Available online at www.sciencedirect.com

ScienceDirect

Journal of Econometrics 141 (2007) 683–703

JOURNAL OF Econometrics

www.elsevier.com/locate/jeconom

Masking identification of discrete choice models under simulation methods

Lesley Chiou[a,*], Joan L. Walker[b]

[a]*Department of Economics, Occidental College, 1600 Campus Road, Los Angeles, CA 90041, USA*
[b]*Center for Transportation Studies, Boston University, 675 Commonwealth Avenue, Boston, MA 02215, USA*

Available online 8 December 2006

Abstract

We present examples based on actual and synthetic datasets to illustrate how simulation methods can mask identification problems in the estimation of discrete choice models such as mixed logit. Simulation methods approximate an integral (without a closed form) by taking draws from the underlying distribution of the random variable of integration. Our examples reveal how a low number of draws can generate estimates that appear identified, but in fact, are either not theoretically identified by the model or not empirically identified by the data. For the particular case of maximum simulated likelihood estimation, we investigate the underlying source of the problem by focusing on the shape of the simulated log-likelihood function under different conditions.

JEL classification: C15; C25

Keywords: Simulation methods; Discrete choice; Identification

1. Introduction

Over the past decade, simulation methods have grown in popularity as advancements in computational speed have allowed researchers to estimate increasingly richer models of consumer and firm behavior. In particular, the literature on consumer choice theory has spread rapidly with the development of numerical techniques such as maximum simulated likelihood and the method of simulated moments (for example, Berry et al., 1995;

[*]Corresponding author. Tel.: +1 323 259 2875; fax: +1 323 259 2704.
E-mail addresses: lchiou@oxy.edu (L. Chiou), joanw@bu.edu (J.L. Walker).

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.10.012

Brownstone and Train, 1999; Goolsbee and Petrin, 2003). While simulation allows estimation of more flexible models, we examine how simulation noise can mask identification problems inherent in the model or data. We explore the issue within the context of mixed logit and maximum simulated likelihood estimation through the use of both synthetic and real data. Furthermore, we investigate the underlying source of the problem by examining the shape of the simulated log-likelihood function under different conditions. While we examine the particular case of maximum simulated likelihood estimation, the caveats presented in this paper could apply to other estimation strategies that employ simulation methods.

Simulation methods rely on approximating an integral (that does not have a closed form) through Monte Carlo integration. Draws are taken from the underlying distribution of the random variable of integration and used to calculate the numeric integral. Poor approximations of numerical techniques have important implications for the interpretation of the estimates and any resulting conclusions or welfare calculations. We show how the practice of applying maximum simulated likelihood estimation can generate misleading results even under 1000 pseudo-random, Halton, or shuffled Halton draws. Most empirical work on mixed logit models typically apply 200–300 draws.

We present examples of mixed logit models where employing a "low" number of draws to construct the simulated integral can generate estimates that are not identified by the model or the data. In each of the examples, we estimate the model under different types of simulation draws: random, Halton, and shuffled Halton. We consider two types of identification problems: theoretical and empirical unidentification. Theoretical unidentification occurs when the model cannot be estimated in principle (regardless of the data at hand). Empirical unidentification occurs when the data cannot support the model even though the model may be estimable in principle.

The first example utilizes a dataset on consumers' choices of telephone plans. It demonstrates how a model that is not theoretically identified can appear to result in identified estimates at a low number of draws. A classic symptom of unidentification, a singular Hessian, does not emerge until a much higher number of draws is employed. In the second set of examples, we generate synthetic datasets to investigate the source of empirical unidentification by examining the shape of the simulated log-likelihood functions under varying numbers of draws and identification conditions. The last two examples use an actual dataset on consumer choices across retail stores and a synthetic dataset to consider empirical unidentification under more complex specifications.

Limited research exists on the empirical identification of discrete choice models under simulation methods. Ben-Akiva and Bolduc (1996) and Walker (2001) note that an identification problem can arise when a low number of draws are used, and they and others (e.g., Hensher and Greene, 2003) emphasize the necessity of verifying the stability of parameter estimates as the number of draws are increased. Walker (2001) and Walker et al. (2006) examine the theoretical framework for the identification of mixed logit models and support their findings with empirical examples. In contrast, this paper focuses primarily on the issue of empirical unidentification by explicitly examining the properties and transformation of the log-likelihood function over a range of number of draws.

The next section presents the theoretical framework of the mixed logit model and describes the estimation procedure. The remaining sections discuss the empirical examples and illustrate the sensitivity of the estimates with respect to simulation.

2. The mixed logit model

2.1. Consumer preferences

In this discrete choice model, the utility that consumer n, $n = 1, \ldots, N$ where N is the sample size, receives from choosing alternative i, $i = 1, \ldots, J_n$ is given by

$$U_{ni} = X_{ni}\beta_n + \varepsilon_{ni}, \tag{1}$$

where X_{ni} is a $(1 \times K)$ vector of observable characteristics for alternative i and consumer n and β_n is a $(K \times 1)$ vector of consumer n's tastes over the attributes of alternative i. The random coefficient β_n contains a subscript n to indicate that preferences over the characteristics of an alternative may vary among individuals in the population. The term ε_{ni} captures consumer n's idiosyncratic and unobservable taste for alternative i. Under a logit model, ε_n (consisting of ε_{ni}, $i = 1, \ldots, J_n$) is an i.i.d. extreme value random vector. The random coefficient β_n can assume any distributional form (see Train, 2003, for further information). McFadden and Train (2000) demonstrate that any random utility model can be "approximated to any degree of accuracy by a mixed logit model with the appropriate choice of variables" and distribution of the random coefficient.

A consumer chooses the alternative that gives her the highest utility. More specifically, the set of values A_{ni} of the idiosyncratic error ε_{ni} that induces consumer n to choose alternative i is given by

$$A_{ni} = \left\{ \varepsilon_{ni} : U_{ni}(X_{ni}, \beta_n, \varepsilon_{ni}) \geqslant \max_{j=1,\ldots,J_n} U_{nj}(X_{nj}, \beta_n, \varepsilon_{nj}) \right\}, \tag{2}$$

where j indexes all possible alternatives in consumer n's choice set. Conditional on the utility parameters β_n, the probability that consumer n chooses alternative i is given by the standard logit formula:

$$L_{ni}(\beta_n) = \int_{A_{ni}} f(\varepsilon)\,\mathrm{d}\varepsilon, \tag{3}$$

$$L_{ni}(\beta_n) = \frac{\exp(X_{ni}\beta_n)}{\sum_{j=1}^{J_n} \exp(X_{nj}\beta_n)}, \tag{4}$$

where $f(\cdot)$ is the density of the extreme value distribution. By convention, the parameters of the distribution of ε_{ni} are normalized to set the level and scale of the utility function (Train, 2003). The location parameter of each extreme value error ε_{ni} is zero, and the scale is set to 1 so that $\mathrm{Var}(\varepsilon_{ni}) = \pi^2/6$.

