SIMULATION-BASED INFERENCE

Michal Kolesár*

April 25, 2024

In 2024, Princeton flew out 14 junior candidates. Five papers were mostly concerned with theory (economic or econometric), and of the remaining nine papers, 5 featured simulation-based inference.[1] In previous years, the fraction of papers using the methods we discuss in this note is similarly around a third.

1. SIMULATION APPROACHES

Suppose we have i.i.d. data $Z_i = (Y_i, X_i)$, $i = 1, \dots, n$, where X_i are exogenous covariates, and the conditional distribution of the outcome Y_i given X_i is known up to a finite-dimensional parameter θ. For concreteness, suppose that this model has the form

$$Y_i = m(X_i, \epsilon_i, \theta), \tag{1}$$

where the distribution F of the vector ϵ_i is known. The outcome Y_i may be vector-valued, so that the setup is general enough so that Y_i may be a vector of equilibrium outcomes, or a time series path of outcomes in a dynamic panel model.

In principle, we should be able to write down the (conditional on X_i) likelihood, and estimate the model using maximum likelihood estimator (MLE). But in models of the form (1), evaluating the likelihood involves integration. If we can't evaluate the integral analytically, doing so numerically is hard if ϵ_i is high-dimensional (say $\dim(\epsilon_i) \geq 5$). There are a lot of examples: models with missing data, models with latent (unobserved) variables, or complicated discrete choice models.

Sometimes even simple-looking discrete choice models end up being surprisingly hard to estimate by likelihood methods:

Example 1 (Multinomial discrete choice). Suppose that the utility from choice j is given by $Y_{ij}^* = X_{ij}'\beta + U_{ij}$, $j = 0, \dots, J$, with the distribution of $U_i = (U_{i0}, \dots, U_{iJ})$ known up to its

*Email: mkolesar@princeton.edu.

1. These were Tim de Silva (finance) "Insurance versus Moral Hazard in Income-Contingent Student Loan Repayment", Aleksei Oskolkov (international finance) "Heterogeneous Impact of the Global Financial Cycle", Anna Russo (environmental, with K. M. Aspelund) "Additionality and Asymmetric Information in Environmental Markets", Charlie Rafkin (public, with Evan Soltas) "Eviction as Bargaining Failure: Hostility and Misperceptions in the Rental Housing Market", and Evan Soltas (public) "Tax Incentives and the Supply of Low-Income Housing".

covariance Σ. We can write this as $U_i = A\epsilon_i$, $\epsilon_i \sim F$, and $\Sigma = AA'$. Let a_j' denote the jth row of A, so that we can equivalently write $U_{ij} = a_j'\epsilon_i$. Let Y_i denote the $(J+1)$-vector of zeros with one in the position of the observed choice. That is, $Y_{ij} = 1$ if $Y_{ij}^* \geq \max_\ell Y_{i\ell}^*$ (assuming no ties), and $Y_{ij} = 0$ otherwise. This is a special case of (1), with vector outcome Y_i, and

$$m_j(X_i, \epsilon_i, \theta) = \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta + a_j'\epsilon_i \geq a_\ell'\epsilon_i\}.$$

The log-likelihood is given by $\sum_{i=1}^n \sum_{j=0}^J \mathbb{1}\{Y_{ij} = 1\} \log P_\theta(Y_{ij} = 1 \mid X_i)$. To build the likelihood, we therefore need to calculate all conditional choice probabilitys (CCPs), $P_\theta(Y_{ij} = 1 \mid X_i)$. Unfortunately,

$$P_\theta(Y_{ij} = 1 \mid X_i) = \int \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\epsilon\} f(\epsilon)\, \mathrm{d}\epsilon$$

is a J-dimensional integral that you don't want to evaluate numerically unless J is very small.

In some special cases, we do have analytic solutions. For example, consider the multinomial logit model, in which $\epsilon_{ij} \sim F(u)$, with $F(u) = e^{-e^{-u}}$. This distribution is called *Gumbel distribution*, or *Type-I extreme value distribution*[2]. Thus, $\Sigma = I$, so that a_i is the unit vector, ϵ_{ij} has density $e^{-u}e^{-e^{-u}}$. Then

$$\begin{aligned}
P_\theta(Y_{i0} = 1 \mid X_i) &= \int_{-\infty}^{\infty} \prod_{\ell=1}^{J} P(\epsilon_\ell \leq (X_{i0} - X_{i\ell})'\beta + \epsilon_0 \mid \epsilon_0) e^{-\epsilon_0} e^{-e^{-\epsilon_0}}\, \mathrm{d}\epsilon_0 \\
&= \int_{-\infty}^{\infty} e^{-e^{-(X_{i0}-X_{i1})'\beta - \epsilon_0}} \cdots e^{-e^{-(X_{i0}-X_{iJ})'\beta - \epsilon_0}} e^{-\epsilon_0} e^{-e^{-\epsilon_0}}\, \mathrm{d}\epsilon_0 \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\epsilon_0}[1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{iJ})'\beta}]\right) e^{-\epsilon_0}\, \mathrm{d}\epsilon_0 \\
&= \frac{1}{1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{iJ})'\beta}} = \frac{e^{X_{i0}'\beta}}{\sum_{j=0}^J e^{X_{ij}'\beta}},
\end{aligned}$$

where the last line follows from

$$\int_{-\infty}^{\infty} e^{-\epsilon} e^{-e^{-\epsilon-c}}\, \mathrm{d}\epsilon = \int_{-\infty}^{\infty} e^{-u+c} e^{-e^{-u}}\, \mathrm{d}u = e^c \int_{-\infty}^{\infty} e^{-u} e^{-e^{-u}} = e^c.$$

by setting $c = -\log(1 + e^{-(X_{i0}-X_{i1})'\beta} + \cdots + e^{-(X_{i0}-X_{i1})'\beta})$.

But what if $\epsilon \sim \mathcal{N}_{J+1}(0, I)$ and we don't restrict Σ?[3] The likelihood for this multinomial probit model is hard to evaluate unless J is very small. ⊠

2. The names come from the fact that if X_i are i.i.d. exponential (so that the cumulative distribution function (CDF) is $1 - e^{-x}$), and we let $M_n = \max_{i \leq n} X_i$ denote the maximum, then Emil Julius Gumbel showed that $M_n - \log(n) \Rightarrow F(u)$. Indeed, $P(M_n - \log(n) \leq u) = (1 - e^{-u}/n)^n \to e^{-e^{-u}}$.

3. The covariance matrix Σ needs some normalization beyond requiring that it be symmetric positive semi-definite; we'll come back to this issue in Section 2.3.

