Estimation Performance of Low Discrepancy Sequences in Stated Preferences

Rodrigo A. Garrido
Department of Transport Engineering
Pontificia Universidad Católica de Chile
Santiago, Chile

Phone: 056-2-686 5893
Fax: 056-2-553 0281
eMail: rgarrido@ing.puc.cl

Abstract

A standard assumption in stated preferences modelling is the independence between repeated responses from an interviewee. The rationale for this strong assumption is that the mathematical treatment of the choices becomes rather cumbersome when dependency is incorporated in the analysis. Assuming dependence between the various responses given by one individual is known as the *repeated observations problem* and its mathematical formulation is similar to that of the autocorrelation in panel data surveys. These type of problems can only be studied with the aid of flexible models such as the multinomial probit (MNP) or mixed logit (MXL). However, these models need the computation of a multidimensional integral to obtain values for the choice probabilities. The integration process is usually complex. Therefore, approximation methods must be applied –typically simulation, to evaluate the choice probabilities. The standard simulation approach relies on the Monte Carlo (MC) method, which basically replaces a continuous average (the integral) by a discrete average over a set of points randomly distributed within the region of integration. The numerical analysis literature shows various procedures to choose *smart* points from a deterministic series instead of random realizations. This type of points are known as low discrepancy sequences (LDS). However, the use of these techniques in econometrics is rather limited and recent, and consequently there are several open questions to be answered before the use of LDS becomes a standard. The evidence found in the fields of mathematics and physics indicates that a LDS called the Sobol sequence, would be a superior alternative to the more known Halton sequences, especially for large dimensions. Nevertheless, the Sobol sequences (to the knowledge of the authors) have not been tested yet in transportation. This paper compares the MC simulation method in three versions: traditional, Halton based, and Sobol based for the estimation of the MXL.

Keywords

Monte Carlo, Sobol, Halton, International Conference on Travel Behaviour Research, IATBR

Preferred citation

Author1, First name author1, first name author2 autthor2, ... and first name author<n> author<n> (2003) Titel of the paper, paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003.

1. INTRODUCTION

The mixed logit (MXL) is a flexible strategy to estimate general discrete choice models (e.g. those that allow the analysis of a variety of complex situations such as an individual facing a selection among correlated alternatives, or the presence of heterogeneity, or a set of dynamic chained decisions, etc.) However, the MXL estimation requires the numerical computation of a complex multiple integral. The most common numerical methods to estimate multiple integrals can be classified as either polynomial cubature methods, or Monte Carlo (MC) simulation methods. The polynomial cubature is a generalization of the standard unidimensional cuadrature, its applicability is restricted to MXL specifications in small dimensions. The standard MC method simulates the multidimensional integral replacing a continuous average by a discrete one, using a series of (pseudo) randomly selected numbers to cover the integration space, and then averaging the distribution of points within the integration space.

The accuracy of the MC simulation method depends on the coverage and the number of points over the integration region. Therefore, for a fixed number of points to be used, they key issue is where to locate them to efficiently cover the integration space. Covering the integration region with pseudo random points requires a large number of draws to attain an even and complete distribution over the whole region, which may imply prohibitively large computational efforts.

Researchers in other fields outside transportation (e.g. finances, or computer sciences) have experimented with the use of non-random series for MC simulation. These series were designed to cover more uniformly the integration region. The generic name of these series is *low discrepancy series* (LDS) (Galanti and Jung, 1997; Ohbuchi and Aono, 1994; Paskov, 1994), also known as quasi-random sequences. The most common LDS are the Hammersley, Halton, Sobol, Faure and Niederreiter sequences.

Only a few papers dealing with LDS in econometrics can be found in the literature (Bhat, 2001a; Bhat, 2001b; Train, 1999). These researches have studied the Halton sequences but none has studied yet the Sobol sequences which have proven to be successful in other fields. Morokoff and Caflisch (1995) showed that Sobol sequences outperform the Halton sequences for integrals in more than six dimensions. Paskov (1994) concludes that Sobol based MC method outperforms both the standard MC and the Halton based MC methods. The aim of this paper is to study the performance of the Sobol sequences in MXL specifications.

