Improved Estimation by Simulated Maximum Likelihood*

Kirill S. Evdokimov
Massachusetts Institute of Technology

Ilze Kalnina
North Carolina State University

February 15, 2019
PRELIMINARY AND INCOMPLETE

Abstract

We propose a method that improves the Simulated Maximum Likelihood Estimator (SMLE). The method does not impose any additional assumptions on the model; instead, it makes a more efficient use of the information available in this framework. Our approach allows reducing the number of simulation draws S without sacrificing the precision of the estimator. In particular, the method provides a semiparametrically efficient estimator when $\sqrt{n}/S \not\to 0$, a situation in which SMLE is biased. Moreover, under certain smoothness restrictions, our estimator can be asymptotically efficient when S is finite. The method should be most useful when the evaluation of the simulated likelihood function is computationally expensive.

1 Introduction

We propose a method that improves the Simulated Maximum Likelihood Estimator (SMLE). The method does not impose any additional assumptions on the model; instead, it makes a more efficient use of the information available in this framework. Our approach allows reducing the number of simulation draws S without sacrificing the precision of the estimator. In particular, the method provides a semiparametrically efficient estimator when $\sqrt{n}/S \not\to 0$, a situation in which SMLE is biased. Moreover, under certain smoothness restrictions, our estimator can be asymptotically efficient when S is finite. The method should be most useful when the evaluation of the simulated likelihood function is computationally expensive.

Estimation of structural models can be computationally demanding. In particular, often some unobserved (state) variables need to be integrated out to obtain the likelihood function. For example, Morellec et al. (2012) estimate agency conflicts in a structural model with unobserved firm heterogeneity using a panel data on leverage ratios; they use SMLE to estimate the model.

SMLE replaces integration over the distribution of (unobserved) variables by simulation. When the number of draws per observation ($S/\sqrt{n}$) is large, SMLE achieves the efficiency of the MLE.

* `kevdokim@mit.edu` and `ikalnin@ncsu.edu`. Support from NSF and UCL is gratefully acknowledged.

SMLE was proposed by Lerman and Manski (1981). Lee (1992) and Gourieroux and Monfort (1993) study the asymptotic properties of the SMLE. Kristensen and Salanié (2017) develop higher-order expansions for a class of estimators that include SMLE. Danielsson (1994) and Durbin and Koopman (2000), among others, apply SMLE to estimate stochastic volatility models.

[[TBC]]

2 Definitions of the Estimators

2.1 SMLE

We first describe the SMLE; Section 2.2 introduces the improved SMLE. We have

$$\widehat{\theta}_{\mathrm{SMLE}} \equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln \left(\overline{f}_{iS}(\theta)\right),$$

$$\overline{f}_{iS}(\theta) \equiv \frac{1}{S} \sum_{s=1}^{S} f_{is}(\theta), \qquad f_{is}(\theta) \equiv f\left(W_i, \varepsilon_{is}, \theta\right). \tag{1}$$

In the above, $W_i \in \mathcal{W}$ are the observed variables; $\varepsilon_i \equiv (\varepsilon_{i1}, \ldots, \varepsilon_{iS})$, where $\varepsilon_{is} \in \mathcal{E}$ are i.i.d. draws (across i and s) from the distribution with some known CDF G_ε.

Let

$$h_i(\theta) \equiv h(\theta, W_i) \equiv E\left[f\left(W_i, \varepsilon_{is}, \theta\right) | W_i\right] = \int f\left(W_i, \varepsilon, \theta\right) dG_\varepsilon(\varepsilon), \tag{2}$$

Ideally, one would have wanted to use the MLE estimator

$$\widehat{\theta}_{\mathrm{MLE}} \equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \ln\left(h\left(W_i, \theta\right)\right). \tag{3}$$

One resorts to SMLE as an approximation of MLE when computing $h(w, \theta)$ analytically is impossible or impractical.

2.2 SM$\mathcal{I}$LE method

This section introduces the new estimator. The simulated densities are informative about the function of interest, but they also contain noise due to simulations. SM$\mathcal{I}$LE reduces the magnitude of this noise by using additional information about simulations.

The additional information about simulations that SM$\mathcal{I}$LE uses is twofold. First, we know the distribution from which ε_{is} is simulated. Second, we know that ε_{is} is drawn independently of W_i. This information can be captured by the basis functions φ introduced below. This additional information about simulations is unrelated to the information about the model, and Assumption 1 requires φ to mirror this fact. In particular, Assumption 1 ensures that φ is orthogonal to the model information as captured by h.

Let $\varphi_K(w,\varepsilon) \equiv (\varphi_{1K}(w,\varepsilon),\ldots,\varphi_{KK}(w,\varepsilon))'$ be a $K\times 1$ vector containing some approximating functions that satisfy the following assumption:

Assumption 1. *$E[\varphi_K(w,\varepsilon_{is})] = 0$ for all $w \in \mathcal{W}$.*

Functions φ_K are chosen by the researcher. Assumption 1 is automatically satisfied if $\varphi_K(w,\varepsilon)$ is constructed as

$$\varphi_K(W_i,\varepsilon_{is}) = \varphi^W_{K_W}(W_i) \otimes \left(\varphi^{\varepsilon}_{K_\varepsilon}(\varepsilon_{is}) - E\left[\varphi^{\varepsilon}_{K_\varepsilon}(\varepsilon_{is})\right]\right),$$

where $\varphi^W_{K_W}(W_i) = \left(\varphi^W_{1K_W}(W_i),\ldots,\varphi^W_{1K_W}(W_i)\right)'$ contains K_W functions of W_i, and $\varphi^{\varepsilon}_{K_\varepsilon}(\varepsilon_{is}) = \left(\varphi^{\varepsilon}_{1K_\varepsilon}(\varepsilon_{is}),\ldots,\varphi^{\varepsilon}_{1K_\varepsilon}(\varepsilon_{is})\right)'$ contains K_ε functions of ε_{is}, so that $K = K_W \cdot K_\varepsilon$.[1]

Let

$$\alpha_K(\theta) \equiv \arg\min_a E\left[\left(f_{is}(\theta) - \varphi'_{is,K}a\right)^2\right], \tag{4}$$

where we denote $\varphi_{is} \equiv \varphi_{is,K} \equiv \varphi_K(W_i,\varepsilon_{is}) \in \mathbb{R}^K$. We can estimate $\alpha_K(\theta)$ by the OLS estimator

$$\widehat{\alpha}_K(\theta) \equiv \left(\mathcal{V}'_{\varphi_K}\mathcal{V}_{\varphi_K}\right)^{-1}\mathcal{V}'_{\varphi_K}\mathcal{V}_{f(\theta)}, \text{ where}$$

$$\mathcal{V}_{f(\theta)} \equiv \begin{pmatrix} f_{11}(\theta) \\ f_{12}(\theta) \\ \vdots \\ f_{1S}(\theta) \\ f_{21}(\theta) \\ f_{22}(\theta) \\ \vdots \\ f_{2S}(\theta) \\ f_{31}(\theta) \\ \vdots \\ f_{nS}(\theta) \end{pmatrix}, \; \mathcal{V}_{\varphi_K} \equiv \begin{pmatrix} \varphi'_{11,K} \\ \varphi'_{12,K} \\ \vdots \\ \varphi'_{1S,K} \\ \varphi'_{21,K} \\ \varphi'_{22,K} \\ \vdots \\ \varphi'_{2S,K} \\ \varphi'_{31,K} \\ \vdots \\ \varphi'_{nS,K} \end{pmatrix} = \begin{pmatrix} \varphi_{11,1K} & \varphi_{11,2K} & \cdots & \varphi_{11,KK} \\ \varphi_{12,1K} & \varphi_{12,2K} & \cdots & \varphi_{12,KK} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{1S,1K} & \varphi_{1S,2K} & \cdots & \varphi_{1S,KK} \\ \varphi_{21,1K} & \varphi_{21,2K} & \cdots & \varphi_{21,KK} \\ \varphi_{22,1K} & \varphi_{22,2K} & \cdots & \varphi_{22,KK} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{2S,1K} & \varphi_{2S,2K} & \cdots & \varphi_{2S,KK} \\ \varphi_{31,1K} & \varphi_{31,2K} & \cdots & \varphi_{31,KK} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{nS,1K} & \varphi_{nS,2K} & \cdots & \varphi_{nS,KK} \end{pmatrix},$$