Example 2 (Random coefficients logit). The multinomial logit model is restrictive: it implies independence of irrelevant alternatives, and it may be able to match the observed CCPs. McFadden and Train (2000) showed that if we allow the coefficients β to be random, the model can match the CCPs generated by *any* discrete choice random utility model. Such a model is called a random coefficients multinomial logit, or sometimes mixed logit, since the choice probabilities are mixtures of logit probabilities. In other words, suppose that each individual draws their own coefficients from some distribution G_θ $\beta_i \sim G_\theta$. Then $P_\theta(Y_{ik} = 1 \mid X_i) = \int \frac{e^{X_{ik}'\beta}}{\sum_{j=0}^J e^{X_{ij}'\beta}} dG_\theta(\beta)$. If G_θ and X_i is sufficiently flexible, we can match any observed choice patterns. A tractable choice for the mixing distribution G_θ is $\beta \sim \mu + \Lambda\epsilon$, where Λ is an $K \times L$ matrix of "factor loadings" onto factors ϵ, distributed $\mathcal{N}(0, I_K)$, say. Then the CCP becomes an L-dimensional integral,

$$P_\theta(Y_{ik} = 1 \mid X_i) = \int \frac{e^{X_{ik}'\mu + X_{ik}'\Lambda\epsilon}}{\sum_{j=0}^J e^{X_{ij}'\mu + X_{ij}'\Lambda\epsilon}} dF(\epsilon). \tag{2}$$

⊠

Example 3 (Panel probit). Another simple model that's non-trivial to estimate is the panel Probit model. For simplicity, suppose there are only two alternatives, so that $Y_{it}^* = X_{ij}'\beta + \epsilon_{it}$, and we observe $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, with $Y_{it} = \mathbb{1}\{Y_{it}^* \geq 0\}$. In general, we'd expect ϵ_{it} to be correlated over time, since factors that are not observed by the researcher can persist over time. Then with the autocovariance structure unrestricted, the likelihood for observing the sequence Y_i is a T-dimensional integral. ⊠

1.1. *Simulated method of moments*

The classic reference is McFadden (1989), see also Pakes and Pollard (1989). Notice that (1) and iterated expectations imply that for any function $s(z)$, we have the moment condition

$$E\left[g(Z_i, \theta_0)\right] = 0, \quad g(Z_i, \theta) = s(Z_i) - E_{\theta_0}[s(Z_i) \mid X_i] = s(Z_i) - \int s(X_i, m(X_i, \epsilon, \theta_0))\, \mathrm{d}F(\epsilon).$$

For example, we may take $s(z) = yg(x)$, so that $g(Z_i, \theta) = (Y_i - E[Y_i \mid X_i])g(X_i)$. The integral may be hard to evaluate. If we could evaluate it, we would be able to use the generalized method of moments (GMM) estimator

$$\hat{\theta}_{\text{GMM}} = \operatorname*{argmin}_\theta \frac{1}{n}\sum_i g(Z_i, \theta)' \hat{W}_n \frac{1}{n}\sum_i g(Z_i, \theta),$$

with asymptotic distribution (under regularity conditions)

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}),$$

where $G = E[\frac{\partial}{\partial\theta} g(Z_i, \theta_0)]$ and

$$\Omega = \text{var}(g(Z_i, \theta_0)) = E[\text{var}_{\theta_0}(s(Z_i) \mid X_i)].$$

The key idea of simulated method of moments (SMM) is that we can replace the hard-to-evaluate integral with an unbiased estimate of it. In particular, simulate M samples of $\{\tilde{\epsilon}_1^m, \ldots, \tilde{\epsilon}_n^m\}_{m=1}^M$ from the known distribution of ϵ, and replace the integral (which is an expectation) by the sample average based on the simulated samples, generating the moment

$$\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta) = s(Z_i) - \frac{1}{M} \sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta)), \tag{3}$$

where $\tilde{\epsilon}_i = (\epsilon_i^1, \ldots, \tilde{\epsilon}_i^M)$ is a vector of simulation samples for the ith observation. We then apply GMM to this moment condition,

$$\hat{\theta}_{\text{SMM}} = \frac{1}{n} \sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)' \hat{W} \frac{1}{n} \sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta).$$

One issue (that we'll come back to) is that if Y_i is discrete as in discrete choice models, $\tilde{g}$ is not differentiable in θ, which creates issues both for minimizing the sample GMM objective function[4], for proving normality, and for estimating the asymptotic variance. Using empirical process theory, it is possible to nonetheless show:

Proposition 1. Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_{SMM} - \theta) \Rightarrow \mathcal{N}(0, (G'WG)^{-1} G'W\Omega_{SMM} WG(G'WG)^{-1}),$$

where (using differentiation under the integral sign)

$$G = \frac{\partial}{\partial\theta} E[\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)] = E\left[\frac{\partial}{\partial\theta} g(Z_i, \theta_0)\right].$$

and

$$\Omega_{SMM} = \text{var}(\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)) = \left(1 + \frac{1}{M}\right) \Omega.$$

So the gradient G in the asymptotic variance is the same as in the GMM case, but simulation increases the variance of the moment condition.[5] The formula for Ω_{SMM} obtains

4. One needs to use non-derivative based methods, such as Nelder-Mead, but minimization can be challenging unless $\dim(\theta)$ is small.

5. Because the simulation draws are ancillary (like the precision of the scale in the example in Cox (1958)), one can make the argument that we should do inference conditional them. In that case, there is no noise, but there is bias! Will need $M \to \infty$ to get consistency.

because the simulation draws are independent of each other and of the data, so that

$$\begin{aligned}
&\operatorname{var}(\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta_0)) \\
&\quad = E\left[\operatorname{var}\left(g(Z_i, \theta_0) + E_{\theta_0}(s(Z_i) \mid X_i) - \frac{1}{M}\sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta_0))\right) \;\middle|\; X_i\right] \\
&\quad = E[\operatorname{var}(g(Z_i, \theta_0)) \mid X_i] + E\left[\operatorname{var}(\frac{1}{M}\sum_{m=1}^{M} s(X_i, m(X_i, \tilde{\epsilon}_i^m, \theta_0)) \;\middle|\; X_i)\right] \\
&\quad = E[\operatorname{var}(g(Z_i, \theta_0)) \mid X_i] + \frac{1}{M} E\left[\operatorname{var}(g(Z_i, \theta_0) \mid X_i)\right].
\end{aligned}$$

You can see that more generally, if say the simulation draws are correlated with each other, the asymptotic variance formula is the sum of the variance of the moment condition, and the simulation variance.

Remark 2 (Aside). The key new regularity condition (relative to the usual regularity conditions for GMM with smooth moment conditions) for Proposition 1 is a stochastic equicontinuity condition allowing us to use the differentiation under the integral sign, that if $\delta_n \to 0$, then

$$\sup_{\|\theta_n - \theta_0\| \le \delta_n} \frac{\sqrt{n}\|\tilde{g}_n(\theta_n) - \tilde{g}_n(\theta_0) - E[g(Z_i, \theta_n)]\|}{1 + \sqrt{n}\|\theta_n - \theta_0\|} \xrightarrow{p} 0, \tag{4}$$

where $\tilde{g}_n(\theta_0) = n^{-1}\sum_{i=1}^n \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)$. Stochastic equicontinuity is a topic for 519 (check Ch 7 in Newey and McFadden 1994, if you're interested). The intuition is that we already know $\tilde{g}_n(\theta) \xrightarrow{p} E[g(Z_i, \theta)]$ by the law of large numbers (LLN), so that eq. (4) holds already pointwise (i.e. without the sup) for any $\theta_n \neq \theta_0$. The condition strengthens the pointwise convergence to hold uniformly.