The rest of the paper is structured as follows. Next section presents a MXL specification for repeated observations (stated preferences modeling). Section 3 introduces the MC method applied to the estimation of MXL using simulated maximum likelihood. Section 4 presents the LDS to be used in this research. Section 5 presents the experimental results. Finally, section 6 summarizes the conclusions.

2. MIXED LOGIT FOR SUCCESSIVE CHOICES

The specification to be summarized in this section has been previously applied by other authors to model heterogeneity, state dependence, and serial correlation effects in stated preferences experiments (Goett et al, 2000; Srinivasan and Mahmassani, 2000; Train, 1998).

The utility of alternative *i* for the individual *q* in the situation *t*, can be expressed as follows:

$$U_{iqt} = \beta_q^{'} \cdot X_{iqt} + \varepsilon_{iqt} \qquad (1)$$

where X_{iqt} is a vector of observed variables related to the alternative i for individual q in the choice situation t, (its dimensionality is K); β_q is a vector of coefficients representing personal tastes for each q, (its dimensionality is K), randomly distributed over the individuals with mean b covariance matrix W, (its dimensionality is KxK); ε_{iqt} is an error term Gumbel distributed, identical and independent over the individuals and choice situations. ε_q is defined as a vector with element ε_{iqt} such that $\varepsilon_q \sim Gumbel(0, \Sigma_g)$ with Σ_g diagonal covariance matrix with dimension JT x JT. The variance for the choice situation t is $\sigma_t^2 = \pi^2 / (6\mu_t^2)$, $t = 1,....T$ with μ_t the scale factor.

The coefficient's vector for each individual, β_q , can be expressed as follows:

$$\beta_q = b + \eta_q \qquad (2)$$

where b is the parameters mean value over the population and η_q is the individual deviation from the mean, representing personal preferences of each individual.

In particular if β_q takes the value β, the choice probability for *q* to choose the alternative *i* in the *t-th* choice situation would be given by:

$$L_{iqt}(\beta) = \frac{\exp(\mu_t(\beta' \cdot X_{iqt}))}{\sum_j \exp(\mu_t(\beta' \cdot X_{jqt}))} \quad (3)$$

where all the elements have been previously defined.

The probability that an individual makes a sequence of choices (conditioned on β) is given by:

$$P_{y_q}(\beta) = \prod_{t=1}^{T} L_{y_{qt}}(\beta) \quad (4)$$

where y_{qt} identifies the alternative chosen by q in the choice situation t, and y_q is the vector $\{y_{q1},.....y_{qT}\}$ that describes the choices sequence for each individual.

To obtain the unconditional choice probability, expression (4) must be integrated over the whole dominium of β:

$$P_{y_q}(b,W) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}.....\int_{-\infty}^{\infty} P_{y_q}(\beta) f(\beta/b,W)\, d\beta \quad (5)$$

The dimension of this integral equals the number of random parameters.

The corresponding log-likelihood function is given by:

$$L(b,W) = \sum_q \ln(P_{y_q}(b,W)) \quad (6)$$

3. MONTE CARLO SIMULATION

For given values of b and W, several values of β can be drawn from its probability distribution. Let β^n be a realization of β. With this value, the choice probabilities, conditioned on the n-th realization of β, can be computed using expression (4):

$$P_{y_q}(\beta^n) = \prod_{t=1}^{T} L_{y_{qt}}(\beta^n) \quad (7)$$

This process is repeated N times to approximate the (unconditional) choice probability as the average of (7) over the N realizations:

$$SP_{y_q}(b,W) = \frac{1}{N}\sum_{n=1}^{N} P_{y_q}(\beta^n) \quad (8)$$

where $SP_{y_q}(b,W)$ is the simulated probability for the choice sequence of individual q. Accordingly, the simulated log-likelihood function is given by:

$$SL(b,W) = \sum_q \ln(SP(y_q / b,W)) \quad (9)$$

The parameters are obtained with the aid of a non-linear optimization algorithm that maximizes expression (9).

4. LOW DISCREPANCY SEQUENCES

The concept on which the LDS are based is the spread of points, successively, in a position as far away as possible from the previously located points. This principle precludes the formation of clusters.

LDMC methods have been successfully applied to diverse fields such as physics, computational graphics, finances among others (Niederreiter, 1992; Morokoff and Caflisch, 1995). Their advantages in lower dimension problems has been established, but Morokoff and Caflisch (1995) point out that the precision and efficiency of LDMC diminishes as the dimensionality increases. This deterioration occurs due to the dependence of the points locations in higher dimensions.