where $\mathcal{V}_{f(\theta)} \in \mathbb{R}^{nS\times 1}$, $\mathcal{V}_{\varphi_K} \in \mathbb{R}^{nS\times K}$, and $\widehat{\alpha}_K(\theta) \in \mathbb{R}^{K\times 1}$. Below we write $\varphi_{is} \equiv \varphi_{is,K} \equiv \left(\varphi_{is,1K},\ldots,\varphi_{is,KK}\right)' \equiv (\varphi_{1K}(W_i,\varepsilon_{is}),\ldots,\varphi_{KK}(W_i,\varepsilon_{is}))'$.

We then use

$$\bar{f}^{\mathcal{I}}_{iS}(\theta) \equiv \frac{1}{S}\sum_{s=1}^{S} f^{\mathcal{I}}_{is}(\theta), \qquad f^{\mathcal{I}}_{is}(\theta) \equiv f_{is}(\theta) - \varphi'_{is,K}\widehat{\alpha}_K(\theta),$$

[1] Instead of using the Kronecker product, one could use monomials.

as the simulated density, and suggest using the estimator $\widehat{\theta}_{\text{SM}\mathcal{I}\text{LE}}$:

$$\widehat{\theta}_{\text{SM}\mathcal{I}\text{LE}} \equiv \underset{\theta\in\Theta}{\text{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} \ln\left(\overline{f}^{\mathcal{I}}_{iS}\left(\theta\right)\right).$$

As we show below, an advantage of $\widehat{\theta}_{\text{SM}\mathcal{I}\text{LE}}$ is that it allows using a smaller number of simulation draws S than $\widehat{\theta}_{\text{SMLE}}$ does, without sacrificing the accuracy of estimation.

3 Efficiency/Improvements of SM$\mathcal{I}$LE: an Overview

The current section reviews SMLE and outlines why SM$\mathcal{I}$LE improves upon SMLE. The formal asymptotic results and proofs are provided in the following sections. We start with a brief outline for why SMLE needs $\sqrt{n}/S \to 0$ for efficiency.

3.1 Reviewing SMLE

Recall that the criterion function of SMLE is

$$\widehat{Q}_n\left(\theta\right) = E_n\left[\ln\left(\overline{f}_{iS}\left(\theta\right)\right)\right] \overset{p}{\to} E\left[\ln\left(h\left(\theta\right)\right)\right],$$

where the latter term is maximized at θ_0, so $\widehat{\theta}^{\text{SMLE}}$ solves the FOC $\nabla_\theta \widehat{Q}_n\left(\theta\right) = 0$. By the usual Taylor expansion

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -E_n\left[\nabla_{\theta\theta'} \ln \overline{f}_{iS}\left(\tilde{\theta}\right)\right]^{-1} \sqrt{n} E_n\left[\nabla_\theta \ln \overline{f}_{iS}\left(\theta_0\right)\right],$$

where $\tilde{\theta}$ lies between $\widehat{\theta}$ and θ_0. Since $\tilde{\theta} \to_p \theta_0$, by the usual arguments, $E_n\left[\nabla_{\theta\theta'} \ln \overline{f}_{iS}\left(\tilde{\theta}\right)\right] \to_p \Sigma \equiv E\left[\frac{\partial h_i(\theta_0)}{\partial\theta}\frac{\partial h_i(\theta_0)}{\partial\theta'}\right]$, and we focus on the term

$$\sqrt{n}E_n\left[\nabla_\theta \ln \overline{f}_{iS}\left(\theta_0\right)\right] = \sqrt{n}E_n\left[\frac{\nabla_\theta \overline{f}_{iS}\left(\theta_0\right)}{\overline{f}_{iS}\left(\theta_0\right)}\right] = \sqrt{n}E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right) + \overline{\Delta}_{iS}}{h_i\left(\theta_0\right) + \overline{\delta}_{iS}}\right],$$

where $\overline{\delta}_{iS}\left(\theta\right) \equiv \frac{1}{S}\sum_{s=1}^{S} \delta_{is}, \delta_{is}\left(\theta\right) \equiv f_{is}\left(\theta\right) - h_i\left(\theta\right)$, $\overline{\delta}_{iS} \equiv \overline{\delta}_{iS}\left(\theta_0\right)$, $\overline{\Delta}_{iS} \equiv \nabla_\theta \overline{\delta}_{iS}\left(\theta_0\right)$. The r.h.s. of the above can be separated into three terms

$$\begin{aligned}\sqrt{n}E_n\left[\nabla_\theta \ln \overline{f}_{iS}\left(\theta_0\right)\right] &= \sqrt{n}E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i\left(\theta_0\right)}\right] + \sqrt{n}E_n\left[\frac{\overline{\Delta}_{iS}}{h_i\left(\theta_0\right)} - \frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\overline{\delta}_{iS}\right] + \sqrt{n}R_n \\ &= T_{0n} + T_{1n} + \sqrt{n}R_n. \end{aligned} \tag{5}$$

The term T_{0n} is the "MLE" term that corresponds to using the true density h

$$T_{0n} \equiv \sqrt{n}E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i\left(\theta_0\right)}\right] \to_d N\left(0, \Sigma\right).$$

Next, the linear approximation term T_{1n} can be written as

$$T_{1n} \equiv \sqrt{n} E_n [\xi_i], \text{ where } \xi_i \equiv \frac{\overline{\Delta}_{iS}}{h_i(\theta_0)} - \frac{\nabla_\theta h_i(\theta_0)}{h_i^2(\theta_0)} \overline{\delta}_{iS}.$$

Since $E[\xi_i] = 0$ and $E\left[\|\xi_i\|^2\right] \to 0$ as $S \to \infty$, we have $T_{1n} = o_P(1)$.

Thus, we want $\sqrt{n} R_n = o_P(1)$, to ensure that

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \Sigma^{-1} T_{0n} + o_P(1) \to_d N\left(0, \Sigma^{-1}\right). \tag{6}$$

One can show that $R_n = O_P\left(E_n\left[\overline{\delta}_{iS}^2\right] + E_n\left[\|\overline{\delta}_{iS}\overline{\Delta}_{iS}\|\right]\right)$. Suppose $E_n\left[\overline{\delta}_{iS}^2\right] + E_n\left[\|\overline{\delta}_{iS}\overline{\Delta}_{iS}\|\right] = O_P(\gamma_n)$ for some $\gamma_n = o(1)$. Then equation (6) holds if

$$\sqrt{n}\gamma_n \to 0. \tag{7}$$

Section 4 shows that for a general class of simulation estimators including SMLE and SM$\mathcal{I}$LE, the condition in equation (7) implies asymptotic efficiency, i.e., equation (6).