Remark 3 (Estimating the asymptotic variance). Ω in the asymptotic variance formula is easy to estimate. On the other hand, since the moment condition $g(Z_i, \theta_0)$ is hard to evaluate, estimating G can be challenging. The problem is that in discrete choice models, $\tilde{g}_n(\theta) = n^{-1}\sum_{i=1}^n \tilde{g}(Z_i, \epsilon_i, \theta)$ is not typically differentiable in θ. One option is to use importance sampling to make it smooth (see Section 2.2 below). Another option is to take a numerical derivative, with a large enough step size s_n, as discussed in Newey and McFadden (1994, Chapter 7.3). In particular, the estimator

$$\hat{G}_j = \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta})}{s_n} \quad \text{or} \quad \hat{G}_j = \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta} - s_n e_j)}{2s_n}$$

of Ge_j (and similarly for other columns of G) will satisfy

$$\hat{G}_j - Ge_j \xrightarrow{p} 0, \tag{5}$$

if the step size s_n satisfies $s_n \to 0$ and $\sqrt{n}s_n \to \infty$. The idea is similar to kernel smoothing. To pick s_n in practice, one possibility is to plot $\hat{G}_j$ as a function of s_n, and then choose

s_n, small, but not in a region where the function is very choppy.

In some cases, one may obtain consistency under even weaker conditions on s_n. See Hong, Mahajan, and Nekipelov (2015) for a thorough treatment of numerical estimation of derivatives.

Proof of eq. (5). By triangle inequality, letting $g(\theta) = E[g(Z_i, \theta)]$ and $\tilde{g}_n(\theta) = n^{-1}\sum_i \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)$,

$$\left\| \frac{\tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\hat{\theta})}{s_n} - G e_j \right\| \\ \leq \frac{\left\| \tilde{g}_n(\hat{\theta} + s_n e_j) - \tilde{g}_n(\theta_0) - g(\hat{\theta} + s_n e_j) \right\|}{s_n} + \left\| \frac{g(\hat{\theta} + s_n e_j)}{s_n} - G e_j \right\| + \frac{\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0)\|}{s_n}$$

Now, by eq. (4) and the fact that $\hat{\theta} - \theta = O_p(n^{-1/2})$, the first term is bounded by

$$o_p(1)\frac{n^{-1/2} + \|\hat{\theta} + s_n e_j - \theta\|}{s_n} \leq o_p(1/s_n\sqrt{n}) + o_p(1) + o_p(1)\|\hat{\theta} - \theta\|/s_n = o_p(1/\epsilon\sqrt{n}).$$

By Taylor's theorem and the fact that g is differentiable at θ_0, $g(\hat{\theta} + s_n e_j) = g(\theta_0) + G(\hat{\theta} + s_n e_j - \theta_0) + o(\|\hat{\theta} + s_n e_j - \theta_0\|)$, the second term is bounded by

$$\left\| \frac{g(\hat{\theta} + s_n e_j)}{s_n} - G e_j \right\| \leq \|G(\hat{\theta} - \theta_0)/s_n\| + o(\|\hat{\theta} + s_n e_j - \theta_0\|/s_n) \\ \leq (\|G\| + o(1))\|\hat{\theta} - \theta_0\|/s_n + o(1) \leq O_p(1/\epsilon\sqrt{n}).$$

Finally, again using the triangle inequality and eq. (4),

$$\frac{\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0)\|}{s_n} \leq s_n^{-1}\|\tilde{g}_n(\hat{\theta}) - \tilde{g}_n(\theta_0) - g(\hat{\theta})\| + s_n^{-1}\|g(\hat{\theta})\| = o_p(1/s_n\sqrt{n}) + O_p(1/s_n\sqrt{n}). \quad \square$$

Example 4. First we for simplicity consider a binomial choice example. We have $Y_i = \mathbb{1}\{X_i'\theta_0 + \epsilon_i \geq 0\}$, with distribution of $\epsilon_i \sim F$ known and independent of X_i. This model delivers the moment (by setting $s(Z_i) = Y_i X_i$)

$$E[(Y_i - P_{\theta_0}(Y_i = 1 \mid X_i))X_i] = 0.$$

Since $P_\theta(Y_i = 1 \mid X_i) = F(X_i'\theta)$, we can estimate θ by GMM, using the moment $g(Z_i, \theta) = (Y_i - F(X_i'\theta))X_i$. Since we're exactly identified, the weight matrix doesn't matter, and the elements of the asymptotic variance would be given by

$$G = E[f(X_i'\theta_0)X_i X_i'],$$
$$\Omega = E[\mathrm{var}(Y_i \mid X_i)X_i X_i'] = E[F(X_i'\theta_0)(1 - F(X_i'\theta_0))X_i X_i'],$$

with $f = \frac{\mathrm{d}}{\mathrm{d}u}F(u)$ denoting the density. Suppose we don't know how to calculate the CDF $F(u)$, but we do know how to draw from the distribution. We could then use a

SMM estimator, with the moment (3) now given by

$$0 = E[\tilde{g}(Z_i, \tilde{\epsilon}_i, \theta)] \qquad \tilde{g}(Z_i, \tilde{\epsilon}_i, \theta) = \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{X_i'\theta + \tilde{\epsilon}_i^m \geq 0\} \right) X_i,$$

replacing the unknown CCP with a sample average based on the M simulated samples. Note that $\tilde{g}$ is a step function as a function of θ. ⊠

Example 1 (continued). In the multinomial model, we could try to match the CCPs,

$$g_j(Z_i, \theta) = (\mathbb{1}\{Y_{ij} = 1\} - P_\theta(Y_{ij} = 1 \mid X_i))X_i \tag{6}$$

(so that $s(Z_i) = X_i \otimes Y_i$). Since the event $\{Y_{ij} = 1\}$ is equivalent to $\bigcap_{\ell=1}^{J} \{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\epsilon_i\}$, we can estimate the CCP by

$$\frac{1}{M} \sum_{m=1}^{M} \prod_{\ell=0}^{J} \mathbb{1}\{(X_{ij} - X_{i\ell})'\beta \geq (a_\ell - a_j)'\tilde{\epsilon}_i^m\},$$

where $\tilde{\epsilon}_i^m$ is drawn from F. In other words, replace $P_\theta(Y_{ij} = 1 \mid X_i)$ with sample frequency with which j is the choice with the highest utility in the simulated data. ⊠

Example 2 (continued). The CCP in eq. (2) can be estimated as

$$\hat{P}_\theta(Y_{ik} = 1 \mid X_i) = \frac{1}{M} \sum_{m=1}^{M} \frac{e^{X_{ik}'\mu + X_{ik}'\Lambda\tilde{\epsilon}_i^m}}{\sum_{j=0}^{J} e^{X_{ij}'\mu + X_{ij}'\Lambda\tilde{\epsilon}_i^m}},$$

which is smooth in the parameters. See McFadden and Train (2000) for a discussion of how to pick the moments. ⊠

Remark 4. Because the moment condition (3) is mean zero even with a single simulation draw per observation ($M = 1$), using the number of simulation draws only affects the asymptotic variance of the estimator. The estimator's consistency or asymptotic normality is not affected by the simulation. Furthermore, the standard errors are only inflated by a factor of $\sqrt{(1 + 1/M)}$, which means that for practical purposes, $M = 10$ (say) is sufficient (leading to 5% larger standard errors).