I. Standard Halton Sequences

The standard Halton sequences (Halton, 1960) in S dimensions are obtained pairing S unidimensional sequences of Van der Corput (Niederreiter, 1992), based in S prime numbers, r_1, r_2,.....r_s (usually the first S prime numbers).

The sequence corresponding to the prime r has cycles of length r with numbers monotonically increasing. This feature makes the initial terms of two sequences highly dependent, at least to the first cycle of each sequence.

4.1 Sobol Sequences

The Sobol sequences (Sobol, 1967) solved the problem of dependency in higher dimensions (long prime numbers) using only the prime 2. The sequences are generated in such a way that the first 2^m terms of each dimension, for $m=0,1,2,....,$ are permutations of the corresponding terms of the Van der Corput (Niederreiter, 1992), sequence with base r=2. If an appropriate permutation is performed, the resulting S-dimensional sequence presents good uniformity (low discrepancy).

Several authors have concluded that the Sobol sequences appear to resist the degradation effect better than the Halton sequences in higher dimensions, Galanti and Jung (1997) showed that the Sobol sequences presented no degradation at all up to the dimension 260. Morokoff and Caflisch (1995) concluded that for lower dimensions (S = 6 or lower) the Halton sequences exhibited the best results, whereas for higher dimensions the Sobol sequences were better off. Cheng and Druzdzel (2000) tested Halton, Sobol and Faure sequences concluding that for higher dimensions the Sobol sequences outperformed the other two.

One way of destroying the higher dimension dependency is to scramble the numbers within each sequence, maintaining the equidistribution property.

The scrambling method implemented in this paper was proposed by Morokoff and Caflisch (1994), it performs pseudo-random arrangements on each dimension such that the new sequence discrepancy equals that of the original sequence

Tuffin (1996) points out that in the LDMC method it is difficult to estimate the integration error, due to the deterministic nature of the LDS. An easy estimation of the integration error can be found by randomizing the sequences, the process must preserve the uniformity and equidistribution. The method used in this paper was proposed by Owen (1995).

5. EXPERIMENTAL ANALYSIS

5.1 Construction of the Simulated Samples

Two samples were built under different MXL specifications, simulating a stated preference experiment. The utility functions considered random parameters and (explicit) correlation between repeated observations. Both samples have 1,000 individuals who make six choices each (i.e. 6,000 pseudo-individuals).

The first sample corresponds to a problem with five integration dimensions –similar to the first published applications of Halton sequences. Three alternatives were defined, characterized by five attributes. The attributes values were generated from a uniform distribution. The utility function parameters considered two constants and the five attributes coefficients were assumed normal.

The second sample corresponds to a problem with 10 integration dimensions, formed by three alternatives with 10 attributes whose coefficients follow a normal distribution. This case does not include specific constants.

5.2 Models Estimation

The estimation software was developed in C++ and it is fully described in Leva and Silva (2002). Even though some published studies perform the simulation starting from a vicinity of the actual parameters values, in this paper the starting points were not that close to the actual solution. The reason for the latter was to measure not only the integration error (standard deviation of various trials) but also the accuracy of the different methods under comparison.

For each sample, the specifications were estimated several times with a different number of repetitions and type of sequences (pseudo-random and LDS). In the case of Halton and Sobol sequences, the estimations were done using 50, 100, 150, and 250 repetitions; for the pseudo-random case 500, 750 and 1,000 repetitions were tried. For each type of sequence and number of repetitions, six sets of values were used for each observation.

To obtain the pseudo-random sequences, different seeds were fed into the generator code. For the Halton sequences, different values were generated permuting the prime numbers that originate each column of points for each dimension. For the Sobol case the sequences corresponding to each dimension were permuted producing the same effect.

5.3 Results for the Five-Dimensional Sample

The first comparison criterion is the mean-square error (MSE) applied to the estimated parameters, computed from the actual parameter values. The MSE were computed for the different simulated sets (six). If the MSE corresponding to 1,000 pseudo-random repetitions is defined as the base error, then the MSE of the LDS can be expressed as a percentage of that base. These values are shown in Table 1. There are practically no differences among the three sequences. Similar results were found by Train (1999), in fact, he found that the mean values

of the parameters, estimated with 100 Halton repetitions was statistically no different from that of 1,000 pseudo-random. The only conclusion that the MSE allows to draw is that 150 points from a LDS seem to be equivalent to 1,000 points from a pseudo-random sequence.