Since β_n is not observed, the unconditional probability of consumer n choosing alternative i is obtained by integrating out β_n over its population distribution:

$$P_{ni}(\theta) = \int L_{ni}(\beta) g(\beta|\theta)\,\mathrm{d}\beta, \tag{5}$$

where $g(\cdot)$ is the density of the distribution of β_n over the population and θ is the vector of parameters of the distribution. For instance, if the joint distribution of the $(K \times 1)$ vector β_n is multi-variate normal, then θ would represent the mean and parameters of the covariance matrix of the joint distribution. The vector θ is an unknown to be estimated.

2.2. Maximum simulated likelihood estimation

To estimate the mixed logit model under maximum simulated likelihood, we construct the log-likelihood by calculating each individual's probability P_{ni} of making her observed choice. When the integral in the expression for P_{ni} does not have a closed form, the probability is often evaluated numerically by taking R draws $\beta^{(r)}$ (with $r = 1,2, \ldots, R$) from the population density $g(\beta)$ and calculating $L_{ni}(\beta^{(r)})$ for each draw. The average of $L_{ni}(\beta^{(r)})$ over R draws gives the simulated probability:

$$\hat{P}_{ni}(\theta) = \frac{1}{R}\sum_{r=1}^{R} L_{ni}(\beta^{(r)}). \tag{6}$$

This simulated probability is an unbiased estimator whose variance decreases as the number of draws R increases; it is smooth (twice-differentiable) and sums to one over all alternatives (Train, 2003). Since it is strictly positive, its logarithm is defined.

The simulated log-likelihood $SL(\theta)$ of the sample is the sum of the logarithm of the simulated probabilities for each consumer making her observed choice:

$$SL(\theta) = \sum_{n=1}^{N}\sum_{j=1}^{J_n} d_{nj} \log \hat{P}_{nj}(\theta), \tag{7}$$

where d_{nj} equals 1 if consumer n chose alternative j and 0 otherwise.

It is a well-known fact that although the simulated log-likelihood function is consistent when the number of draws increases with the sample size, it is simulated with a downward bias under a finite number of draws (Börsch-Supan and Hajivassiliou, 1993). Our paper focuses on a different phenomenon: namely, how the shape of the simulated log-likelihood is affected by the number of draws.

2.3. Methods of generating draws for simulation

We focus on three common procedures for generating draws from a density. The most straightforward approach obtains draws through a pseudo-random number generator available in most statistical software.

An alternative approach creates draws based on a deterministic Halton sequence (Halton, 1960). Train (1999, 2003) provide an explanation and an example of the construction of a Halton sequence. In general, a Halton sequence can be created from any prime number p. The unit interval [0,1] is divided into p equally sized segments, and the endpoints or "breaks" of these segments form the first p numbers in the Halton sequence. Successive numbers in sequence are generated by further subdividing each segment into p equally sized segments and adding the breaks in a particular order.

The resulting Halton draws achieve greater precision and coverage for a given number of draws than random draws, since successive Halton draws are negatively correlated and therefore tend to be "self-correcting" (Train, 2003). In fact, Bhat (2001) demonstrates that for a particular mixed logit model, 100 Halton draws provided results that were more accurate than 1000 random draws.

Since each Halton sequence is constructed from a prime number, each dimension of simulation corresponds to a different sequence or prime. For instance, if β_n is a 2×1 vector where the two components are independently distributed, then the first component of the vector is generated from a Halton sequence based on the prime 2, and the second component is generated from a Halton sequence based on the prime 3. Higher dimensions of simulation require using higher primes. Unfortunately, under higher primes, Halton draws can become highly correlated, leading to "poor multi-dimensional coverage" (Walker, 2001; Hess and Polak, 2003b).

To ameliorate the poor coverage of Halton draws for higher dimensions, researchers have adopted modified procedures such as shuffled and scrambled Halton draws. Hess and Polak (2003a, b) describe the construction of the shuffled sequence, which creates multi-dimensional sequences from randomly shuffled versions of the one-dimensional standard Halton sequence. They find that the shuffled Halton sequence typically outperforms the standard Halton and scrambled Halton sequences.

3. A simple example of theoretical unidentification

3.1. Telephone services

As a way of introducing the problem of interest, we present an example of a model that is not theoretically identified, and we show estimation results in which an identification problem is masked at low numbers of simulation draws.

This example uses a dataset of households' choices over telephone services. The choice set comprises five alternatives, which are categorized into two groups: flat and measured. The three flat alternatives consist of a fixed monthly charge for unlimited calls within a specified geographic area, and the two measured alternatives include a reduced fixed monthly charge for a limited number of calls as well as usage charges for additional calls. A consumer's tastes over the five alternatives are assumed to be correlated within groups (or nests). Consumer n's utility from choosing telephone service i is given by

$$U_{ni} = X_{ni}\alpha + NEST_i\,\beta_n + \varepsilon_{ni}. \tag{8}$$

X_{ni} is a (1×5) vector consisting of four alternative-specific constants and the log of the cost of the service, and α is the (5×1) vector of associated taste parameters. The variable $NEST_i$ is a (1×2) vector of dummies for each nest. The parameter β_n is a (2×1) vector consisting of $\beta 1_n$ distributed N(0, σ_1^2) and $\beta 2_n$ distributed N(0, σ_2^2) where $\beta 1_n$ and $\beta 2_n$ are independent. More detail on the dataset and model can be found in Train et al. (1987) and Walker (2001).

Walker (2001) discusses conditions for identification of the model and shows that only the value $(\sigma_1^2+\sigma_2^2)$ is identified. That is, when exactly two nests exist, only one nesting parameter is identified and to estimate the model an identifying constraint must be imposed, e.g., $\sigma_1 = \sigma_2$, $\sigma_1 = 0$, or $\sigma_2 = 0$.

Table 1 shows the estimation results[1] for a specification that does not include the necessary identifying constraint. Even without a necessary identifying restriction, the estimation procedure generates estimates that appear identified under a low number of

[1] All estimation results in this paper were estimated using either (1) BIOGEME using the DONLP2 optimization routine (see Bierlaire et al., 2004, and http://roso.epfl.ch/biogeme) or (2) a MATLAB implementation of Kenneth Train's GAUSS code using a BHHH-algorithm (Berndt et al., 1974) to compute the Hessian (for further information, see Chiou, 2005). Both estimation programs were shown to produce similar results.