Consider SMLE. When would equation equation (7) and hence equation (6) be satisfied? We have $E\left[\overline{\delta}_{iS}\right] = 0$ and $E\left[\overline{\Delta}_{iS}\right] = 0$, and $E\left[\overline{\delta}_{iS}^2\right] \propto 1/S$ and $E\left[\left\|\overline{\Delta}_{iS}\right\|^2\right] \propto 1/S$, so $\gamma_n = 1/S$. Hence, condition (7) requires $\sqrt{n}/S = o(1)$.

3.2 SM$\mathcal{I}$LE

This section describes the source of the efficiency gains of SM$\mathcal{I}$LE.

SM$\mathcal{I}$LE uses additional information about simulations to reduce the magnitude of the simulation errors $\delta_{is}(\theta) \equiv f_{is}(\theta) - h_i(\theta)$. More precisely, SM$\mathcal{I}$LE can be described as follows: SM$\mathcal{I}$LE removes from the simulated density $\overline{f}_{iS}(\theta)$ in (1) the component of the simulation errors $\delta_{is}(\theta)$ that is driven by the simulations alone and is unrelated to the model. The latter component is obtained by projecting the simulation errors $\delta_{is}(\theta)$ on the information about the simulations that is unrelated to the model information.

Such projection of $\delta_{is}(\theta)$ is not feasible because $\delta_{is}(\theta)$ is not observable, as it contains the function we want to estimate $h_i(\theta)$. The above description of SM$\mathcal{I}$LE is nevertheless valid for the following reasons. Recall that φ_{is} captures the information about simulations that is unrelated to the model information. Recall that Assumption 1 ensures that φ_{is} is uncorrelated with $h_i(\theta)$. As a result, a projection of $\delta_{is}(\theta) = f_{is}(\theta) - h_i(\theta)$ on φ_{is} is equivalent to a projection of $f_{is}(\theta)$ on φ_{is}, which can be readily implemented. The latter projection is used in the definition of $\alpha_K(\theta)$ in (4) that enters the definition of SM$\mathcal{I}$LE.

We now explain why SM$\mathcal{I}$LE does not require $\sqrt{n}/S = o(1)$. First, the Assumption 1 ensures that the basis functions φ_{is} are uncorrelated with the density of the data $h_i(\theta)$, which allows us to show that $\alpha_K(\theta)$ in (4) can identically be written as

$$\alpha_K(\theta) = \arg\min_a E\left[\left(\delta_{is}(\theta) - a'\varphi_{is,K}\right)^2\right]. \tag{8}$$

Next, let us decompose the SM$\mathcal{I}$LE density estimation errors,

$$\begin{aligned}
\overline{\delta}_{iS}^{\mathcal{I}}(\theta) &\equiv \overline{f}_{iS}^{\mathcal{I}}(\theta) - h_i(\theta) \\
&= \frac{1}{S}\sum_{s=1}^{S}\left(\delta_{is}(\theta) - \varphi_{is,K}'\widehat{\alpha}_K(\theta)\right) \\
&= \frac{1}{S}\sum_{s=1}^{S}\left(\delta_{is}(\theta) - \varphi_{is,K}'\alpha_K(\theta)\right) + \left(\frac{1}{S}\sum_{s=1}^{S}\varphi_{is,K}\right)'\left(\alpha_K(\theta) - \widehat{\alpha}_K(\theta)\right).
\end{aligned}$$

For the first term, we assume that $\delta_{is}(\theta)$ can be approximated by an increasing number of basis functions; by (8), this can be written as $E\left[\left(\delta_{is}(\theta) - \varphi_{is,K}'\alpha_K(\theta)\right)^2\right] \to 0$ as $K \to \infty$. In the second term, $\widehat{\alpha}$ is the slope in a regression with nS observations and K covariates, so we can show that $\|\widehat{\alpha}_K(\theta) - \alpha_K(\theta)\|^2 = O_P\left(\frac{K}{nS}\right)$. Moreover, since the approximating functions $\varphi_{is,K}$ are mean zero and uncorrelated across s, one can show that the square of the second term above is of order $O_P\left(\frac{1}{S} \times \frac{K}{nS}\right)$. Hence, we obtain

$$E_n\left[\left(\overline{\delta}_{iS}^{\mathcal{I}}(\theta)\right)^2\right] \leq O_P\left(\frac{1}{S}E\left[\left(\delta_{is}(\theta) - \varphi_{is,K}'\alpha_K(\theta)\right)^2\right]\right) + O_P\left(\frac{1}{S} \times \frac{K}{nS}\right) = o_P\left(\frac{1}{S}\right).$$

Thus, assuming that $\sqrt{n}/S$ is bounded is sufficient to satisfy equation (7) and hence for SM$\mathcal{I}$LE to achieve asymptotic efficiency. In contrast, SMLE requires $\sqrt{n}/S \to 0$, i.e., a faster rate of growth of S, to have no asymptotic bias (e.g., see Lee (1995)). We present a more precise relationship between S and n in the next section.

4 Large Sample Theory

4.1 General Lemma

Assumption 2. (i) *For some $C > 0$, W.p.a.1,* $\min_{i \leq n}\left(h_i(\theta_0) + \overline{\delta}_{iS}\right) \geq C > 0$;
(ii) $\|\nabla_\theta h_i(\theta_0)\|$ *is bounded.*

Assumption 3. *(i) The data $\{W_i\}_{i=1}^n$ is i.i.d. and has (possibly conditional) density $h(w, \theta_0)$;*

(ii) $\Theta \subset \mathbb{R}^p$ is compact, and $\theta_0 \in int(\Theta)$;

(iii) $\varepsilon_{is} \sim$ i.i.d. and are independent from $\{W_i\}_{i=1}^n$, and equation (2) holds;

(iv) $f(w, \varepsilon, \theta)$ is twice continuously differentiable in θ;

(v) [[the regularity conditions needed for consistency and asymptotic normality of the (infeasible) MLE estimator (3) hold; for normality conditions in the time series case, see Theorem 5.2 of Wooldridge handbook chapter.]].

Estimator $\widehat{\theta}_{\mathrm{SMLE}}$ solves

$$\widehat{\theta}_{\mathrm{SMLE}} \equiv \arg\max_{\theta\in\Theta} \widehat{Q}_n(\theta), \qquad \widehat{Q}_n(\theta) \equiv E_n\left[\ln\left(\overline{f}_S(X_i, \varepsilon_i, \theta)\right)\right].$$

and hence $\widehat{\theta}$ satisfies FOC:

$$0 = E_n \left[\nabla_\theta \ln \overline{f}_S \left(X_i, \varepsilon_i, \widehat{\theta} \right) \right].$$

To state a theorem that applies to both estimators, SMLE and SM$\mathcal{I}$LE, we introduce the following notation. Let $\widehat{\theta}$ denote either $\widehat{\theta}_{\text{SMLE}}$ or $\widehat{\theta}_{\text{SM}\mathcal{I}\text{LE}}$. Also, let $\ddot{f}_{iS}(\theta)$ denote either $\bar{f}_{iS}(\theta)$ or $\bar{f}^{\mathcal{I}}_{iS}(\theta)$, $\ddot{\delta}_{iS}(\theta) \equiv \ddot{f}_{iS}(\theta) - h_i(\theta)$, and $\ddot{\Delta}_{iS}(\theta) \equiv \nabla_\theta \ddot{\delta}_{iS}(\theta)$.

