8704	0.8986	0.9196	0.9276	0.9324	0.9358	0.9382
	0.5858	0.8462	0.8936	0.898	0.9272	0.9342	0.9434	0.9494	0.9516	0.9534
200	0.75	0.8646	0.8908	0.8978	0.9128	0.9316	0.934	0.9386	0.9404	0.9384
	0.7778	0.893	0.9052	0.9156	0.9302	0.9334	0.9438	0.9456	0.9482	0.9482
400	0.832	0.8932	0.9076	0.917	0.926	0.9324	0.9324	0.9376	0.9376	0.9378
	0.8722	0.914	0.9292	0.9316	0.9402	0.9474	0.9476	0.9522	0.9522	0.9532
800	0.8908	0.9196	0.924	0.9302	0.9334	0.94	0.9422	0.944	0.9458	0.9442
	0.901	0.9318	0.936	0.9368	0.9426	0.9498	0.9472	0.95	0.9488	0.9504

The total number of Monte Carlo repetitions is 5000.

Table 7: Step size parameter (α) minimizing difference between numerical and asymptotic variance in each trial of (n, κ)

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.75	0.75	0.75	0.5	0.5	0.5	0.5	0.5	0.5	0.5
100	0.75	0.75	0.75	0.5	0.5	0.5	0.5	0.5	0.5	0.5
200	0.75	0.75	0.75	0.5	0.5	0.5	0.5	0.067	0.067	0.067
400	0.75	0.75	0.75	0.5	0.5	0.067	0.067	0.067	0.067	0.067
800	0.75	0.75	0.75	0.125	0.067	0.067	0.067	0.067	0.067	0.067

The total number of Monte Carlo repetitions is 5000.

Table 8: Coverage with $\alpha = 1/2$, overlapping, correct variance

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.854	0.871	0.881	0.891	0.910	0.932	0.937	0.947	0.951	0.954
	0.963	0.949	0.948	0.945	0.953	0.956	0.958	0.964	0.967	0.968
100	0.846	0.873	0.890	0.891	0.907	0.927	0.930	0.938	0.938	0.941
	0.948	0.939	0.935	0.940	0.940	0.945	0.948	0.953	0.951	0.954
200	0.863	0.890	0.906	0.906	0.915	0.928	0.936	0.932	0.939	0.941
	0.944	0.935	0.929	0.931	0.932	0.939	0.940	0.944	0.945	0.949
400	0.887	0.900	0.912	0.920	0.922	0.934	0.930	0.935	0.935	0.935
	0.942	0.936	0.941	0.939	0.944	0.943	0.948	0.951	0.949	0.953
800	0.897	0.914	0.922	0.915	0.936	0.935	0.941	0.939	0.942	0.942
	0.945	0.940	0.934	0.938	0.942	0.948	0.947	0.947	0.952	0.950

The total number of Monte Carlo repetitions is 5000.

lidity 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 5 and 6 we report the couterparts with independent draws. When R is relatively small compared to n, the point estimate is sufficiently biased and the confidence interval does not center on the true parameter value. Larger R reduces bias and leads to more accurate empirical coverage frequencies. In particular, the independent draws method does not perform well relative to overlapping draws when both N ad R/N are small. In unreported simulations we also experimented with the EM algorithm. While the EM algorithm converges quickly, we are not able to obtain empirical coverage frequencies that are close to the nominal level.

We also report the mean bias and root mean square error for both overlapping and independent draws in tables 9 to 12. Especially for small n and R/n, overlapping draws tend to have smaller bias. While its intercept term tends to have a larger RMSE in overlapping draws, the larger variance can be accounted for in constructing confidence intervals.

Table 9: Mean Bias, overlapping draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.0636	0.0777	0.0782	0.077	0.0726	0.0639	0.0637	0.0592	0.0579	0.0575
	0.0027	0.0015	0.00066	-0.00051	0.0010	0.00036	-5.4E-05	0.00048	-2.2E-05	0.00022
100	0.0323	0.033	0.0331	0.0312	0.0299	0.0279	0.0269	0.027	0.0256	0.0257
	0.0009	0.0005	0.0004	0.0004	0.0006	0.0011	0.0005	0.0006	0.0007	0.0005
200	0.0129	0.0139	0.015	0.0147	0.0145	0.0142	0.014	0.0127	0.0131	0.0131
	-0.0009	-0.0011	-0.0008	-0.00065	-0.00078	-0.00067	-0.0009	-0.0009	-0.00104	-0.0011
400	0.0056	0.0047	0.0062	0.0051	0.0062	0.0062	0.0051	0.0052	0.0052	0.0052
	-0.00035	-0.0003	-0.0004	-0.0004	-0.00045	-0.00053	-0.00053	-0.00066	-0.00059	-0.00057
800	0.0023	0.0025	0.0023	0.0029	0.0029	0.0026	0.0024	0.0028	0.0025	0.0026
	-0.00023	-0.00033	-0.00023	-0.00038	-0.00034	-0.00031	-0.00039	-0.00040	-0.00049	-0.00045

The total number of Monte Carlo repetitions is 5000.