1.2. *Simulated maximum likelihood*

This idea goes back to Lerman and Manski (1981): if the likelihood contains an integral that's hard to evaluate, replace it with a sample average based on simulated samples.

Example 4 (continued). We can also estimate θ by MLE,

$$\hat{\theta}_{\text{MLE}} = \operatorname*{argmin}_\theta \frac{1}{n} \sum_i q(Y_i, F(X_i, \theta)) \quad q(Y_i, F) = -\mathbb{1}\{Y_i = 1\} \log F - \mathbb{1}\{Y_i = 0\} \log(1 - F).$$

If F is hard to compute, we replace it with a simulated version,

$$\hat{\theta}_{\text{SMLE}} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_i q(Y_i, \hat{F}(X_i, \theta)), \qquad \hat{F}(X_i, \theta) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{X_i'\theta + \tilde{\epsilon}_i^m \geq 0\}. \qquad \boxtimes$$

More generally, for whatever discrete choice model, we can approximate the log likelihood as

$$\sum_{i=1}^{n} \sum_{j=0}^{J} \mathbb{1}\{Y_{ij} = 1\} \log \hat{P}_\theta(Y_{ij} = 1 \mid X_i).$$

This is particularly popular for estimating the random coefficient logit, see, for example Huber and Train (2001).

The main problem with this method is that because the log-likelihood is non-linear, even if $\hat{F}$ is an unbiased estimate of F, (i.e. $E[\hat{F}(X_i, \theta) \mid X_i] = F(X_i'\theta)$) it's still the case that $q(Y_i, E[\hat{F}(X_i, \theta) \mid X_i]) \neq E[q(Y_i, \hat{F}_i(\theta)) \mid X_i]$, so that the limiting objective function is not correct (and the SMLE is inconsistent) unless $M \to \infty$ as $n \to \infty$. For SMLE to be asymptotically equivalent to MLE, we need $M^2/n \to \infty$. For these reasons, it's a good idea to use the sandwich variance estimator when estimating the standard errors, as discussed in McFadden and Train (2000).

The methods of simulated moments and that of simulated scores, discussed next, were initially motivated by the desire for a simulation-based estimator that is consistent even for a fixed number of draws.

1.3. *Method of simulated scores*

This method is due to Hajivassiliou and McFadden (1998). One issue with SMM relative to MLE is that it is less efficient, even if we didn't incur a simulation error (because we have a parametric model, unless there is endogeneity in X_i). Another is that it is unclear how to pick the moments. If, as the moments, we used the score, we would solve both issues.

For concreteness, let us consider a discrete choice model, with the CCP given by $P(Y_{ij} = 1 \mid X_i) = p_j(\theta, X_i)$, and $\sum_{j=0}^{J} p_j(\theta, X_i) = 1$. Then the likelihood is given by $\sum_{i,j} Y_{ij} \log p_j(\theta, X_i)$, with the score given by

$$\sum_{i,j} Y_{ij} \frac{1}{p_j(\theta, X_i)} \frac{\partial p_j(\theta, X_i)}{\partial \theta}.$$

Since the expected value of the score is zero, this gives a moment condition. If the CCPs are hard to evaluate, we could instead try to get independent unbiased estimates of $\partial p_j(\theta, X_i)/\partial\theta$ and of $1/p_j(\theta, X_i)$. The first one is easy, since we can get unbiased estimates of $p_j(\theta, X_i)$ and the derivative is a linear operator. The trouble comes in getting reliable unbiased estimates of $1/p_j(\theta, X_i)$. Since $1/p$ is the mean of a geometric distribution (number of Bernoulli trials needed to get $Y_{ij} = 1$), one estimator is the number

of trials needed to generate $Y_{ij} = 1$. The disadvantage of this approach is that it can be slow, especially if some CCPs are small.

1.4. *Indirect inference*

This method is due to Smith (1993) (part of his 1990 PhD thesis), further developed in Gourieroux, Monfort, and Renault (1993). It was originally developed in a time-series context; we focus here on the cross-section and panel case.

The intuition for the SMM method is that the sample moments $\frac{1}{n}\sum_{i=1}^{n} s(Z_i)$ from the real data should match sample moments from the simulated data, if the simulated data is generated using the true θ_0. The insight of indirect inference is that we don't need to restrict ourselves to moments: if we have the right model generating the data, then *any feature* of the simulated data should match the real data. In particular, we can focus on matching features of the real data that we think are important.

To make this concrete, let $\hat{\pi}$ be some easily-computable reduced-form parameter $\pi_0 = h(\theta_0)$, which satisfies

$$\sqrt{n}(\hat{\pi} - \pi_0) \Rightarrow \mathcal{N}(0, \Omega). \tag{7}$$

For example, π could be a reduced-form regression, quantile regression parameters, reduced-form choice probabilities, or even maximum likelihood estimates based on a simpler (and therefore potentially misspecified) model (the case studied in Gallant and Tauchen 1996), or GMM estimates based on such model. The only thing to be careful about is that in deriving Ω, we don't want to assume that the model for π is correctly specified. So if π is, say, the MLE estimand, we want to compute Ω using the sandwich formula.

The model defining the reduced-form parameter is called an *auxiliary model*. To fix ideas, suppose $\hat{\pi}$ is an M-estimator

$$\hat{\pi} = \operatorname*{argmin}_{\pi} \frac{1}{n}\sum_{i=1}^{n} q(Z_i, \pi), \tag{8}$$

which defines $\pi = h(\theta) = \operatorname{argmin}_{\varpi} \int q(z, \varpi)\, \mathrm{d}P_\theta(z)$. This map $h\colon \theta \to \pi$ is called the *binding function* by Gourieroux, Monfort, and Renault (1993). If it were a known function, then we could use minimum distance to get an estimate of θ.

Typically, however, it is hard to figure out the form of h, for the same reason that it is hard to evaluate the likelihood in the first place. On the other hand, it's easy to simulate the data: generate draws $\{\tilde{\epsilon}_i^m\}_{m=1}^{M}$ from the (known) distribution of ϵ_i, and set

$$\tilde{Y}_i^m(\theta) = m(X_i, \tilde{\epsilon}_i^m, \theta)$$

and $\tilde{X}_i^m = X_i$.[6] This gives us M simulated samples $\{\tilde{Z}_i^m(\theta) = (\tilde{Y}_i^m(\theta), X_i)\}_{i=1}^n$. On each of these simulated samples, we can estimate π using the same estimator that we used to get $\hat{\pi}$ on the real data: $\tilde{\pi}^m(\theta) = \operatorname{argmin}_\pi n^{-1} \sum_{i=1}^n q(\tilde{Z}_i^m(\theta), \pi)$. We then estimate θ by minimizing the distance between the reduced-form estimate from the data and those from the simulated samples,

$$\hat{\theta}_{\text{II}} = \underset{\theta}{\operatorname{argmin}} \left(\hat{\pi} - \tilde{\pi}(\theta)\right)' \hat{W} \left(\hat{\pi} - \tilde{\pi}(\theta)\right), \qquad \tilde{\pi}(\theta) = \frac{1}{M} \sum_{m=1}^{M} \tilde{\pi}^m(\theta).$$

An alternative (and asymptotically equivalent) approach is to pool all the simulated data together and estimate $\tilde{\pi}$ based on one big sample of size Mn.