Table 1
The demand for telephone service

Draws	2000 Random		5000 Random		1000 Halton		2000 Halton	
Parameter	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Alternative specific constants								
Budget measured (1)	−3.81	(0.66)	−3.81	(0.66)	−3.80	(0.67)		
Standard measured (2)	−3.01	(0.61)	−3.01	(0.61)	−3.01	(0.61)	Singular	
Local flat (3)	−1.09	(0.30)	−1.09	(0.30)	−1.09	(0.30)		
Extended flat (4)	−1.19	(0.85)	−1.19	(0.85)	−1.19	(0.85)		
Log cost	−3.25	(0.53)	−3.26	(0.53)	−3.25	(0.53)		
σ_1	0.81	(0.81)	3.07	(1.06)	2.65	(0.85)		
σ_2	2.91	(0.94)	0.24	(1.20)	1.51	(0.69)		
$(\sigma_1^2+\sigma_2^2)^{1/2}$	3.02		3.08		3.05			
Simulated log-likelihood	−472.73		−472.66		−473.02			
Number of observations	434		434		434		434	

Note: Uses robust standard errors.

draws (1000 Halton, 5000 pseudo-random). A large number of simulation draws (in this case, 2000 Halton draws) are necessary before resulting in a singular Hessian.

Not realizing the identification condition can lead to incorrect conclusions drawn from hypothesis tests based on standard errors. Since the parameter estimates and standard errors are poorly approximated under a low number of draws, they are a function of the specific draws and the starting values that are used. For example, the 2000 pseudo-random draw results for the telephone dataset would lead the modeler to incorrectly conclude that there is no correlation within the first nest (measured), but there is correlation within the second nest (flat). However, the estimation results under 5000 pseudo-random draws lead to the opposite conclusion (correlation among the measured alternatives but not among the flat alternatives).

4. Empirical unidentification and the log-likelihood function

In this section, we use extremely simple, synthetic datasets to examine the source of the empirical unidentification by investigating the properties of the simulated log-likelihood function as the number of draws increase. We show that regardless of whether the model is empirically identified, the simulated log-likelihood for mixed logit is always globally concave under only one draw because the model is analogous to a standard logit. When a model is not empirically identified, the simulated log-likelihood function begins to flatten and exhibit a singular Hessian only as the number of draws increases. The obfuscation of the identification problem occurs in the intermediate cases where the log-likelihood still exhibits the concavity as when only one draw is used. For the discussion, we consider two examples of the most common ways in which mixed logit is applied: first error components with a nesting formulation and then random coefficients on continuous explanatory variables.

4.1. Random coefficient on a nest dummy

We consider a simplified case where only one parameter is estimated. The discrete choice model consists of five alternatives that are divided into two nests, and the first three alternatives comprise Nest 1. Consumer n's utility of choosing alternative i is given by

$$U_{ni} = NEST1_i \beta_n + \varepsilon_{ni}, \tag{9}$$

where $NEST1$ is a dummy for whether alternative i lies in Nest 1. We specify the random coefficient β_n with a normal distribution, N(0, σ^2); thus only one parameter σ must be estimated. Strictly speaking, the resulting standard deviation is calculated as $\sqrt{\sigma^2}$, and therefore the sign of the estimated coefficient σ is irrelevant. As the number of draws approaches infinity, the simulated log-likelihood will be symmetric about zero.

We generate the synthetic data according to the true value $\sigma = 2.0$ by creating observations for N consumers. For each consumer, we calculate the utility of each alternative by taking a single draw of β_n from the N(0, σ^2) distribution and ε_{ni} from the extreme value distribution; the consumer chooses the alternative with the highest utility.

Table 2 reports the estimates of the standard deviation σ when the dataset contains only $N = 50$ observations. Under lower number of draws, the estimation procedure converges to a non-singular Hessian, but under a higher number of draws, the empirical unidentification becomes apparent. Figs. 1 and 2 graph the simulated log-likelihood as a function of the parameter σ for a varying number of random and Halton draws.

As shown in Figs. 1 and 2, the simulated log-likelihood is always concave when only one draw is used. In general, including only one draw in the simulation is equivalent to estimating a standard logit model with an additional variable included in the utility equation. Consider a model with a single explanatory variable X_{ni} and a random coefficient β_n. We express consumer n's utility from alternative i as $U_{ni} = X_{ni}\beta_n + \varepsilon_{ni}$. If the number of draws R is equal to one, then the simulated probability of consumer n choosing alternative i is

$$\hat{P}_{ni} = \frac{\exp(X_{ni}\beta_n)}{\sum_{j=1}^{J_n} \exp(X_{nj}\beta_n)}. \tag{10}$$

In the estimation procedure, the random coefficient for a single draw is decomposed as $\beta_n = \bar{\beta} + \sigma \upsilon_n$ where υ_n, $n = 1, \ldots, N$, are independent draws from a standard normal distribution. The coefficients to be estimated are $\bar{\beta}$ and σ, which are the population mean and standard deviation for the random taste. Substituting this expression into (10), we obtain:

$$\hat{P}_{ni} = \frac{\exp(X_{ni}(\bar{\beta} + \sigma\upsilon_n))}{\sum_{j=1}^{J_n} \exp(X_{nj}(\bar{\beta} + \sigma\upsilon_n))} = \frac{\exp(X_{ni}\bar{\beta} + X_{ni}\upsilon_n\sigma)}{\sum_{j=1}^{J_n} \exp(X_{nj}\bar{\beta} + X_{nj}\upsilon_n\sigma)} = \frac{\exp(X_{ni}\bar{\beta} + W_{ni}\sigma)}{\sum_{j=1}^{J_n} \exp(X_{nj}\bar{\beta} + W_{nj}\sigma)}. \tag{11}$$

where $W_{ni} = X_{ni}\upsilon_n$ is defined as a "new" variable created from X_{ni} and the particular draw for the consumer n. This is the standard logit formula where fixed parameters $\bar{\beta}$ and σ are estimated over the variables X_{ni} and W_{ni}. Since the standard logit model is globally concave (Train, 2003), estimation will always return a non-singular Hessian.