Lemma 1. *Suppose Assumptions 1 – ... and Assumption 2 hold and* $E_n \left[\ddot{\delta}^2_{iS} \right] + E_n \left[\left\| \ddot{\Delta}_{iS} \right\|^2 \right] = O_P(\gamma_n)$ *for all l and equation (7) holds. Then* $\sqrt{n} \left(\widehat{\theta} - \theta_0 \right) \to_d N(0, \Sigma^{-1})$, *where* $\Sigma \equiv E\left[\nabla_{\theta\theta} h_i(\theta_0) \right]$.

Remark 1. *Note that Lemma 1 applies to all estimators of the SMLE type. In particular, it applies to SM$\mathcal{I}$LE. What is crucial in the proof of the Lemma, is that the simulation errors satisfy* $E[\delta_i(\theta_0)] = 0$ *and* $E[\Delta_i(\theta_0)] = 0$, *so that the FOC of the estimator does not have a first-order bias due to simulation. The form of approximating functions* $\varphi_K(\cdot)$ *is taken specifically to ensure that the simulation errors satisfy this property.*

Thus, the key is to investigate conditions that ensure $\gamma_n = o_P\left(n^{-1/2}\right)$.

4.2 SMLE

Was discussed above.

4.3 SM$\mathcal{I}$LE

The current Section expands on the intuitive explanation of Section 3.2, and states the formal asymptotic results about the SM$\mathcal{I}$LE method.

Recall that $\varphi_{is} \equiv \varphi_{is,K} \equiv \varphi_K(W_i, \varepsilon_{is}) \in \mathbb{R}^K$. Let $\Omega_{\varphi\varphi} \equiv E[\varphi_{is}\varphi'_{is}]$. We will also require that

Assumption 4. $1/C \leq \lambda_{\min}(\Omega_{\varphi\varphi}) \leq \lambda_{\max}(\Omega_{\varphi\varphi}) \leq C$ *for some constant* $C > 0$.

Let $\omega_{is}(\theta) \equiv f_{is}(\theta) - h_i(\theta)$ and notice that $E[\omega_{is}(\theta) | W_i] = 0$. Using equation (8) we can writes

$$\alpha_K(\theta) \equiv \underset{a \in \mathbb{R}^K}{\operatorname{argmin}}\, E\left[\left(\omega_{is}(\theta) - \varphi'_{is,K} a \right)^2 \right] = E\left[\varphi_{is,K}\varphi'_{is,K} \right]^{-1} E\left[\varphi'_{is,K}\omega_{is}(\theta) \right].$$

Note also that $\operatorname{argmin}_{b \in \mathbb{R}^K} E\left[\left(\nabla_\theta \omega_{is}(\theta) - \varphi'_{is,K} b \right)^2 \right] = E\left[\varphi_{is,K}\varphi'_{is,K} \right]^{-1} E\left[\varphi'_{is,K} \nabla_\theta \omega_{is}(\theta) \right] = \nabla_\theta \alpha_K(\theta)$.

Assumption 5. *(i) As* $K \to \infty$, $\sup_{\theta \in \Theta} E\left[\left(\omega_{is}(\theta) - \varphi'_{is,K}\alpha_K(\theta) \right)^2 \right] = o(1)$.

(ii) For some positive constants τ,*and* τ_∇, *as* $K \to \infty$,

$$E\left[\left(\omega_{is}(\theta_0) - \varphi'_{is,K}\alpha_K(\theta_0) \right)^2 \right]^{1/2} = O\left(K^{-\tau}\right) \text{ and } E\left[\left(\nabla_\theta \omega_{is}(\theta_0) - \varphi'_{is,K}\nabla_\theta \alpha_K(\theta_0) \right)^2 \right]^{1/2} = O\left(K^{-\tau_\nabla}\right). \quad (9)$$

The value of τ in Assumption 5(ii) can usually be found as follows. Suppose f_{is} depends on $W_{ic} \in \mathbb{R}^{d_{W,c}}$ continuous observables ($d_{W,c} \leq \dim(W_i)$) and $\varepsilon_{is} \in \mathbb{R}^{d_\varepsilon}$. Suppose f_{is} has bounded continuous (cross-)derivatives of order k_f w.r.t. $(W_{ic}, \varepsilon_{is})$. Also, $(W_{ic}', \varepsilon_{is}')'$ has bounded support Then, $\tau = k_f / (d_c + d_\varepsilon)$ in equation (9). In practice, functions f_{is} are often infinitely smooth. If one uses cubic splines as approximating functions φ_K, one should take $\tau = \min\{k_f, 3\} / (d_c + d_\varepsilon)$. [2]

Remark 2. *Suppose* $f_{is}(\theta) \equiv f(W_i, \varepsilon_{is}, \theta) \equiv f\left(Y_i, \theta_{1:d_1}' X_{i,1:d_1}, X_{i,d_1+1:d_X}, \theta_{d_1+1:d_\theta}, \varepsilon_{is}\right)$, *i.e.,* f_{is} *depends on covariates* $X_{i,1:d_1} \equiv (X_{i,1}, \ldots, X_{i,d_1})'$ *only through the linear index* $\theta_{1:d_1}' X_{i,1:d_1}$. *If* Y_i *has finite support (i.e., is a discrete RV or a a vector of discrete RVs), then* $d_c \equiv \dim(\varepsilon_{is}) + (d_X - d_1) + 1$. *If* $Y_i \in \mathbb{R}^{d_Y}$ *is a vector of continuously distributed RVs, then* $d_c \equiv \dim(\varepsilon_{is}) + (d_X - d_1) + 1$.

Theorem 1. *Suppose Assumptions 1 – 5, ## hold, and* $K^2/(nS) \to 0$. *Then, for* $E_n\left[\left(\overline{\delta}_{iS}^{\mathcal{I}}\right)^2\right] + E_n\left[\left\|\overline{\delta}_{iS}^{\mathcal{I}} \overline{\Delta}_{iS}^{\mathcal{I}}\right\|\right] = O_P\left(\gamma_{n,SM\mathcal{I}LE}\right)$ *for all l with*

$$\gamma_{n,SM\mathcal{I}LE} = \frac{1}{S}\left(K^{-2\tau} + \frac{K}{nS}\right).$$

Remark 3. *The proof of the theorem is pointwise in* θ, *but the result extends to hold uniformly in* θ *in the usual fashion.*

The optimal choice of K is $K_n^* \equiv C(nS)^{1/(2\tau+1)}$, in which case

$$\gamma_{n,\text{SM}\mathcal{I}\text{LE}}^* = \frac{1}{S^{1+2\tau/(2\tau+1)} n^{2\tau/(2\tau+1)}} = \left(\frac{1}{S^{1+4\tau} n^{2\tau}}\right)^{\frac{1}{1+2\tau}}.$$

Hence, with the optimal choice of K, condition (7) holds when

$$\sqrt{n} = o\left(S^{1+\frac{6\tau}{1-2\tau}}\right) \quad \Leftrightarrow \quad S \Big/ n^{\frac{1-2\tau}{2+8\tau}} \to \infty. \tag{10}$$

□

Remark 4. *The standard SMLE condition* $\sqrt{n} = o(S)$ *obtains by taking* $\tau = 0$, *but since* $\tau \in (0, 1/2)$, *the restriction is weaker for SM$\mathcal{I}$LE estimator, i.e., condition (7) is satisfied when* S *grows at a slower rate (or, in practice, for smaller values of* S*). In particular, as long as* $K \to \infty$, *SM$\mathcal{I}$LE estimator is efficient and unbiased even if* $\sqrt{n}/S \to C \in (0, \infty)$, *a situation in which the SMLE estimator is known to be biased.*

Remark 5. *Remarkably, when* $\tau > 1/2$, *condition (10) is satisfied for* fixed S. *That is, SM$\mathcal{I}$LE can be asymptotically equivalent to MLE estimator (and in particular,* $\sqrt{n}$*-consistent, and asymptotically*

[2] Univariate cubic splines: for a grid of knot points $t_1 < t_2 < \ldots < t_m$ $(K = m + 4)$ let

$$\varphi_K(x) = \left(1, x, x^2, x^3, \max\{x - t_1, 0\}^3, \ldots, \max\{x - t_m, 0\}^3\right)'.$$

For example, one can take $t_j = Q_X\left(\frac{j}{m+1}\right)$.