Proposition 5. Suppose that as $n \to \infty$,

1. $\pi(\theta) \neq \pi(\theta_0)$ *if* $\theta \neq \theta_0$
2. $\pi(\theta)$ *is continuous*
3. Θ *is compact*
4. $\sup_\theta |\tilde{\pi}(\theta) - \pi(\theta)| \xrightarrow{p} 0$
5. $\hat{\pi} \xrightarrow{p} h(\theta_0)$, *and* $\hat{W} \xrightarrow{p} W$ *positive definite.*

and that $M \geq 1$ is fixed. Then $\hat{\theta}_{II} \xrightarrow{p} \theta_0$.

Proof. This follows by verifying conditions UC and ID in the theorem for consistency of extremum estimators that is discussed in 519. □

Using empirical process methods (see Theorem 7.2 in Newey and McFadden 1994), and assuming (7), it follows that

$$\sqrt{n}(\tilde{\pi}^m(\theta) - \pi_0) \Rightarrow \mathcal{N}(0, \Omega),$$

with the limiting distribution

$$\sqrt{n}(\hat{\theta}_{\text{II}} - \theta_0) \Rightarrow \mathcal{N}\left(0, \left(1 + \frac{1}{M}\right)(G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}\right).$$

Similar to the SMM case, we would need $M \to \infty$ to be as efficient as a minimum distance estimator.

Remark 6 (Picking the reduced form). If the auxiliary model is the true model (that is, if $q(\cdot)$ in (8) is the true log-likelihood), with $\dim(\pi) = \dim(\theta)$, then solving $\hat{\theta} = h(\hat{\pi})$ gives the MLE of θ. This suggests that it should be a good idea to pick the auxiliary model to be close to the true model. However, so long as $h(\cdot)$ is invertible (which may be in practice hard to check!), the indirect inference estimator will be consistent—it doesn't

6. One could also make the simulated samples have sample size different from n, in which case X_i^m would be drawn from the empirical distribution of X

matter if the auxiliary model is misspecified. The choice of a particular auxiliary model only affects efficiency of the estimator. The choice is often driven by what aspect of the data you want to fit.

Remark 7. Indirect inference is similar to calibration in macro—except that we do get standard errors as well. See the working paper https://arxiv.org/abs/2109.08109 for how to get the standard errors in calibration exercises.

Remark 8. Notice that the efficient weight matrix is proportional to Ω^{-1}, which, since Ω is just the variance of the reduced-form estimator, can be estimated directly from the data—we do not need a two-step estimator, in which we first get a consistent estimate of θ, and then use the estimate to form an efficient weight matrix.

Often, people set $\hat{W} = \text{diag}(\hat{\Omega})^{-1}$. It may have something to do with misspecification of the model (1).

Research Question. How should one deal with such potential misspecification? ⊠

2. PRACTICAL ISSUES

2.1. *An important implementation note*

Remark 9. For the simulation methods to work, it is important to draw the sequence $\{\tilde{\epsilon}_i^1, \ldots, \tilde{\epsilon}_i^M\}_{i=1}^n$ once, and hold it fixed as we search over β. Otherwise, the objective function will jump wildly as we move over the Θ space, and consistency will fail (and, on a more basic level, the estimator is not well-defined). See Figure 1. In fact, in the derivation of the asymptotics, we have implicitly treated $(Y_i, X_i, \tilde{\epsilon}_i)$ as the data.

One way of ensuring the draws are fixed is to set the seed to a particular value. A better way is to draw the sequence of epsilons once at the beginning, store them, and then use the set each time you need to evaluate the objective function. In other words, treat the simulation draws as part of the data after you generate them. It's more transparent, marginally quicker (since you don't need to re-generate the draws), and it also makes it easy to share the draws along with the data with another person or another programming language if the whole estimation uses more than one language.

2.2. *Improving on frequency simulators*

In the description of the simulation methods above, we have used a simple frequency simulator to estimate the CCPs $p_j(\theta, X_i) := P_\theta(Y_{ij} = 1 \mid X_i)$,

$$\hat{p}_j(\theta, X_i) = \frac{1}{M} \sum_{m=1}^{M} \tilde{Y}_{ij}^m(\theta).$$

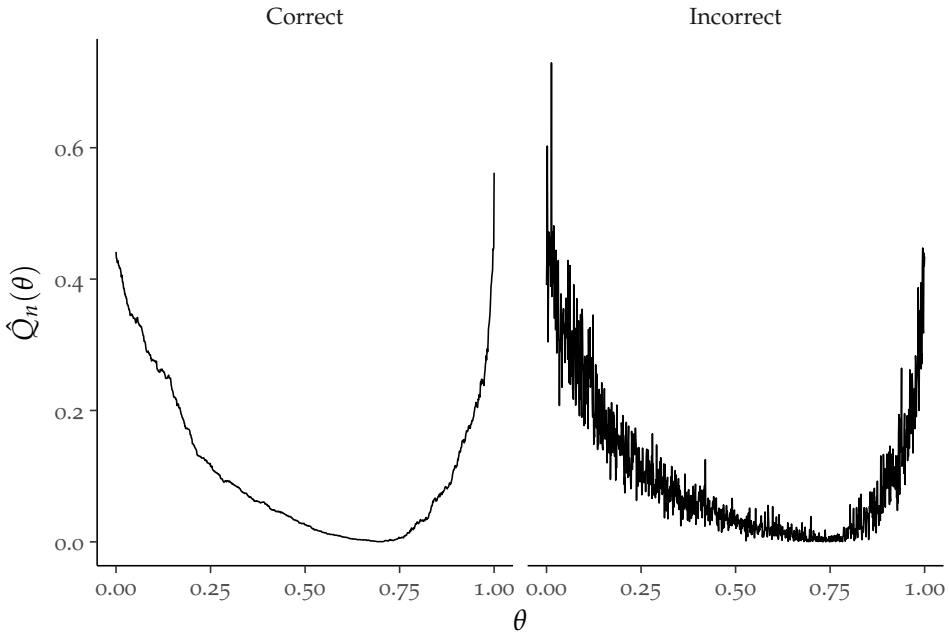

Figure 1: If $\tilde{\epsilon}$ is not held fixed, the objective function oscillates wildly ("Incorrect" panel). When $\tilde{\epsilon}$ is held fixed ("Correct" panel), the objective function is much better behaved, although it is still not differentiable.

This has multiple problems:

1. Slow if J is large
2. Sample objective function is discontinuous in θ, since $\tilde{Y}_{ij}^m(\theta)$ will be a step function. Apart from complications with asymptotic theory, this makes the sample objective function hard to optimize.
3. There is a positive probability that $\hat{p}_j(\theta, X_i) = 0$, which creates issues with some estimators (such as simulated maximum likelihood). A related problem is that the frequency sampler will be imprecise if the CCP is close to zero.