In our example, a draw υ_n is taken from a N(0,1) distribution, and β_n is calculated as $\beta_n = \sigma\upsilon_n$. The "new" variable is $\upsilon_n NEST1_i$. In other words, we can reinterpret the utility function as

$$U_{ni} = \sigma(NEST1_i\,\upsilon_n) + \varepsilon_{ni}, \tag{12}$$

Table 2
Empirical unidentification with a random coefficient on a nest dummy

	True value	1 Random	10 Random	35 Random	100 Random	1000 Random	1 Halton	10 Halton	35 Halton	100 Halton
σ	2.0	0.065	0.763	4.556	No convergence	No convergence	0.168	21.456	No convergence	No convergence
		(0.254)	(0.848)	(9.050)			(0.279)	(101.142)		
Simulated log-likelihood		−80.44	−80.26	−78.58	—	—	−80.32	−78.73	—	—
Number of observations		50	50	50	50	50	50	50	50	50

Notes: Standard error in parentheses. Uses non-robust standard errors.

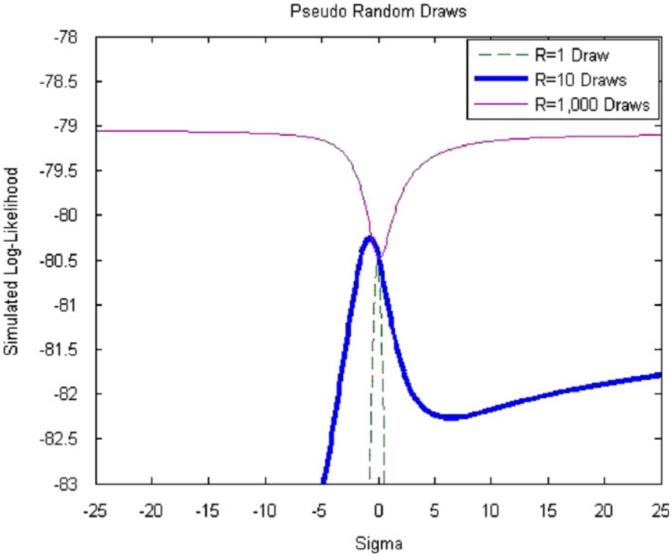

Fig. 1. Unidentified model with random coefficient on a nest dummy (50 observations, random draws).

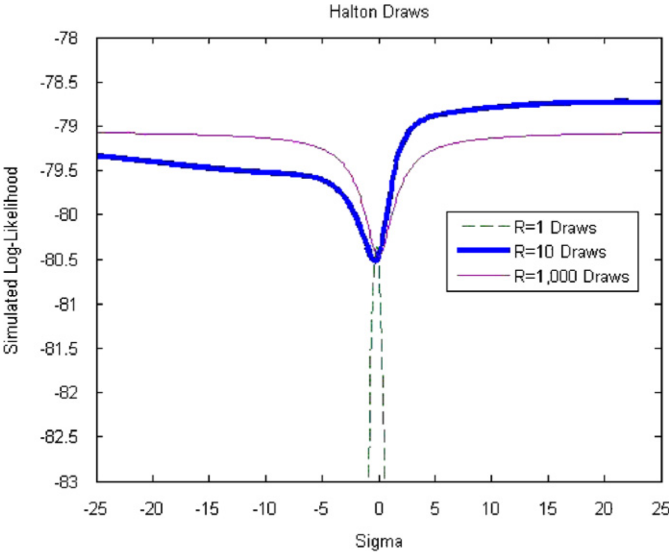

Fig. 2. Unidentified model with random coefficient on a nest dummy (50 observations, Halton draws).

where σ is a fixed coefficient on the variable $\upsilon_n NEST1_i$. Not surprisingly, the local maximum when 1 draw is used occurs near the origin, since the purely random draw has no explanatory power.

On the other hand, the singularity of the Hessian is evident under 1000 random draws. In Fig. 1, the simulated log-likelihood function rises away from 0, reflecting that the true value of σ is not zero, but the log-likelihood function flattens at higher magnitudes of σ. The data cannot empirically distinguish among the higher magnitudes of σ. In the

Table 3
Empirical identification with a random coefficient on a Nest dummy

	True value	1 Random	10 Random	100 Random	1000 Random	1 Halton	10 Halton	100 Halton
σ	2.0	0.003	0.152	1.770	2.378	0.004	2.112	2.464
		(0.020)	(0.064)	(0.223)	(0.319)	(0.020)	(0.274)	(0.336)
Simulated Log-likelihood		−16094.37	−16092.56	−16071.31	−16052.96	−16094.36	−16059.71	−16049.65
Number of observations		10,000	10,000	10,000	10,000	10,000	10,000	10,000

Notes: Standard error in parentheses. Uses non-robust standard errors.

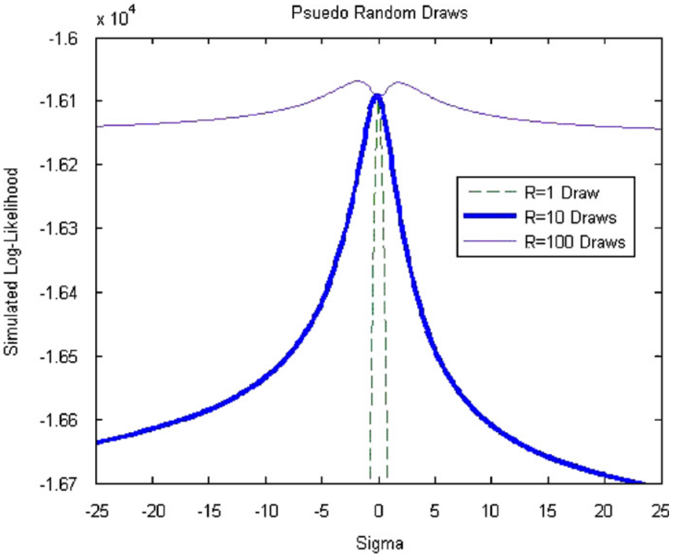

Fig. 3. Identified model with random coefficient on a nest dummy (10,000 observations, random draws).

intermediate case of 10 random draws, the simulated log-likelihood still exhibits a single peak as in the case of 1 draw. The local concavity gives rise to a convergence of the maximization routine.

Due to the efficiency of Halton draws relative to random draws, the Halton draws achieve the same unidentification properties at a lower threshold. In Table 2, the singularity of the Hessian occurs at 35 Halton draws whereas 35 random draws still generate a local maximum. Moreover, the large standard error of 101.142 under 10 Halton draws suggests the presence of an identification problem.