If x is multivariate, take $\varphi_K(x) = \varphi_L(x_1) \otimes \varphi_L(x_2) \otimes \cdots \otimes \varphi_L(x_{d_X})$, so $K = d_X^L$.

See Chen (2007) for details on approximation and splines.

unbiased normal). In stark contrast, when S is finite, standard SMLE is inconsistent*. To the best of our knowledge, our paper is the first to show that it is possible to construct a version of SMLE that is asymptotically equivalent to MLE with finite S. We are only aware of one other estimator in the SMLE settings that is consistent with finite S – the estimator proposed in Lee and Song (2013); but their estimator has the rate of convergence $n^{-1/3}$.*

4.3.1 Adding Bias Correction

Recall $\ddot{\delta}_{iS}(\theta) \equiv \ddot{f}_{iS}(\theta) - h_i(\theta)$ If $E\left[\left(\ddot{\delta}_{iS}\right)^2\right] \to 0$ as $S, K, n \to \infty$, we can expand $E_n\left[\ln\left(\ddot{f}_{iS}(\theta)\right)\right]$ around $h_i(\theta)$

$$\ddot{Q}_n(\theta) \equiv E_n\left[\ln\left(\ddot{f}_{iS}(\theta)\right)\right] = E_n\left[\ln\left(h_i(\theta) + \ddot{\delta}_{iS}(\theta)\right)\right] = E_n\left[\ln(h_i(\theta)) + \frac{\ddot{\delta}_{iS}(\theta)}{h_i(\theta)} - \frac{1}{2}\frac{\ddot{\delta}_{iS}(\theta)^2}{h_i(\theta)^2}\right].$$

The second term on the right-hand side quickly converges to zero, and the third term is the source of bias and converges to

$$B_S(\theta) \equiv -\frac{1}{2}E\left[\frac{\ddot{\delta}_{iS}(\theta)^2}{h_i(\theta)^2}\right] = -\frac{1}{2S}E\left[\frac{\ddot{\delta}_{is}(\theta)^2}{h_i(\theta)^2}\right].$$

If $S \geq 2$ we can estimate $B_S(\theta)$ by

$$\ddot{B}_{n,S}(\theta) \equiv -\frac{1}{2S}E_n\left[\frac{\frac{1}{S-1}\sum_{s=1}^{S}\left(\ddot{f}_{is}(\theta) - \ddot{f}_{iS}(\theta)\right)^2}{\ddot{f}_{iS}(\theta)^2}\right].$$

Hence, the following estimator based on the bias-corrected criterion function could have lower bias:

$$\ddot{\theta}_{\text{BC}} = \underset{\theta\in\Theta}{\operatorname{argmin}}\,\ddot{Q}_{n,\text{BC}}(\theta), \qquad \ddot{Q}_{n,\text{BC}}(\theta) \equiv \ddot{Q}_n(\theta) - \ddot{B}_{n,S}(\theta).$$

A Appendix

Notation: "w.p.a.1" stands for "with probability approaching 1". $E_n[a_i] = \frac{1}{n}\sum_{i=1}^{n} a_i$, $E_{nS}[a_{is}] = \frac{1}{nS}\sum_{i=1}^{n}\sum_{s=1}^{S} a_{is}$.

A.1 Proof of Lemma 1

To complete the analysis of Section 3.1 we need to bound the term R_n in equation (5). Using the identity $\frac{A+a_n}{B+b_n} - \{\frac{A}{B} + \frac{1}{B}(a_n - \frac{A}{B}b_n)\} = -\frac{b_n(a_n - \frac{A}{B}b_n)}{B(b_n+B)}$ we have

$$R_n \equiv E_n[R_i],\ R_i \equiv -\frac{\bar{\delta}_{iS}\left(\overline{\Delta}_{iS} - \frac{\nabla_\theta h_i(\theta_0)}{h_i(\theta_0)}\bar{\delta}_{iS}\right)}{h_i(\theta_0)\left(h_i(\theta_0) + \bar{\delta}_{iS}\right)}. \tag{11}$$

By Assumption 2, w.p.a.1, $\min_{i\leq n}\left(h_i(\theta_0)+\overline{\delta}_{iS}\right)\geq C>0$, so

$$\|R_n\|\leq E_n\left[\|R_i\|\right]\leq CE_n\left[\left\|\overline{\delta}_{iS}\overline{\Delta}_{iS}\right\|+\left\|\frac{\nabla_\theta h_i(\theta_0)}{h_i(\theta_0)}\right\|\overline{\delta}_{iS}^2\right]$$

Also, $\|\nabla_\theta h_i(\theta_0)\|\leq C$ by Assumption 2, and hence

$$R_n\leq C\left(E_n\left[\overline{\delta}_{iS}^2\right]+E_n\left[\left\|\overline{\delta}_{iS}\overline{\Delta}_{iS}\right\|\right]\right).$$

Hence, $\sqrt{n}R_n=O_P\left(\sqrt{n}\gamma_n\right)=o_P(1)$, by the condition of the Lemma and equation (7). Hence, $\sqrt{n}\left(\widehat{\theta}-\theta_0\right)=\mathcal{I}^{-1}T_{0n}+o_P(1)\rightarrow_d N\left(0,\mathcal{I}^{-1}\right)$. ■

A.2 Proof of Theorem 1

1. The simulation errors are

$$\begin{aligned}\overline{\delta}_{iS}^{\mathcal{I}}(\theta) &\equiv \overline{f}_{iS}^{\mathcal{I}}(\theta)-h_i(\theta)=\frac{1}{S}\sum_{s=1}^{S}\left(f_{is}(\theta)-h_i(\theta)-\varphi_{is}'\widehat{\alpha}_K(\theta)\right)=\overline{f}_{iS}(\theta)-\overline{\varphi}_{iS}'\widehat{\alpha}_K(\theta)-h_i(\theta)\\ \overline{\Delta}_{iS}^{\mathcal{I}}(\theta) &\equiv \nabla_\theta\overline{\delta}_{iS}^{\mathcal{I}}(\theta)=\nabla_\theta\overline{f}_{iS}(\theta)-\overline{\varphi}_{iS}'\widehat{\beta}_K(\theta)-\nabla_\theta h_i(\theta),\end{aligned}$$

where $\overline{\varphi}_{iS}\equiv\frac{1}{S}\sum_{s=1}^{S}\varphi_{is}$.Remember that $\alpha_K(\theta)\equiv\arg\min_a E\left[\left(f_{is}(\theta)-\varphi_{is}'a\right)^2\right]$, and we estimate $\widehat{\alpha}_K(\theta)$ by the regression $\widehat{\alpha}_K(\theta)\equiv\left(\mathcal{V}_{\varphi_K}'\mathcal{V}_{\varphi_K}\right)^{-1}\mathcal{V}_{\varphi_K}'\mathcal{V}_{f(\theta)}$. We do not estimate $\widehat{\beta}_K(\theta)$ explicitly, rather, $\widehat{\beta}_K(\theta)=\nabla_\theta\widehat{\alpha}_K(\theta)$.