We'll briefly discuss two approaches to ameliorate this problem: kernel smoothing and importance sampling (or, more precisely, importance sampling coupled with a change of variables), both of which were suggested in McFadden (1989).

Remark 10. Besides these two methods, there have been other proposals for implementing simulation-based methods. For example, the frequency simulator uses $n \times M$ independent draws $\tilde{\epsilon}_i^m$. If n is very large, this may be computationally prohibitive. An alternative is to use the same set of M draws $\{\tilde{\epsilon}^m\}_{m=1}^M$ for each observation. See, for example, Hong, Li, and Li (2021) for an implementation of this idea in the context of the Berry, Levinsohn, and Pakes (1995) model, and Armstrong et al. (2017) for a general theory.

KERNEL SMOOTHING The idea of kernel smoothing is to replace the discrete choice indicators by a smooth function of the underlying continuous latent variables that determine the model's discrete outcomes. An obvious problem is how to choose the amount of smoothing, and that the smoothing induces finite-sample bias. See Bruins et al. (2018) for an application of this idea in the context of indirect inference.

IMPORTANCE SAMPLING This idea was imported to econometrics by Kloek and van Dijk (1978). To explain it, consider Example 4, and use a change of variables to rewrite the CCP as

$$P_\theta(Y_i = 1 \mid X_i = x) = E[\mathbb{1}\{X_i'\theta + \epsilon_i \geq 0\} \mid X_i = x] = \int \mathbb{1}\{u \geq 0\} f(u - x'\theta)\,\mathrm{d}u$$
$$= \int \mathbb{1}\{u \geq 0\} \frac{f(u - x'\theta)}{g(u \mid x)} g(u \mid x)\,\mathrm{d}u,$$

where f is the density of ϵ_i, and $g(u \mid x)$ is some other density. If we draw samples $\{\tilde{u}_i^m\}_{m=1}^M$ this density, we can estimate the CCP by

$$\frac{1}{M}\sum_m \mathbb{1}\{\tilde{u}_i^m \geq 0\} \frac{f(\tilde{u}_i^m - X_i'\theta)}{g(\tilde{u}_i^m \mid X_i)},$$

which is differentiable in θ so long as f is differentiable. Instead of holding the draws $\tilde{\epsilon}_i^m$ (and their implicit weights, $1/M$) constant as we change θ, we're holding the utility $\tilde{u}_i^m$ constant, but vary the importance weights $\frac{1}{M}\frac{f(\tilde{u}_i^m - X_i'\theta)}{g(\tilde{u}_i^m \mid X_i)}$ that we put on each simulated observation as we search over θ.

- Since the integrand is only non-zero if u is positive, the support of g only needs to be on $\mathbb{R}_+$: for instance, we can set it to a truncated normal, or an exponential distribution. Then we can get rid of the indicator in the preceding display.
- A possible choice for $g(u \mid x)$ is to set it to $f(u - x'\tilde{\theta})$, truncated to the positive part of the real line, where $\tilde{\theta}$ is an initial guess of θ.
- Since $\mathbb{1}\{\tilde{u}_i^m \geq 0\}/g(\tilde{u}_i^m \mid X_i)$ doesn't vary with θ, we can store it along with the simulation draws as we search over θ.

Example 1 (continued). Suppose $F = \Phi$, so we're in the multinomial probit case. Let $V_{ij} = (Y_{i0}^* - Y_{ij}^*, \ldots, Y_{i,j-1}^* - Y_{ij}^*, Y_{i,j+1}^* - Y_{ij}^*, \ldots, Y_{iJ}^* - Y_{ij}^*)$ denote the vector of utility differences relative to choice j. Then V_{ij} conditional on X_i is multivariate normal with mean and variance depending on $\theta = (\beta, \Sigma)$. Let $f(V_{ij} \mid X_i, \theta)$ denote its pdf. Then

$$P_\theta(Y_{ij} = 1 \mid X_i = x) = \int \prod_{\ell=1}^J \mathbb{1}\{v_\ell \leq 0\} f(v \mid x, \theta)\,\mathrm{d}v$$
$$= \int \prod_{\ell=1}^J \mathbb{1}\{v_\ell \leq 0\} \frac{f(v \mid x, \theta)}{g(v \mid x)} g(v \mid x)\,\mathrm{d}v,$$

where g is a density supported on $(\mathbb{R}_-)^J$, such as the product of truncated normal variables or a product of exponentials. If we can draw $\{\tilde{v}_i^m\}_{m=1}^M$ from it, then we can estimate the CCP by

$$\frac{1}{M}\sum_{m=1}^{M}\frac{f(\tilde{v}_i^m \mid x,\theta)}{g(\tilde{v}_i^m \mid x)}.$$

Again, this will yield a smooth moment condition. ⊠

Remark 11. In some cases, it may be hard to compute $m(X_i,\epsilon_i,\theta)$. This is the case when Y_i is the equilibrium of a game, or solution to a dynamic program. It would then be very useful if we only had to compute the equilibrium once, and not every time we change θ as we're optimizing. See Ackerberg (2009) for how one can use change of variables coupled with importance sampling to avoid having to recompute the equilibrium each time.

Remark 12. Importance sampling is also a variance reduction method. Consider the problem of calculating

$$p(\theta) = \int h(\epsilon,\theta)f(\epsilon)\,\mathrm{d}\epsilon,$$

where f is a density. In Example 4, with $p(\theta) = P_\theta(Y_i = 1 \mid X_i = x)$, we had $h(\epsilon,\theta) = \mathbb{1}\{x'\theta + \epsilon \geq 0\}$, and f is the density of ϵ. A frequency simulator estimates this quantity as

$$\hat{p} = \frac{1}{M}\sum_{m=1}^{M} h(\epsilon^m,\theta), \qquad \operatorname{var}(\hat{p}) = \frac{1}{M}\int (h(\epsilon,\theta) - p(\theta))^2 f(\epsilon)\,\mathrm{d}\epsilon.$$

Suppose that we use importance sampling, and we simulate u^m from density g, yielding the estimate

$$\hat{p} = \frac{1}{M}\sum_{m=1}^{M}\frac{h(u^m,\theta)f(u^m)}{g(u^m)}, \qquad \operatorname{var}(\hat{p}) = \frac{1}{M}\int \left(\frac{h(u,\theta)f(u)}{g(u)} - p(\theta)\right)^2 g(u)\,\mathrm{d}u.$$

Notice that we can make the variance zero if we set $g(u) = g(u,\theta) = h(u,\theta)f(u)/p(\theta)$ (this integrates to one, so it's a density), which, unfortunately, requires knowledge of p. However, the variance expression suggests that we want to make $g(u)$ large if $f(u)h(u,\theta)$ is large, that is, make g large for those u's that contribute a lot to the integral: hence the name "importance sampling". Intuitively, we want to make the integrand $h(u,\theta)f(u)/g(u)$ to be as close to constant as possible. Other observations:

- The importance distribution does not have to be positive everywhere. It is enough to have $g(u) > 0$ whenever $f(u)h(u,\theta) \neq 0$ (we used this previously)
- We do need to prevent $h(u,\theta)f(u)/g(u)$ from getting very large (to keep the variance from exploding), so that the tails of $g(u)$ should be at least as thick as those of $h(u,\theta)f(u)$.