Table 10: Mean Bias, Independent draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	-.16800	.02191	.05585	.05852	.05531	.06072	.06648	.06469	.06100	.05951
	.00202	.00049	.00000	-.00143	.00001	.00059	-.00016	-.00009	.00046	-.00002
100	-.06624	.00892	.01600	.02348	.03354	.03154	.02970	.02825	.02566	.02617
	-.00037	.00120	.00018	.00101	.00005	.00109	.00091	.00058	.00061	.00067
200	-.02114	.01151	.01541	.01802	.01637	.01534	.01412	.01387	.01352	.01381
	-.00156	-.00128	-.00135	-.00106	-.00110	-.00104	-.00089	-.00111	-.00096	-.00084
400	-.00027	.00542	.00358	.00488	.00515	.00538	.00487	.00527	.00567	.00506
	-.00094	-.00046	-.00049	-.00005	-.00059	-.00055	-.00066	-.00082	-.00050	-.00061
800	-.00361	.00095	.00253	.00220	.00267	.00242	.00309	.00313	.00267	.00284
	-.00099	-.00057	-.00033	-.00050	-.00040	-.00058	-.00055	-.00041	-.00040	-.00038

The total number of Monte Carlo repetitions is 5000.

Table 11: Root Mean Square Error, Overlapping draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.5407	0.5407	0.5407	0.5407	0.5407	0.5407	0.5407	0.5407	0.5407	0.5407
	0.1577	0.1533	0.1489	0.1489	0.1398	0.1347	0.1327	0.1301	0.1287	0.1284
100	0.3832	0.2829	0.2523	0.2386	0.2092	0.1878	0.1787	0.1749	0.1718	0.1711
	0.1024	0.0954	0.0923	0.0909	0.0879	0.0857	0.0844	0.0833	0.0828	0.0825
200	0.2632	0.1929	0.1690	0.1592	0.1397	0.1260	0.1204	0.1181	0.1153	0.1148
	0.0660	0.0630	0.0618	0.0611	0.0598	0.0583	0.0578	0.0573	0.0569	0.0566
400	0.1851	0.1350	0.1175	0.1106	0.0972	0.0872	0.0844	0.0822	0.0814	0.0804
	0.0449	0.0427	0.0413	0.0415	0.0404	0.0395	0.0391	0.0388	0.0386	0.0386
800	0.1316	0.0944	0.0819	0.0780	0.0670	0.0604	0.0579	0.0567	0.0562	0.0557
	0.0305	0.0292	0.0289	0.0286	0.0280	0.0274	0.0272	0.0272	0.0270	0.0270

The total number of Monte Carlo repetitions is 5000.

Table 12: Root Mean Square Error, Independent draws

n, κ	0.2	0.5	0.8	1	2	5	10	20	50	100
50	0.3204	0.2869	0.2987	0.3025	0.2864	0.2756	0.2712	0.2694	0.2661	0.2640
	0.1384	0.1427	0.1418	0.1407	0.1390	0.1340	0.1323	0.1299	0.1293	0.1280
100	0.1862	0.1902	0.1931	0.1947	0.1851	0.1773	0.1753	0.1731	0.1708	0.1700
	0.0936	0.0939	0.0924	0.0925	0.0886	0.0855	0.0841	0.0833	0.0824	0.0823
200	0.1407	0.1336	0.1293	0.1260	0.1221	0.1171	0.1163	0.1150	0.1148	0.1152
	0.0680	0.0633	0.0624	0.0605	0.0592	0.0586	0.0573	0.0569	0.0567	0.0568
400	0.0971	0.0894	0.0853	0.0849	0.0821	0.0814	0.0814	0.0810	0.0804	0.0803
	0.0456	0.0417	0.0413	0.0406	0.0395	0.0392	0.0391	0.0389	0.0387	0.0385
800	0.0629	0.0595	0.0587	0.0577	0.0568	0.0559	0.0556	0.0549	0.0551	0.0551
	0.0309	0.0287	0.0285	0.0283	0.0280	0.0274	0.0273	0.0271	0.0271	0.0270

The total number of Monte Carlo repetitions 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, 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.

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 (2010): "Higher order improvements for 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.

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.