Then

$$\begin{aligned}\overline{f}_{iS}^{\mathcal{I}}(\theta) &\equiv \frac{1}{S}\sum_{s=1}^{S}\left(f_{is}(\theta)-\varphi_{is}'\widehat{\alpha}_K(\theta)\right)\\ &= \frac{1}{S}\zeta_{iS}'\left(\mathcal{V}_{f(\theta)}-\mathcal{V}_{\varphi_K}\widehat{\alpha}_K(\theta)\right)=\frac{1}{S}\zeta_{iS}'\left(I_{nS}-P_{\varphi_K}\right)\mathcal{V}_{f(\theta)},\end{aligned}$$

where $\zeta_{iS}\equiv\left(0_{1\times(i-1)S},1_{1\times S},0_{1\times(n-i)S}\right)'$ (equivalently, $\zeta_{iS}\equiv e_i\otimes 1_{S\times 1}$), and $P_{\varphi_K}\equiv\mathcal{V}_{\varphi_K}\left(\mathcal{V}_{\varphi_K}'\mathcal{V}_{\varphi_K}\right)^{-1}\mathcal{V}_{\varphi_K}'$

Write $\omega_{is}(\theta)\equiv f_{is}(\theta)-h_i(\theta)$, so $E\left[\omega_{is}(\theta)|W_i\right]=0$. Then

$$\overline{\delta}_{iS}^{\mathcal{I}}(\theta)\equiv\overline{f}_{iS}^{\mathcal{I}}(\theta)-h_i(\theta)=\frac{1}{S}\sum_{s=1}^{S}\left(f_{is}(\theta)-h_i(\theta)-\varphi_{is}'\widehat{\alpha}_K(\theta)\right)=\frac{1}{S}\sum_{s=1}^{S}\left(\omega_{is}(\theta)-\varphi_{is}'\widehat{\alpha}_K(\theta)\right).$$

Note that for any $a\in\mathbb{R}^K$,

$$\begin{aligned}E\left[\left(f_{is}(\theta)-\varphi_{is}'a\right)^2\right] &= E\left[\left(h_i(\theta)+\omega_{is}(\theta)-\varphi_{is}'a\right)^2\right]\\ &= E\left[\left(\omega_{is}(\theta)-\varphi_{is}'a\right)^2\right]+E\left[h_i(\theta)^2+2h_i(\theta)\left(\omega_{is}(\theta)-\varphi_{is}'a\right)\right]\\ &= E\left[\left(\omega_{is}(\theta)-\varphi_{is}'a\right)^2\right]+E\left[h_i(\theta)^2\right],\end{aligned}$$

so $\alpha_K(\theta)$ can be equivalently written as

$$\alpha_K(\theta) = \operatorname*{argmin}_a E\left[\left(\omega_{is}(\theta) - \varphi'_{is}a\right)^2\right]. \tag{12}$$

2. We can write

$$\overline{\delta}^{\mathcal{I}}_{iS}(\theta) = \frac{1}{S}\sum_{s=1}^{S}\left(\omega_{is}(\theta) - \varphi'_{is}\alpha_K(\theta)\right) + \frac{1}{S}\sum_{s=1}^{S}\varphi'_{is}\left(\alpha_K(\theta) - \widehat{\alpha}_K(\theta)\right) = \overline{\delta}^{\mathcal{I}}_{iS,A}(\theta) + \overline{\delta}^{\mathcal{I}}_{iS,B}(\theta)$$

and hence

$$E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS}(\theta)\right)^2\right] \le 2E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,A}(\theta)\right)^2\right] + 2E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,B}(\theta)\right)^2\right].$$

Here

$$E\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,A}(\theta)\right)^2\right] = \frac{1}{S}E\left[\left(\omega_{is}(\theta) - \varphi'_{is}\alpha_K(\theta)\right)^2\right] \le C\frac{1}{S}K^{-2\tau},$$

where the last inequality holds by Assumption 5. Hence, by Markov inequality $E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,A}(\theta)\right)^2\right] = O_P\left(\frac{1}{S}K^{-2\tau}\right)$. Next,

$$\begin{aligned} E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,B}(\theta)\right)^2\right] &= E_n\left[\left(\overline{\varphi}'_{iS}\left(\alpha_K(\theta) - \widehat{\alpha}_K(\theta)\right)\right)^2\right] \\ &= \left(\alpha_K(\theta) - \widehat{\alpha}_K(\theta)\right)' E_n\left[\overline{\varphi}_{iS}\overline{\varphi}'_{iS}\right]\left(\alpha_K(\theta) - \widehat{\alpha}_K(\theta)\right). \end{aligned}$$

Here $\lambda_{\max}\left(E_n\left[\overline{\varphi}_{iS}\overline{\varphi}'_{iS}\right]\right) \le_{\text{w.p.a.1}} \frac{1}{S}\left(\lambda_{\max}\left(E\left[\varphi_{is}\varphi'_{is}\right]\right) + 1\right) \le \frac{C}{S}$, so w.p.a.1

$$E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS,B}(\theta)\right)^2\right] \lesssim \frac{1}{S}\left\|\widehat{\alpha}_K(\theta) - \alpha_K(\theta)\right\|^2.$$

Hence

$$E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS}(\theta)\right)^2\right] = O_P\left(\frac{1}{S}\left(K^{-2\tau} + \left\|\widehat{\alpha}_K(\theta) - \alpha_K(\theta)\right\|^2\right)\right). \tag{13}$$

3. Let $\eta_{is}(\theta) \equiv f_{is}(\theta) - \varphi'_{is}\alpha_K(\theta)$, so $\widehat{\alpha}_K(\theta) - \alpha_K(\theta) = \left(\mathcal{V}'_{\varphi_K}\mathcal{V}_{\varphi_K}\right)^{-1}\mathcal{V}'_{\varphi_K}\mathcal{V}_\eta$, where $E\left[\varphi_{is}\eta_{is}(\theta)\right] = 0$ by equation (4). $\widehat{\alpha}_K(\theta)$ is the vector of OLS coefficients in the regression with nS observations, hence $\left\|\widehat{\alpha}_K(\theta) - \alpha_K(\theta)\right\|^2 = O_P\left(\frac{K}{nS}\right)$, since $E\left[\left\|\varphi_{is}\eta_{is}(\theta)\right\|^2\right] \le C$ [[#cond: $\left\|\varphi(w,e)\right\|_\infty \le C$]], and hence

$$E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS}(\theta)\right)^2\right] = O_P\left(\frac{1}{S}K^{-2\tau} + \frac{K}{nS^2}\right) = O_P\left(\frac{1}{S}\left(K^{-2\tau} + \frac{K}{nS}\right)\right).$$

4. Note that $\operatorname{argmin}_{b\in\mathbb{R}^K} E_{nS}\left[\left(\nabla_\theta\omega_{is}(\theta) - \varphi'_{is,K}b\right)^2\right] = E_{nS}\left[\varphi_{is,K}\varphi'_{is,K}\right]^{-1}E_{nS}\left[\varphi'_{is,K}\nabla_\theta\omega_{is}(\theta)\right] = \nabla_\theta\hat{\alpha}_K(\theta)$. Thus, the arguments of parts 2-3 hold with $\overline{\delta}^{\mathcal{I}}_{iS}$ replaced by $\overline{\Delta}^{\mathcal{I}}_{iS}$, giving $E_n\left[\left(\overline{\delta}^{\mathcal{I}}_{iS}(\theta)\right)^2\right] = O_P\left(\frac{1}{S}\left(K^{-2\tau_\nabla} + \frac{K}{nS}\right)\right)$. Thus,

$$\gamma_{n,\text{SM}\mathcal{I}\text{LE}} \equiv \frac{1}{S}\left(K^{-\tau-\min(\tau,\tau_\nabla)} + K^{-\min(\tau,\tau_\nabla)}\sqrt{\frac{K}{nS}} + \frac{K}{nS}\right).$$

When $\tau_\nabla = \tau$ we can simply set $\gamma_{n,\text{SM}\mathcal{I}\text{LE}} \equiv \frac{1}{S}\left(K^{-2\tau} + \frac{K}{nS}\right)$.