See Sauer and Taber (2021) for an application of a version of this idea to indirect inference.

This is a specific method evaluating the probability that a correlated multivariate normal random variable falls into a rectangular region. Since the choice probabilities in discrete choice models with normal errors (the panel probit model, or the multinomial probit model) take this form, it is commonly used in such models. The Geweke-Hajivassiliou-Keane (GHK) simulator is named after Geweke (1989) and Hajivassiliou and McFadden (1998) and Keane (1994). It reduces the problem of calculating a J-dimensional integral to a sequence of univariate integrals via a conditioning argument. It is a special version of importance sampling, and as such it has the advantage of producing simulated probabilities that are smooth functions of the model parameters.

GENERAL ALGORITHM We first describe the general algorithm. Suppose we want to evaluate the probability that $Z \sim \mathcal{N}_J(0, LL')$, falls into a rectangular region: $A = \{a_j \leq Z_j \leq b_j\}$. Here L is lower-triangular, corresponding to the Cholesky decomposition of the covariance matrix. Write $Z = L\epsilon$, so that $\{Z \in A\}$ is equivalent $\{\epsilon \in B\}$, with $B = \{L^{-1}a \leq \epsilon \leq L^{-1}b\}$.

Now, $\{a_j \leq Z_j \leq b_j\} = \{a_j \leq \sum_{k=1}^{j} L_{jk}\epsilon_j \leq b_j\}$. Therefore, $\epsilon_j \mid \epsilon_1, \ldots, \epsilon_{j-1}, B$, is a univariate standard normal, truncated to

$$\frac{a_j - \mu_j}{L_{jj}} \leq \epsilon_j \leq \frac{b_j - \mu_j}{L_{jj}}, \quad \mu_j = \begin{cases} \sum_{k=1}^{j-1} L_{jk}\epsilon_j & \text{if } j > 1, \\ 0 & \text{if } j = 1. \end{cases}$$

Hence, by the law of total probability, the density of ϵ given B is simply

$$\begin{aligned} g(\epsilon \mid B) &= g(\epsilon_1 \mid B)g(\epsilon_2 \mid \epsilon_1, B) \cdots g(\epsilon_J \mid \epsilon_1, \ldots, \epsilon_{j-1}, B) \\ &= \prod_{j=1}^{J} \frac{\phi(\epsilon_j)}{\Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})} \end{aligned}$$

Since $g(\epsilon \mid B)$ only has support over B, so that $g(\epsilon \mid B) = \mathbb{1}\{\epsilon \in B\}g(\epsilon \mid B)$, it follows that

$$\begin{aligned} P(Z \in A) &= \int \mathbb{1}\{\epsilon \in B\} \prod_j \phi(\epsilon_j) \mathrm{d}\epsilon \\ &= \int \left[\prod_{j=1}^{J} (\Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})) \right] g(\epsilon \mid B) \mathrm{d}\epsilon, \end{aligned}$$

where the second line follows by multiplying and dividing by $\prod_{j=1}^{J} \Phi((b_j - \mu_j)/L_{jj}) - \Phi((a_j - \mu_j)/L_{jj})$.

Since sampling from $g(\epsilon \mid B)$ is simple, we can evaluate $P(Z \in A)$ by sampling from this density, and using importance weights given in the square parentheses in the above expression. This is the GHK algorithm:

1. Draw $\tilde{\epsilon}_1^m, \ldots, \tilde{\epsilon}_{J-1}^m$ from the density $g(\epsilon \mid B)$. In particular, for $j = 1, \ldots, J-1$, draw $\tilde{\epsilon}_j^m$ from standard normal, truncated above at $(b_j - \tilde{\mu}_j^m)/L_{jj}$, and below at $(a_j - \tilde{\mu}_j^m)/L_{jj}$, where $\tilde{\mu}_j^m = \sum_{k=1}^{j-1} L_{jk}\tilde{\epsilon}_j^m$. This is done by drawing a $(J-1)$-vector $\tilde{U}^m$ of independent uniforms, and setting

$$\tilde{\epsilon}_j^m = \Phi^{-1}(\tilde{U}_j^m(\Phi((b_j - \tilde{\mu}_j^m)/L_{jj}) - \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})) + \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})).$$

2. Repeat the previous step M times, and form the estimate

$$\hat{P}(Z \in A) = \frac{1}{M}\sum_{m=1}^{M}\prod_{j=1}^{J}(\Phi((b_j - \tilde{\mu}_j^m)/L_{jj}) - \Phi((a_j - \tilde{\mu}_j^m)/L_{jj})).$$

A detailed description, tailored to the multinomial probit model, can be found in Train (2009, Chapter 4).

MULTINOMIAL PROBIT Let us now show how this simulator can be applied to the multinomial probit model, as in Example 1. Let $Y_{ij}^* = V_{ij} + U_{ij}$ denote the utility from choice $j = 0, \ldots, J$, with $U_i = A\epsilon_i$, and $\epsilon_i \sim \mathcal{N}_{J+1}(0, I)$. Here $V_{ij} = X_{ij}'\beta$ denotes the non-stochastic part of the utility.

- The attractive feature of this model is that it doesn't restrict which choices are close: this allows for rich substitution patterns. In a random effects approach, for example, only choices that are close in terms of observables can be close.
- The problem is that if J is large, there are too many parameters in the covariance matrix to estimate.

VARIANCE NORMALIZATION First, we'll need some normalization on the covariance matrix $\Sigma = AA'$. Let A denote the Cholesky decomposition of Σ, so that A is lower triangular. Let $\tilde{U}_{ik} = U_{i,-k} - U_{ik}$ denote the differences relative to alternative k. Then $\tilde{U}_{ik} \sim \mathcal{N}_J(0, \tilde{\Sigma}_k)$, where $\tilde{\Sigma}_k = M_k\Sigma M_k'$, where M_k is the matrix $I_{J+1} - \iota_{J+1}e_k'$ with the kth row (which is all zeros), removed. Here ι is a $(J+1)$-vector of ones, and e_k is a vector of zeros with 1 in the kth position. For instance, if $J = 2$, then

$$M_0 = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}.$$

It is clear that since only utility differences matter, we could just parametrize the model in terms of $\tilde{\Sigma}_0$, say the utility differences with respect to the outside option. This is still not enough, since we need to normalize the scale of the utility. Typically, this is done by setting $\tilde{\Sigma}_{0,11} = 1$. To ensure $\tilde{\Sigma}_0$ is positive definite, it is convenient to parametrize the model terms of its Cholesky factor, the lower-triangular matrix A_0 with $A_{0,11} = 1$ so that $\tilde{\Sigma}_{0,11} = 1$. The remaining elements of A_0 can vary freely. So while Σ has $(J+1)J/2$ parameters, after normalization, we are left with $J(J-1)/2 - 1$ free parameters. This

means we can normalize $J+1$ parameters in the original matrix. A convenient choice is to set the first row and column to zero, in which case

$$\Sigma(A_0) = \begin{pmatrix} 0 & 0 \\ 0 & A_0 A_0' \end{pmatrix}', \qquad A_0 = \begin{pmatrix} 1 & 0 & \cdots & \\ a_{21} & a_{22} & 0 & \cdots \\ \vdots & & \ddots & \end{pmatrix}.$$

With this normalization, we can derive the matrix $\tilde{\Sigma}_j$ of covariances when we take utility differences with respect to any alternative j.