Table 3 reports the estimates of the standard deviation σ when the dataset contains $N = 10,000$ observations. In contrast to the previous case of only 50 observations, the dataset of 10,000 observations is sufficient to empirically identify the model. The parameter estimates stabilize and approach the true value of $\sigma = 2.0$ as the number of draws increases. Figs. 3 and 4 graph the simulated log-likelihood as a function of the parameter σ for a varying number of random and Halton draws, indicating a unique maximum (disregarding the sign) even for large number of draws.

4.2. Random coefficients on continuous variables

The dataset generated for this example uses three alternatives ($i = 1, 2, 3$) and two explanatory variables ($X1$ and $X2$). Unlike the previous example with a nest dummy, the two explanatory variables here are drawn from a continuous distribution. The utility function is as follows:

$$U_{ni} = X1_{ni}\beta_{1n} + X2_{ni}\beta_{2n} + \varepsilon_{ni}. \qquad (13)$$

The parameters β_{1n} and β_{2n} are independently and identically distributed N(0,σ_k), resulting in two parameters to estimate: σ_1 and σ_2. As with the previous case, two different synthetic datasets were generated that produced an empirically unidentified and identified

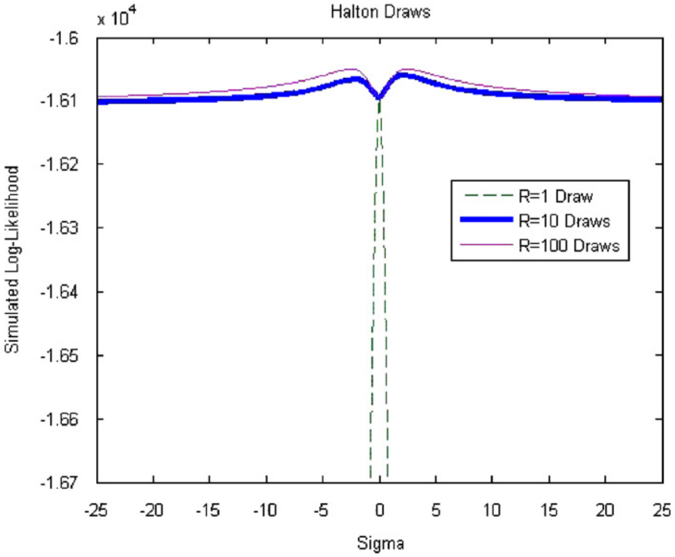

Fig. 4. Identified model with random coefficient on a nest dummy (10,000 observations, Halton draws).

model. The true values used for the parameters were $\sigma_1 = 1$ and $\sigma_2 = 1$. For the unidentified model, 100 observations were generated, and $X1$ and $X2$ were drawn from a bivariate normal distribution (variance of each equal to 1.0, covariance equal to 0.5, and non-zero means). The identified model used an identical specification except that 10,000 observations were generated, and the covariance of $X1$ and $X2$ was set to 0.[2]

For this example, only pseudo-random draws were used. Table 4 provides the results for the unidentified model. Fig. 5 plots the maximum of the simulated log-likelihood over σ_2 with respect to a given value of σ_1 for 1 draw, 5 draws, and 500 draws. First note that the model is, indeed, unidentified as indicated by the exploding parameter estimates and flat log-likelihood function when 500 draws are used in estimation. However, note that for a low number of draws (5, in this case), the simulated log-likelihood exhibits two local maxima on each side of zero, and therefore estimation of the model with 5 draws results in a seemingly identified model.

Table 5 and Fig. 6 report an analogous set of results for an identified model estimated with 10,000 observations. The model is identified as indicated by stable parameter estimates for 1000 and 2000 draws; these estimates are also close to the true values used to generate the synthetic data.

4.3. Discussion of simulated log-likelihood analysis

The simulated log-likelihood functions for mixed logit models are not as well behaved as a standard logit model; mixed logit likelihood functions are not globally concave and are sensitive to poor approximations of the simulated probabilities for low number of draws.

[2] *Cov(X*1,*X*2) was set to 0.0 in order to be able to visually see curvature in the log-likelihood plots in Fig. 2. However, 10,000 observations using the same specification as the "unidentified example" (i.e., with *Cov*(*X*1,*X*2) = 0.5) is, indeed, identified.

Table 4
Empirical unidentification with random coefficients on continuous variables

	True value	1 Random	5 Random	500 Random
σ_1	1.0	0.002	0.698	5972.361
		(0.017)	(0.222)	(67105.679)
σ_2	1.0	0.042	−0.305	2147.465
		(0.066)	(0.212)	(24133.121)
Simulated log-likelihood		−109.6	−94.1	−85.5
Number of observations		100	100	100

Notes: Standard error in parentheses. Uses robust standard errors.

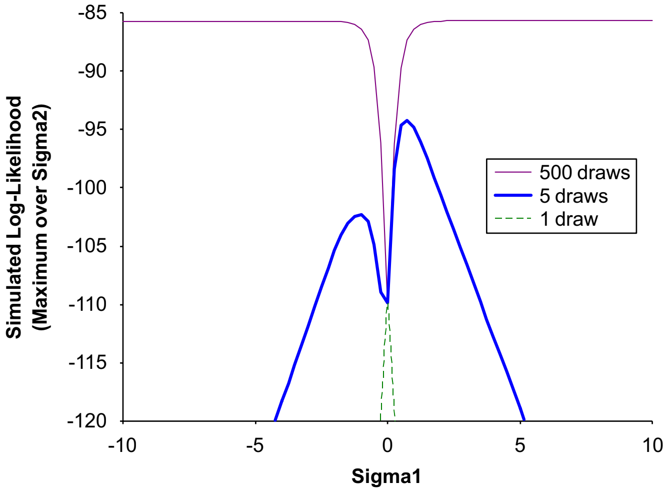

Fig. 5. Unidentified model with random coefficient on normally distributed variable (100 observations, random draws). *Note*: This figure plots the maximum of the simulated log-likelihood over Sigma2 for a given value of Sigma1.

Table 5
Empirical identification with random coefficients on continuous variables

	True value	1 Random	25 Random	1000 Random	2000 Random
σ_1	1.0	0.000	−0.496	0.909	0.920
		(0.002)	(0.020)	(0.123)	(0.127)
σ_2	1.0	−0.001	0.123	0.866	0.877
		(0.007)	(0.059)	(0.178)	(0.181)
Simulated Log-likelihood		−10,986	−10,083	−9917	−9911
Number of observations		10,000	10,000	10,000	10,000

Notes: Standard error in parentheses. Uses robust standard errors.