The second comparison criterion is the standard deviation of the parameter values over the various estimations (six in this paper) for each type of sequence and number of repetitions. Table 2 shows the average standard deviation for each case.

The results indicate that 150 Sobol repetitions reach the same level of accuracy than 1,000 pseudo-random repetitions. In addition, 250 repetitions yield standard deviations (in average) 15% lower than those corresponding to 1,000 pseudo-random repetitions.

As for the Halton sequences, 150 repetitions yield standard deviations (in average) 15% greater than 1,000 pseudo-random. Furthermore, 250 repetitions are not enough to reach the same level of accuracy of 1,000 pseudo-random repetitions (the Halton sequence standard deviations were, in average, 7% greater).

Models were also estimated using 500 repetitions of Halton and Sobol sequences. The standard deviation of the Halton sequences were 74% lower than that of 1,000 pseudo-random repetitions, reaching the same level of accuracy found for the Sobol sequences, which had a standard deviation 75% lower than that of 1,000 pseudo-random repetitions.

When a lower number of repetitions was tested, counterintuitive results were found. In fact, when the number of Halton repetitions were increased from 100 to 150, the standard deviation increased 6%; when increasing the number of Sobol repetitions from 50 to 100 the standard deviation augmented by 230%. Similar results have been reported by Train (1999) and Bhat (2001a).

5.4 Results for the Ten-Dimensions Sample

Table 3 presents the RMS results obtained for the different value sets. As in the five-dimension sample, there is no statistical difference among the values obtained from the different methods. Therefore, the relevant comparison criterion is the standard deviation, shown in Table 4. It was observed that 100 to 150 Sobol repetitions yield the minimum standard deviation, 80% lower than that of 1,000 pseudo-random repetitions.

For the Halton case, 150 repetitions reach standard deviations 49% lower than those obtained with 1,000 pseudo-random repetitions. Thus, it can be concluded that the Halton sequences are an inferior alternative when compared to the Sobol sequences.

6. Conclusions

In the experiments presented in this paper, the use of Sobol sequences for the estimation of MXL specifications emerged as the best available option. The Sobol sequences presented better resistance to the higher dimensions degradation and they covered the integration space more efficiently than both the pseudo-random sequences and the Halton sequences. In fact, 150 repetitions of the Sobol based Monte Carlo method, allowed to reach a better level of accuracy than that obtained with 1,000 pseudo-random repetitions. In addition, 150 repetitions of the Sobol based Monte Carlo method yielded a standard deviation (in average) 58% lower than that of the Halton sequences.

More research is needed to find out the root of the differences between the Sobol and Halton sequences (or other LDS), especially to explain counterintuitive results found when increasing the number of repetitions in Halton sequences from 100 to 150 and Sobol sequences from 50 to 100. In both cases the standard deviation of the estimated parameters increased with no apparent cause.

The role of model specification should also be studied when different variance-covariance combinations are included. The multinomial probit model should also be tested to find a suitable LDS that could substantially reduce the computational effort involved in the estimation process.

Table 1: Percentage of Error w/r to the Base Error for Sample 1

REPETITIONS	HALTON	SOBOL	RANDOM
50	**97.3%**	**100.9%**	**-**
100	**97.0%**	**99.3%**	**-**
150	**98.5%**	**100.5%**	**-**
250	**97.9%**	**100.2%**	**-**
500	**100.1%**	**100.2%**	**100.7%**
750	**-**	**-**	**102.2%**
1,000	**-**	**-**	**100.0%**

Table 2: Averaged Standard Deviations of Estimated Parameters in Sample 1

REPETITIONS	HALTON	SOBOL	RANDOM
50	**0.000113**	**0.000048**	**-**
100	**0.000108**	**0.000158**	**-**
150	**0.000115**	**0.000099**	**-**
250	**0.000107**	**0.000085**	**-**
500	**0.000026**	**0.000025**	**0.000100**
750	**-**	**-**	**0.000159**
1,000	**-**	**-**	**0.000100**