Example 5. For instance, with $J = 2$, we have 2 free parameters,

$$\Sigma = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & a_{12} \\ 0 & a_{12} & a_{12}^2 + a_{22}^2 \end{pmatrix}, \quad \tilde{\Sigma}_0 = A_0 A_0' = \begin{pmatrix} 1 & a_{12} \\ a_{12} & a_{12}^2 + a_{22}^2. \end{pmatrix}, \quad A_0 = \begin{pmatrix} 1 & 0 \\ a_{12} & a_{22} \end{pmatrix}.$$

and

$$\tilde{\Sigma}_1 = \begin{pmatrix} 1 & 1 - a_{12} \\ 1 - a_{12} & (1 - a_{12})^2 + a_{22}^2 \end{pmatrix}, \quad \tilde{\Sigma}_2 = \begin{pmatrix} a_{12}^2 + a_{22}^2 & -a_{12} + a_{12}^2 + a_{22}^2 \\ -a_{12} + a_{12}^2 + a_{22}^2 & (1 - a_{12})^2 + a_{22}^2 \end{pmatrix}. \quad \boxtimes$$

APPLYING THE GHK SIMULATOR The GHK simulator works with utility differences with respect to the choice the probability of which we are simulating. To simulate $p_j(\theta, X_i) := P_\theta(Y_{ij} = 1 \mid X_i)$, let $\tilde{\Sigma}_j = LL'$ denote the Cholesky decomposition, so that $\tilde{U}_{ij} = L\epsilon_{ij}$, where $\epsilon_{ij} \sim \mathcal{N}_J(0, I)$, and $\theta = (\beta, A_0)$, and let $\tilde{V}_{ik} = V_{ik} - V_{ij}$. Given the normalization above, for a given matrix A_0, we can calculate L easily as the Cholesky decomposition of $M_j \Sigma(A_0) M_j'$.

$$p_j(\theta, X_i) = P(\tilde{V}_i + L\epsilon_i \leq 0) = P(L\epsilon_i \leq -\tilde{V}_i).$$

Thus, we apply the GHK simulator with $a = -\infty$, $b = -\tilde{V}_i$, L given by the Cholesky decomposition of $\tilde{\Sigma}_j$, and the uniform draws $\tilde{U}^m$ given by a $(J-1)$ vector of uniforms $\tilde{U}_{ij}^m$, which we keep fixed if we optimize over θ.

REFERENCES

Ackerberg, Daniel A. 2009. "A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation." *Quantitative Marketing and Economics* 7, no. 4 (December): 343–376. https://doi.org/10.1007/s11129-009-9074-z.

Armstrong, Tim, A Ronald Gallant, Han Hong, and Huiyu Li. 2017. "The Asymptotic Distribution of Estimators with Overlapping Simulation Draws." December. https:

//cpb-us-w2.wpmucdn.com/campuspress.yale.edu/dist/2/277/files/2017/12/olsim_dec2017-1fvorn4.pdf.

Berry, Steven T., James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63, no. 4 (July): 841–890. https://doi.org/10.2307/2171802.

Bruins, Marianne, James A. Duffy, Michael P. Keane, and Anthony A. Smith Jr. 2018. "Generalized Indirect Inference for Discrete Choice Models." *Journal of Econometrics* 205, no. 1 (July): 177–203. https://doi.org/10.1016/j.jeconom.2018.03.010.

Cox, David Roxbee. 1958. "Some Problems Connected with Statistical Inference." *The Annals of Mathematical Statistics* 29, no. 2 (June): 357–372. https://doi.org/10.1214/aoms/1177706618.

Gallant, A Ronald, and George Tauchen. 1996. "Which Moments to Match?" *Econometric Theory* 12, no. 4 (October): 657–681. https://doi.org/10.1017/S0266466600006976.

Geweke, John. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57, no. 6 (November): 1317–1339. https://doi.org/10.2307/1913710.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. "Indirect Inference." *Journal of Applied Econometrics* 8, no. S1 (December): S85–S118. https://doi.org/10.1002/jae.3950080507.

Hajivassiliou, Vassilis A., and Daniel L. McFadden. 1998. "The Method of Simulated Scores for the Estimation of LDV Models." *Econometrica* 66, no. 4 (July): 863–896. https://doi.org/10.2307/2999576.

Hong, Han, Huiyu Li, and Jessie Li. 2021. "BLP Estimation Using Laplace Transformation and Overlapping Simulation Draws." *Journal of Econometrics* 222, no. 1A (May): 56–72. https://doi.org/10.1016/j.jeconom.2020.07.026.

Hong, Han, Aprajit Mahajan, and Denis Nekipelov. 2015. "Extremum Estimation and Numerical Derivatives." *Journal of Econometrics* 188, no. 1 (September): 250–263. https://doi.org/10.1016/j.jeconom.2014.05.019.

Huber, Joel, and Kenneth Train. 2001. "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths." *Marketing Letters* 12 (3): 259–269. https://doi.org/10.1023/A:1011120928698.

Keane, Michael P. 1994. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica* 62, no. 1 (January): 95–116. https://doi.org/10.2307/2951477.

Kloek, T., and Herman K. van Dijk. 1978. "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo." *Econometrica* 46, no. 1 (January): 1–19. https://doi.org/10.2307/1913641.

Lerman, Stveven R., and Charles F. Manski. 1981. "On the Use of Simulated Frequencies to Approximate Choice Probabilities." In *Structural Analysis of Discrete Data with Econometric Applications*, edited by Charles F. Manski and Daniel L. McFadden, 305–319. Cambridge: MIT Press. https://eml.berkeley.edu/~mcfadden/discrete.html.

McFadden, Daniel L. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration." *Econometrica* 57, no. 5 (September): 995–1026. https://doi.org/10.2307/1913621.

McFadden, Daniel L., and Kenneth E. Train. 2000. "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics* 15, no. 5 (September): 447–470. https://doi.org/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Chap. 36 in *Handbook of Econometrics*, edited by Robert F. Engle and Daniel L. McFadden, 4:2111–2245. New York, NY: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.

Pakes, Ariel, and David Pollard. 1989. "Simulation and the Asymptotics of Optimization Estimators." *Econometrica* 57, no. 5 (September): 1027–1057. https://doi.org/10.2307/1913622.

Sauer, Robert M., and Christopher Taber. 2021. "Understanding Women's Wage Growth Using Indirect Inference with Importance Sampling." *Journal of Applied Econometrics* 36, no. 4 (June): 453–473. https://doi.org/10.1002/jae.2818.

Smith, Anthony A. 1993. "Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions." *Journal of Applied Econometrics* 8, no. S1 (December): S63–S84. https://doi.org/10.1002/jae.3950080506.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge, UK: Cambridge University Press. https://eml.berkeley.edu/books/choice2.html.