The results from these simple examples show that for a small number of draws, an empirically unidentified model can appear identified. It is only after a sufficient number of draws is used that the shape of the log-likelihood reveals the singularity of the Hessian.

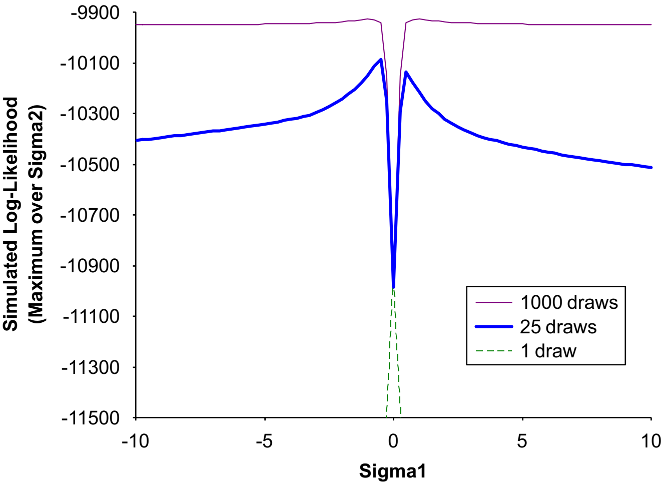

Fig. 6. Identified model with random coefficient on normally distributed variable (10,000 observations, random draws). *Note*: This figure plots the maximum of the simulated log-likelihood over Sigma2 for a given value of Sigma1.

5. Examples of empirical unidentification

The simple examples from the previous section were used to explore the source of empirical unidentification and to demonstrate the behavior of the log-likelihood function under different numbers of draws and under different conditions of identification. This section considers estimation results for more realistic datasets, including one real dataset regarding households' purchases of DVDs and one synthetic dataset.

5.1. Retail stores

In the first example, we apply data from Chiou (2005) to examine a household's choice of retail store to purchase a DVD. We estimate a consumer's choice of store conditional on the purchase of a DVD. The consumer's choice set consists of retail stores from the top 15 chains that sell DVDs, and each retail store is classified under one of five store types: mass merchant, video specialty, electronics, music, and online. Consumer *n*'s utility from traveling to store *i* to purchase her chosen video is given by

$$U_{ni} = X_{ni}\alpha + \sum_{k=1}^{5} TYPE_{ik}\,\beta_{nk} + \varepsilon_{ni}, \tag{14}$$

where X_{ni} contains interactions of observable store and consumer characteristics (such as price, distance to store, income, education), and $TYPE_i$ is a vector of store type dummies. The explanatory variables in vector X_{ni} do not have random coefficients ($\alpha_n \equiv \alpha$ for all n); the vector X_{ni} contains variables for which tastes are constant across the population. In contrast, consumers' marginal utilities over store types vary by unobservable consumer characteristics. More specifically, the coefficients (β_n) on the store type dummies are assumed to be independently and normally distributed. The random coefficients can be expressed as

$$\beta_{nk} = \bar{\beta}_k + \sigma_k \upsilon_{nk} \quad \text{with } k = 1, 2, \ldots, 5, \tag{15}$$

where k indexes the store's type, and υ_{nk} are independent standard normal variables. The coefficients $\bar{\beta}_k$ and σ_k are the population mean and standard deviation for the marginal utility of store type k. Note that the standard deviations are allowed to differ by store type. The unknown parameters to be estimated are α, $\bar{\beta}$ and σ_k (for $k = 1,2, \ldots, 5$). For further details on the data, specification, and estimation procedure, refer to Chiou (2005).[3]

Table 6 presents the results under 1, 100, 200, and 1000 draws with pseudo-random, Halton, and shuffled Halton draws.[4] Although the model specification is theoretically identified, the results with high numbers of draws indicate that the parameters are not empirically identified by the data. Nonetheless, optimizations from 200 random or 100 Halton draws converge and generate estimates that appear identified. Without checking the robustness of the estimates to varying number of draws, the unidentification issue would not be apparent. For instance, under 200 random draws, the mean and standard deviation of the population distribution of tastes over video specialists are 19.391 (6.381) and 1.501 (0.446), and the coefficients are significant at the 1% level. Similarly, under 100 Halton draws, the mean and standard deviation of the random coefficient on video specialists are 19.576 (8.817) and 1.244 (0.507).

The identification issue becomes readily apparent under 100 shuffled Halton draws; the optimization routine does not converge as parameter estimates explode. The results suggest that shuffled Halton draws expose the identification issue at a lower number of draws relative to other types of draws.

5.2. Synthetic data

The above retail and synthetic datasets consist of relatively simple specifications. In this section, we present an example to illustrate how simulation difficulties become more apparent as additional complexities are introduced into the model. The example uses a dataset that consists of 2000 observations, each making a choice among 4 alternatives. The utility function is given by

$$U_{ni} = X_{ni}\beta_n + \varepsilon_{ni}, \tag{16}$$

where X_{ni} is a (1×8) vector containing 3 alternative specific constants and 5 explanatory variables. The explanatory variables are drawn from a multivariate normal distribution (means ranging from 1 to 3, variances approximately equal to 1, and covariances ranging from 0.0 to 0.6). The random coefficient β_n is (8×1), and each parameter is independent and normally distributed, $\mathrm{N}(\bar{\beta}_k, \sigma_k^2)$. This specification provides for alternative specific variances as well as taste heterogeneity with respect to the five explanatory variables. The variance for the fifth explanatory variable was set to 0.0, based on the conventional wisdom to fix at least one parameter in estimation. The true values of the parameters used for data generation are provided in Table 7, along with estimation results.

[3] Chiou (2005) estimates a mixed nested logit which is analogous to constraining the standard deviations of the marginal utilities across all store types to be equal.

[4] The shuffling procedure is not valid when there is only 1 draw. See Hess and Polak (2003b) for the shuffled Halton procedure.