Table 3: Percentage of Error w/r to the Base Error for Sample 2

REPETITIONS	HALTON	SOBOL	RANDOM
50	**99.37%**	**99.76%**	**-**
100	**101.29%**	**99.95%**	**-**
150	**100.01%**	**100.00%**	**-**
250	**99.93%**	**99.85%**	**-**
500	**-**	**-**	**101.15%**
750	**-**	**-**	**102.13%**
1,000	**-**	**-**	**100.00%**

Table 4: Averaged Standard Deviations of Estimated Parameters in Sample 2

REPETITIONS	HALTON	SOBOL	RANDOM
50	**0.00062**	**0.00091**	**-**
100	**0.00861**	**0.00016**	**-**
150	**0.00045**	**0.00019**	**-**
250	**0.00065**	**0.00073**	**-**
500	**-**	**-**	**0.00817**
750	**-**	**-**	**0.01046**
1,000	**-**	**-**	**0.000883**

7. References

BHAT, C. (2001a). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. **Transportation Research, 35B**, 677-693.

BHAT, C. (2001b). Simulation estimation of discrete choice models using randomized and scrambled Halton sequences. Working Paper, Department of Civil Engineering, University of Texas, Austin.

CHENG, J., DRUZDZEL, M.J. (2000). Computational investigation of low-discrepancy sequences in simulation algorithms for Bayesian networks, Working Paper, Decision Systems Laboratory, School of Information Sciences and Intelligent Systems Program, University of Pittsburgh, Pittsburgh.

GALANTI, S., JUNG, A. (1997). Low-discrepancy sequences: Monte Carlo simulation of option prices. **Journal of Derivatives**, **Fall 1997**, 63-83.

GOETT, A., HUDSON, K., TRAIN, K. (2000). Customer's choice among retail energy suppliers: the willingness-to-pay for service attributes. **Energy Journal, 21**, 1-28.

HALTON, J.H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. **Numerische Mathematik, 2**, 84-90.

LEVA, M., and SILVA, M. (2002). Computational Code for the Estimation of Discrete Choice Models through Monte Carlo Simulation and Low Discrepancy Sequences. Working Paper Department of Transport Engineering, Pontificia Universidad Católica de Chile.

MOROKOFF, W.J., and CAFLISCH, R.E. (1994). Quasi-random sequences and their discrepancies. **SIAM Journal on Scientific Computing, 15**, 1251-1279.

MOROKOFF, W.J., and CAFLISCH, R.E. (1995). Quasi-Monte Carlo Integration. **Journal of Computational Physics, 122**, 218-230.

NIEDERREITER, H. (1992). Random Number Generation and Quasi-Monte Carlo Methods. **CBMS-NFS Regional Conference Series in Applied Mathematics, 63,** SIAM, Philadelphia, Pennsylvania.

OHBUCHI, R., and AONO, M. (1994). Quasi-Monte Carlo (QMC) rendering with adaptive sampling. Working Paper, Tokyo Research Laboratory, IBM Corporation.

OWEN, A.B. (1995). Randomly permuted (t,m,s)-nets and (t,s)-sequences. In H. Niederreiter, and J.S. Shiue (Eds.), **Monte Carlo Methods in Scientific Computing,** 299-317, Springer-Verlag, Nueva York.

PASKOV, S.H. (1994). Computing high dimensional integrals with applications to finance. Technical Report CUCS-023-94. Department of Computer Science, Columbia University, New York.

SOBOL, I.M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. **U.S.S.R. Computational Mathematics and Mathematical Physics, 7**, 86-112.

SRINIVASAN, K., and MAHMASSANI, H. (2000). Analyzing heterogeneity and unobserved structural effects in route switching behavior under ATIS: a dynamic kernel logit

(DKL) formulation. Working Paper, Department of Civil Engineering, University of Texas, Austin.

TRAIN, K. (1998). Recreation demand models with taste differences over people. **Land Economics, 74**, 230-239.

TRAIN, K. (1999). Halton sequences for mixed logit. Working Paper Nº E00-278, Department of Economics, University of California, Berkeley.

TUFFIN, B. (1996). Improvement of Halton sequences distribution. Publication Interne N° 998, Institut de Recherche en Informatique et Systèmes Alèatoires, IRISA, France.