5. Consider term T_{1n} in equation (5) for SM$\mathcal{I}$LE. Let $T^{\mathcal{I}}_{1n} \equiv \sqrt{n}E_n\left[\xi^{\mathcal{I}}_i\right]$, where $\xi^{\mathcal{I}}_i \equiv \frac{\overline{\Delta}^{\mathcal{I}}_{iS}}{h_i(\theta_0)} -$

$\frac{\nabla_\theta h_i(\theta_0)}{h_i^2(\theta_0)}\overline{\delta}_{iS}^{\mathcal{I}}$. Write

$$\begin{aligned}
& E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\overline{\delta}_{iS}^{\mathcal{I}}\right] \\
= \;& E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\left(\overline{\delta}_{iS,A}^{\mathcal{I}}+\overline{\delta}_{iS,B}^{\mathcal{I}}\right)\right] \\
= \;& E_{nS}\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\left(\omega_{is}\left(\theta_0\right)-\varphi_{is}'\alpha_K\left(\theta_0\right)\right)\right]+E_{nS}\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\varphi_{is}'\right]\left(\alpha_K\left(\theta_0\right)-\widehat{\alpha}_K\left(\theta_0\right)\right) \\
= \;& A_{1n}+A_{2n}.
\end{aligned}$$

Here, using Assumptions 2, 5

$$E\left[\left(\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\left(\omega_{is}\left(\theta_0\right)-\varphi_{is}'\alpha_K\left(\theta_0\right)\right)\right)^2\right]\leq CE\left[\left(\omega_{is}\left(\theta_0\right)-\varphi_{is}'\alpha_K\left(\theta_0\right)\right)^2\right]\leq CK^{-2\tau},$$

so $A_{1n}=O_P\left(\frac{1}{\sqrt{nS}}K^{-\tau}\right)$. Also, $E_{nS}\left[\frac{\nabla_\theta h_i(\theta_0)}{h_i^2(\theta_0)}\varphi_{is}'\right]=O_P\left(\sqrt{\frac{K}{nS}}\right)$, and hence $A_{2n}=O_P\left(\frac{K}{nS}\right)$. Thus,

$$\sqrt{n}E_n\left[\frac{\nabla_\theta h_i\left(\theta_0\right)}{h_i^2\left(\theta_0\right)}\overline{\delta}_{iS}^{\mathcal{I}}\right]=O_P\left(\frac{1}{\sqrt{S}}K^{-\tau}+\frac{K}{\sqrt{n}S}\right).$$

Same arguments apply to

.

■

A.3 Example: Discrete Outcome

For example, consider the case of discrete outcome $Y_i \in \{1, \ldots, L\}$. The observables are $W_i \equiv (Y_i, X_i')'$, where X_i is the vector of covariates X_i. The notation in part follows Lee (1992). Let $D_{il} \equiv 1\{Y_i = l\}$, $l = 1, \ldots, L$; and let $f_{is,l}(\theta) \equiv f_l(X_i, \varepsilon_{is}, \theta)$ denote the choice probability of alternative l given X_i and ε_{is}, with $p_{i,l}(\theta) \equiv p_l(X_i, \theta) \equiv E[f_{is,l}(\theta)|X_i]$ being the choice probability of alternative l given X_i, so

$$\begin{aligned} h_i(\theta) &= p_{i,1}^{D_{i,1}}(\theta) \cdot p_{i,2}^{D_{i,2}}(\theta) \cdot \ldots p_{i,L}^{D_{i,L}}(\theta) = \sum_{l=1}^{L} D_{i,l} p_{i,l}(\theta), \text{ and} \\ \log(h_i(\theta)) &= \sum_{l=1}^{L} D_{i,l} \log(p_{i,l}(\theta)). \end{aligned}$$

Likewise,

$$\begin{aligned} f_{is}(\theta) &\equiv \sum_{l=1}^{L} D_{i,l} f_{is,l}(\theta), \ f_{is,l}(\theta) \equiv f_l(X_i, \varepsilon_{is}, \theta), \\ \overline{f}_{iS}(\theta) &= \frac{1}{S}\sum_{s=1}^{S} f_{is}(\theta) = \sum_{l=1}^{L} D_{i,l}\left(\frac{1}{S}\sum_{s=1}^{S} f_{is,l}(\theta)\right), \\ \overline{f}_{iS,l}(\theta) &\equiv \frac{1}{S}\sum_{s=1}^{S} f_{is,l}(\theta), \\ \log \overline{f}_{iS}(\theta) &= \sum_{l=1}^{L} D_{i,l} \log\left(\frac{1}{S}\sum_{s=1}^{S} f_{is,l}(\theta)\right). \end{aligned}$$

B Monte Carlo

The current section investigates the finite sample properties of the new estimator. We use the Mixed Logit model Monte Carlo design from from Kristensen and Salanié (2017). The econometrician observes i.i.d. draws (X_i, Y_i) for $i = 1, ..., n$, with X_i a centered normal of variance τ^2, and

$$Y_i = 1\left\{\theta_3 + (\theta_1 + \theta_2 \varepsilon_i) X_i > \eta_i\right\},$$

where η_i has a logistic distribution, ε_i is standard normal, and X_i, η_i, and ε_i are mutually independent. We set $n = 10000$, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 0$, and use two values for the parameter τ, $\tau = 1$ or 2. The number of replications is 1000. Notice that in this model, the choice probability takes the form of an integral

$$P(Y_i = 1|X_i; \theta) = \int \frac{\phi(\varepsilon)}{1 + \exp(-(\theta_3 + (\theta_1 + \theta_2 \varepsilon) x))} d\varepsilon,$$

so $h_i(\theta) \equiv P(Y_i = 1|X_i; \theta)^{Y_i} (1 - P(Y_i = 1|X_i; \theta))^{1-Y_i}$, while the p.m.f. of Y_i given both X_i and ε_i has the explicit expression:

$$f(Y_i|X_i, \varepsilon_i; \theta) = \left[\frac{1}{1 + \exp(-\theta_3 - (\theta_1 + \theta_2 \varepsilon_i) X_i)}\right]^{Y_i} \left[1 - \frac{1}{1 + \exp(-\theta_3 - (\theta_1 + \theta_2 \varepsilon_i) X_i)}\right]^{1-Y_i}.$$

A closed-form expression for the choice probabilities is not available, which motivates the use of a simulation-based methods. Following Kristensen and Salanié (2017), we also include the MLE estimator with h approximated by quadratures as a benchmark.