Table 6
Retail stores

		Parameter	1 Draw		100 Draws		200 Draws		1000 Draws	
			Estimate	Std. error	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Random	Mass									
	Mean	$\bar{\beta}_1$	5.281	(0.656)	19.515	(7.510)	20.488	(6.386)		
	Std. dev.	σ_1	0.032	(0.040)	0.419	(0.375)	0.038	(0.547)		
	Video									
	Mean	$\bar{\beta}_2$	4.672	(0.677)	18.530	(7.504)	19.391	(6.381)	No convergence	
	Std. dev.	σ_2	0.004	(0.056)	1.209	(0.405)	1.501	(0.446)		
	Electronics									
	Mean	$\bar{\beta}_3$	2.077	(0.851)	16.026	(7.582)	17.262	(6.439)		
	Std. dev.	σ_3	0.086	(0.133)	0.878	(0.930)	0.487	(1.636)		
	Music									
	Mean	$\bar{\beta}_4$	5.631	(0.671)	19.726	(7.516)	20.816	(6.391)		
	Std. dev.	σ_4	0.069	(0.048)	1.317	(0.351)	0.967	(0.373)		
	Online									
	Mean	$\bar{\beta}_5$	0	—	0	—	0	—		
	Std. dev.	σ_5	0.021	(0.140)	8.873	(3.406)	8.737	(2.818)		
	Simulated Log-likelihood		−5254.20		−5229.80		−5231.59			

Halton	Mass							
	Mean	$\bar{\beta}_1$	5.282	(0.657)	20.559	(8.812)		
	Std. dev.	σ_1	0.020	(0.040)	0.183	(0.934)		
	Video							
	Mean	$\bar{\beta}_2$	4.681	(0.678)	19.576	(8.817)	No convergence	No convergence
	Std. dev.	σ_2	0.095	(0.054)	1.244	(0.507)		
	Electronics							
	Mean	$\bar{\beta}_3$	2.084	(0.852)	17.437	(8.854)		
	Std. dev.	σ_3	0.002	(0.182)	0.057	(6.371)		
	Music							
	Mean	$\bar{\beta}_4$	5.639	(0.673)	20.739	(8.827)		
	Std. dev.	σ_4	0.023	(0.050)	1.358	(0.393)		
	Online							
	Mean	$\bar{\beta}_5$	0	—	0	—		
	Std. dev.	σ_5	0.042	(0.126)	9.442	(4.361)		
	Simulated log-likelihood		−5253.91		−5236.63			
Shuffled Halton			—		No convergence		No convergence	No convergence
Number of observations			3132		3132		3132	3132

Note: Uses non-robust standard errors.

Table 7
Synthetic results

	Parameter	True value	250 Draws		1000 Draws		2000 Draws	
			Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Random	ASC1_Mean	0.10	−0.032	(0.154)	−0.244	(0.500)	−0.559	(0.656)
	ASC1_StDev	1.00	0.067	(0.722)	−1.133	(1.296)	1.944	(1.582)
	ASC2_Mean	0.10	−0.094	(0.370)	−0.581	(0.817)	−0.080	(0.543)
	ASC2_StDev	1.00	−0.838	(1.051)	2.024	(1.594)	−1.132	(1.880)
	ASC3_Mean	0.20	0.066	(0.220)	0.029	(0.228)	−0.350	(0.742)
	ASC3_StDev	1.00	0.428	(1.038)	0.320	(1.160)	1.996	(2.206)
	Var1_Mean	0.05	0.006	(0.043)	0.002	(0.051)	0.021	(0.068)
	Var1_StDev	0.50	0.430	(0.120)**	0.571	(0.198)**	0.747	(0.438)
	Var2_Mean	0.05	0.118	(0.048)**	0.140	(0.065)**	0.154	(0.092)
	Var2_StDev	0.50	0.382	(0.129)**	0.411	(0.198)**	0.424	(0.262)
	Var3_Mean	0.05	0.054	(0.047)	0.066	(0.071)	0.104	(0.090)
	Var3_StDev	0.50	0.490	(0.103)**	0.634	(0.249)**	0.751	(0.384)
	Var4_Mean	0.10	0.020	(0.046)	0.009	(0.059)	0.032	(0.068)
	Var4_StDev	0.50	0.338	(0.139)**	0.494	(0.221)**	0.474	(0.292)
	Var5_Mean	0.02	0.084	(0.045)*	0.119	(0.064)*	0.111	(0.080)
	Sim. log-likelihood		−2726.07		−2725.08		−2724.91	
Halton	ASC1_Mean	0.10	−0.410	(0.654)	−0.901	(1.544)	−0.402	(0.665)
	ASC1_StDev	1.00	−1.442	(1.504)	3.210	(4.805)	−1.414	(1.552)
	ASC2_Mean	0.10	−0.097	(0.762)	−0.540	(1.718)	−0.094	(0.506)
	ASC2_StDev	1.00	0.927	(2.609)	2.892	(5.654)	0.897	(1.666)
	ASC3_Mean	0.20	−0.003	(0.421)	−0.881	(1.650)	0.055	(0.286)
	ASC3_StDev	1.00	−0.687	(2.021)	3.780	(5.558)	0.296	(2.605)
	Var1_Mean	0.05	0.004	(0.050)	0.045	(0.124)	0.003	(0.048)
	Var1_StDev	0.50	0.546	(0.276)**	1.222	(1.659)	0.532	(0.231)
	Var2_Mean	0.05	0.129	(0.079)*	0.240	(0.318)	0.124	(0.062)
	Var2_StDev	0.50	0.374	(0.180)**	0.630	(0.843)	0.378	(0.181)
	Var3_Mean	0.05	0.081	(0.071)	0.159	(0.236)	0.081	(0.072)
	Var3_StDev	0.50	0.613	(0.310)**	1.148	(1.489)	0.620	(0.305)
	Var4_Mean	0.10	0.023	(0.055)	0.046	(0.109)	0.020	(0.053)
	Var4_StDev	0.50	−0.400	(0.294)	−0.858	(1.341)	0.390	(0.231)
	Var5_Mean	0.02	0.101	(0.078)	0.181	(0.275)	0.097	(0.062)
	Sim. Log-likelihood		−2725.88		−2725.64		−2726.30	
Shuffled Halton	ASC1_Mean	0.10	−0.746	(0.561)	−0.814	(1.362)		
	ASC1_StDev	1.00	2.204	(1.066)**	2.423	(3.090)		
	ASC2_Mean	0.10	−0.116	(0.402)	−0.312	(0.667)		
	ASC2_StDev	1.00	−1.061	(1.138)	1.674	(2.129)	No convergence	
	ASC3_Mean	0.20	0.008	(0.269)	−0.092	(0.617)		
	ASC3_StDev	1.00	−0.778	(0.897)	1.163	(2.805)		
	Var1_Mean	0.05	0.001	(0.054)	0.010	(0.065)		
	Var1_StDev	0.50	0.596	(0.202)**	0.698	(0.556)		
	Var2_Mean	0.05	0.140	(0.065)**	0.150	(0.113)		
	Var2_StDev	0.50	0.380	(0.187)**	0.435	(0.302)		
	Var3_Mean	0.05	0.102	(0.069)	0.114	(0.129)		
	Var3_StDev	0.50	0.716	(0.216)**	0.830	(0.696)		

Table 7 (*continued*)

	Parameter	True value	250 Draws		1000 Draws		2000 Draws	
			Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
	Var4_Mean	0.10	0.024	(0.058)	0.022	(0.068)		
	Var4_StDev	0.50	−0.431	(0.194)**	0.486	(0.403)		
	Var5_Mean	0.02	0.105	(0.060)*	0.124	(0.101)		
	Sim. log-likelihood		−2725.07		−2726.13			
Number of observations			2000		2000		2000	

**Significant at 5% level of significance.
*Significant at 10% level of significance.
Note: Uses robust standard errors.

The estimation results appear identified and statistically significant (particularly on the distribution for the random coefficients) for 250 draws, whether they are pseudo-random, Halton, or shuffled Halton. The parameter estimates appear to be fairly stable at 1000 draws, but the standard errors rise significantly, resulting in insignificant parameter estimates in all but the pseudo-random case and suggesting a problem. At 2000 shuffled Halton draws, the model does not converge as the parameters began to explode with successive iterations. The results also indicate that the pseudo-random case is heading in this direction (albeit more slowly) in that successive increases in draws result in increasing standard errors. The Halton results do not exhibit such a trend, and this may be due to either starting values (therefore resulting in different local maximum) or because of correlations among the Halton sequences for this 8-dimensional integral.

6. Conclusion

With advancements in computational speed, simulation techniques have vastly improved the ability to estimate complex models to answer a myriad of questions. However, the implementation of simulation can often mask problems of identification. A low number of draws can result in estimates that appear identified, but in fact are not identified either theoretically by the model or empirically by the data.

To highlight the issue, we present examples of maximum simulated likelihood estimation of mixed logit models under actual and synthetic datasets. Although each of these models was unidentified, estimation with too few draws resulted in coefficients that appeared identified. The first was a telephone dataset that was theoretically unidentified, but the problem was masked at 5000 pseudo-random draws and 1000 Halton draws. The second was a model using a retail dataset of DVD store purchases, which was empirically unidentified, but the problem was masked at 200 pseudo-random draws and 100 Halton draws. The third was a model using a synthetic dataset that was empirically unidentified, but the problem was masked at 2000 pseudo-random draws, 2000 Halton draws, and 1000 shuffled Halton draws. These estimation results also provide further evidence of the benefit of variance reduction methods such as shuffled Halton, which were shown to uncover identification issues at a lower number of draws.

The important lesson is that it is critical not to stop at 200 draws, whether pseudo-random, Halton, or shuffled Halton. No general rule of thumb exists for what constitutes a "high" or "low" number of draws as it depends on the data, specification and type of draw. One has to verify the stability of the parameter estimates as well as the standard errors as the number of draws increases.

In addition to the empirical results, the underlying source of these issues was investigated by examining the shape of the log-likelihood function under varying numbers of draws and different identifying conditions. We demonstrate that under one draw, the mixed logit model is equivalent to a standard logit model, and, therefore, the simulated log-likelihood is globally concave and always uniquely identified. Thus, for an unidentified model, the simulated log-likelihood is globally concave under one draw, and only a sufficient number of draws reveals the flatness of the likelihood function and results in either exploding parameter estimates or a singular Hessian.

While we examined the case of mixed logit models under maximum simulated likelihood, the findings can be extended to other discrete choice models and simulation estimators. This applies, for example, to mixed logit under method of simulated moments as well as generalized extreme value models under maximum simulated likelihood. The number of draws affects the precision of the numerical integration and, therefore, the simulated log-likelihood. The obfuscation that leads to masked identification occurs for any situation in which the model is estimable under one draw regardless of whether the model is identified or not.

Acknowledgements

The authors would like to thank Moshe Ben-Akiva, Denis Bolduc, Victor Chernozhukov, Glenn Ellison, John Geweke, Nancy Rose, and anonymous referees for their insights into this work.

References

Ben-Akiva, M., Bolduc, D., 1996. Multinomial probit with a logit kernel and a general parametric specification of the covariance structure. Working paper, Massachusetts Institute of Technology.

Berndt, E., Hall, B., Hall, R., Hausman, J., 1974. Estimation and inference in nonlinear structural models. Annals of Economics and Social Measurement 3/4, 653–665.

Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. Econometrica 63 (4), 841–890.

Bhat, C., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research B 35, 677–693.

Bierlaire, M., Bolduc, D., Godbout, M.-H., 2004. An introduction to BIOGEME (Version 1.0). Working paper, Ecole Polytechnique Federale de Lausanne.

Börsch-Supan, A., Hajivassiliou, V., 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. Journal of Econometrics 58, 347–368.

Brownstone, D., Train, K., 1999. Forecasting new product penetration with flexible substitution patterns. Journal of Econometrics 89, 109–129.

Chiou, L., 2005. Empirical analysis of retail competition: spatial differentiation at Wal-Mart, Amazon.com, and their competitors. Working paper, Occidental College.

Goolsbee, A., Petrin, A., 2003. The consumer gains from direct broadcast satellites and the competition with cable television. Econometrica 72, 351–381.

Halton, J., 1960. On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik 2, 84–90.

Hensher, D., Greene, W., 2003. The mixed logit model: the state of practice. Transportation 30, 133–176.

Hess, S., Polak, J.W., 2003a. A comparison of scrambled and shuffled Halton sequences for simulation based estimation. Centre for Transport Studies Working paper, Centre for Transport Studies, Imperial College London.

Hess, S., Polak, J.W., 2003b. An alternative to the scrambled Halton sequence for removing correlation between standard Halton sequences in high dimensions. Paper presented at the European Regional Science Conference, Jyväskylä.

McFadden, D., Train, K., 2000. Mixed MNL models of discrete response. Journal of Applied Econometrics 15, 447–470.

Train, K., 1999. Halton sequences for mixed logit. Working paper, University of California, Berkeley.

Train, K., 2003. Discrete choice methods with simulation. Cambridge University Press, New York.

Train, K., McFadden, D., Ben-Akiva, M., 1987. The demand for local telephone service: a fully discrete model of residential calling patterns and service choices. RAND Journal of Economics 18 (1), 109–123.

Walker, J.L., 2001. Extended discrete choice models: integrated framework, flexible error structures, and latent variables. Ph.D. dissertation, Massachusetts Institute of Technology, Department of Civil and Environmental Engineering.

Walker, J.L., Ben-Akiva, M., Bolduc, D., 2006. Identification of parameters in normal error components logit mixture models. Journal of Applied Econometrics, forthcoming.