The SMLE and SM$\mathcal{I}$LE use the following estimators of the density of the density $h(Y_i|X_i, \theta)$:

$$\overline{f}_{iS}(\theta) = \frac{1}{S} \sum_{s=1}^{S} f_{is}(\theta) \quad \text{and} \quad \overline{f}_{iS}^{\mathcal{I}}(\theta) = \frac{1}{S} \sum_{s=1}^{S} \left(f_{is}(\theta) - \varphi_{is}' \widehat{\alpha}(\theta)\right),$$

where we choose

$$\varphi_{is} = \begin{pmatrix} 1 \\ Y_i \end{pmatrix} \otimes \begin{pmatrix} 1 \\ X_i \\ X_i^2 \end{pmatrix} \otimes \begin{pmatrix} \varepsilon_{is} \\ \varepsilon_{is}^2 - 1 \\ 1\{\varepsilon_{is} \leq 0\} - \frac{1}{2} \end{pmatrix}.$$

Tables 1 - 4 contain the results for $\tau = 1$, and Tables 5 - 8 contain the results for $\tau = 2$. The performance of SM$\mathcal{I}$LE is clearly superior to that of SMLE. For S = 50, 100, and 200, it slightly increases the standard deviation of the estimator, but the large reduction in bias offsets that, and leads to substantially smaller RMSE's. For $S = 500$, the performance of the estimators is virtually identical.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.259	-0.125	0.006	-0.870	-0.313	-0.000	-0.000	-0.000
STD	0.074	0.026	0.061	0.177	0.059	0.180	0.024	0.022	0.023
RMSE	0.074	0.260	0.139	0.177	0.872	0.361	0.024	0.022	0.023

Table 1: Simulation results with $S = 10$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.208	-0.059	0.006	-0.635	-0.144	-0.000	-0.000	-0.000
STD	0.074	0.050	0.063	0.177	0.183	0.155	0.024	0.022	0.024
RMSE	0.074	0.214	0.086	0.177	0.660	0.212	0.024	0.022	0.024

Table 2: Simulation results with $S = 25$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.132	-0.029	0.006	-0.369	-0.073	-0.000	-0.000	-0.000
STD	0.074	0.060	0.067	0.177	0.168	0.161	0.024	0.023	0.024
RMSE	0.074	0.146	0.073	0.177	0.405	0.177	0.024	0.023	0.024

Table 3: Simulation results with $S = 50$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.075	-0.014	0.006	-0.203	-0.036	-0.000	-0.000	-0.000
STD	0.074	0.066	0.070	0.177	0.164	0.168	0.024	0.023	0.024
RMSE	0.074	0.100	0.072	0.177	0.261	0.172	0.024	0.023	0.024

Table 4: Simulation results with $S = 100$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.039	-0.005	0.006	-0.105	-0.014	-0.000	-0.000	-0.000
STD	0.074	0.070	0.072	0.177	0.168	0.172	0.024	0.024	0.024
RMSE	0.074	0.079	0.072	0.177	0.198	0.172	0.024	0.024	0.024

Table 5: Simulation results with $S = 200$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.004	-0.014	0.001	0.006	-0.039	-0.001	-0.000	-0.000	-0.000
STD	0.074	0.073	0.073	0.177	0.174	0.175	0.024	0.024	0.024
RMSE	0.074	0.074	0.073	0.177	0.178	0.175	0.024	0.024	0.024

Table 6: Simulation results with $S = 500$ and $\tau = 1$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.406	-0.280	0.020	-0.729	-0.402	0.001	0.000	0.001
STD	0.068	0.033	0.080	0.112	0.077	0.129	0.029	0.024	0.028
RMSE	0.069	0.407	0.292	0.113	0.733	0.423	0.029	0.024	0.028

Table 7: Simulation results with $S = 10$ and $\tau = 2$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.238	-0.136	0.020	-0.397	-0.196	0.001	0.000	0.000
STD	0.068	0.046	0.067	0.112	0.078	0.100	0.029	0.026	0.028
RMSE	0.069	0.243	0.152	0.113	0.404	0.220	0.029	0.026	0.028

Table 8: Simulation results with $S = 25$ and $\tau = 2$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.136	-0.068	0.020	-0.223	-0.099	0.001	0.000	0.000
STD	0.068	0.054	0.061	0.112	0.090	0.096	0.029	0.027	0.028
RMSE	0.069	0.147	0.091	0.113	0.241	0.138	0.029	0.027	0.028

Table 9: Simulation results with $S = 50$ and $\tau = 2$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.072	-0.030	0.020	-0.117	-0.045	0.001	0.001	0.001
STD	0.068	0.060	0.062	0.112	0.098	0.100	0.029	0.028	0.029
RMSE	0.069	0.093	0.069	0.113	0.153	0.110	0.029	0.028	0.029

Table 10: Simulation results with $S = 100$ and $\tau = 2$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.036	-0.013	0.020	-0.058	-0.019	0.001	0.001	0.001
STD	0.068	0.063	0.064	0.112	0.103	0.104	0.029	0.029	0.029
RMSE	0.069	0.072	0.065	0.113	0.118	0.106	0.029	0.029	0.029

Table 11: Simulation results with $S = 200$ and $\tau = 2$.

	θ_1			θ_2			θ_3		
	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE	MLE	SMLE	SM$\mathcal{I}$LE
BIAS	0.010	-0.012	-0.002	0.020	-0.019	-0.003	0.001	0.001	0.001
STD	0.068	0.065	0.066	0.112	0.107	0.107	0.029	0.029	0.029
RMSE	0.069	0.066	0.066	0.113	0.108	0.107	0.029	0.029	0.029

Table 12: Simulation results with $S = 500$ and $\tau = 2$.

References

Armstrong, T., R. Gallant, H. Hong, and H. Li (2013): "The Asymptotic Distribution of Estimators with Overlapping Simulation Draws," *Working Paper*, 1–27.

Chen, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5549–5632.

Danielsson, J. (1994): "Stochastic volatility in asset prices: Estimation with simulated maximum likelihood," *Journal of Econometrics*, 64, 375–400.

Durbin, J. and S. J. Koopman (2000): "Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives," *Journal of the Royal Statistical Society Series B*, 62, 3–56.

Evdokimov, K. S. (2014): "Efficient Estimation with a Finite Number of Simulation Draws per Observation," *Working Paper, Princeton University.*

Gourieroux, C. and A. Monfort (1993): "Simulation-based inference," *Journal of Econometrics*, 59, 5–33.

Gourieroux, C., A. Monfort, and E. Renault (1993): "Indirect inference," *Journal of Applied Econometrics*, 8, S85–S118.

Hajivassiliou, V. (1999): "Some Practical Issues in Maximum Simulated Likelihood," *WP*.

Heiss, F. and V. Winschel (2008): "Likelihood approximation by numerical integration on sparse grids," *Journal of Econometrics*, 144, 62–80.

Kristensen, D. and B. Salanié (2017): "Higher-order properties of approximate estimators," *Journal of Econometrics*, 198, 189–208.

Lee, D. and K. Song (2013): "Simulated maximum likelihood estimation for discrete choices using transformed simulated frequencies," *WP*, 1–49.

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.

Lerman, S. R. and C. F. Manski (1981): *On the Use of Simulated Frequencies to Approximate Choice Probabilities*, The MIT Press.

Morellec, E., B. Nikolov, and N. Schürhoff (2012): "Corporate Governance and Capital Structure Dynamics," *The Journal of Finance*, 67, 803–848.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, IV.

Roşu, I. (2009): "A Dynamic Model of the Limit Order Book," *Review of Financial Studies*, 22, 4601–4641.

Sandmann, G. and S. J. Koopman (1998): "Estimation of stochastic volatility models via Monte Carlo maximum likelihood," *Journal of Econometrics*, 87, 271–301.

Stern (1997): "Simulation-Based Estimation," *Journal of economic literature*, 35, 2006–2039.