FOI Working Paper

Maintaining symmetry of simulated likelihood functions

Laura Mørch Andersen

2010 / 16

Maintaining symmetry of simulated likelihood functions[*]

Laura Mørch Andersen[†]

April 2010

Abstract

This paper suggests solutions to two different types of simulation errors related to Quasi-Monte Carlo integration. Likelihood functions which depend on standard deviations of mixed parameters are symmetric in nature. This paper shows that antithetic draws preserve this symmetry and thereby improves precision substantially. Another source of error is that models testing away mixing dimensions must replicate the relevant dimensions of the quasi-random draws in the simulation of the restricted likelihood. These simulation errors are ignored in the standard estimation procedures used today and this paper shows that the result may be substantial estimation- and inference errors within the span of draws typically applied.

JEL classification: C15; C25

Key words: Quasi-Monte Carlo integration; Antithetic draws; Likelihood Ratio tests; simulated likelihood; panel mixed multinomial logit; Halton draws

---

[†] Contact information: Laura Mørch Andersen, Institute of Food and Resource Economics, University of Copenhagen, www.foi.dk e-mail: LA@foi.dk

1.   Introduction

Models allowing for heterogeneity have been developing rapidly during the last couple of decades thanks to advances in computational speed and understanding of simulation methods for approximating integrals, (see e.g. Ben-Akiva et al., 1993, Ben-Akiva and Bolduc, 1996, Berry et al., 1995, Bhat, 1996 or Brownstone and Train, 1999). One way of introducing heterogeneity is to assume that values of one or more parameters follow a specified distribution, rather than assuming a fixed value. This is known as 'mixing' parameters, and the result of the estimation is the parameters describing the distribution rather than the value of the mixed parameter. Many models are estimated by simulated maximum likelihood, and restrictions easily tested using Likelihood Ratio tests. As an example, the Mixed MultiNomial Logit (MMNL, McFadden and Train, 2000) is becoming an attractive way for researchers to introduce heterogeneity into discrete models.

Calculating the likelihood of a mixed model means that a conventional likelihood must be integrated over all possible values of the mixed parameters. Often, this integral does not have a closed form and the integral is therefore approximated by Monte Carlo integration. This means that random values of the mixed parameter are drawn from the underlying distribution and used to calculate the numerical integral which is then used as an approximation.

Clearly, reliable estimation, validation and inference techniques are a prerequisite for sound models and analysis. The use of simulated likelihoods are bound to induce some approximation error, and it is therefore important to validate the results, e.g. by varying the starting values and checking the stability of the results. One example of a simulation error is investigated in Chiou and Walker (2007), who illustrated that a low number of draws in the simulation of the integral may lead to unidentified estimates. Ben-Akiva and Boulduc (1996), Walker (2001), Hensher and Greene (2003) and Walker et al. (2007) confirm that in general

the validity of the results is greatly influenced by the number of draws, which should therefore also be varied.

When working with real data the estimated log-likelihood values are often highly dependent on the starting values. The phenomenon of local maxima is familiar in complicated models, but in a mixed model the variance in likelihood sometimes cannot be explained in this way. In some cases, the estimated likelihood values given different starting values varies a lot, even when the parameter results are very similar. This means that results close to each other may lead to large differences in likelihood value, a fact which appears to be inconsistent with the continuous nature of the likelihood function, and is not consistent with local maxima. In some cases the problem is small enough to be ignored, but in other cases it rules out usable Likelihood Ratio tests, e.g. when Likelihood Ratio test statistics of restrictions become negative. This has been observed on real data, but is theoretically impossible.

One simulation error which has not previously been investigated is the potential lack of symmetry of the simulated likelihood function. If one of the mixing parameter in describes a standard deviation, the mixed distribution will be symmetric around zero for this parameter. The simulated likelihood must therefore also be symmetric around zero for this parameter. This paper illustrates that if the conventional likelihood is symmetric this will always be the case, but if the conventional likelihood is asymmetric this is not true unless the draws are symmetric around zero. Symmetric draws are usually not used in standard optimization procedures, and the simulated likelihoods are therefore usually not symmetric. As the number of draws increases the degree of symmetry will increase, but as it will also be illustrated in the following, the degree of symmetry is not always sufficient for the number of draws usually applied. One example of an asymmetric conventional likelihood is the likelihood of a logit model, which will be used as an example in this paper.

The purpose of this paper is to investigate the problem of inconsistent likelihood results systematically. Using data simulated under a mixed logit model, we are able to understand how approximation errors cause the observed types of simulated likelihood function instability, its implications for estimation and inference, and to propose solutions that dramatically reduce the approximation error for estimation and inference. We show that the same mechanisms appear in a real data set with invalidating implications for Likelihood Ratio tests.

Erroneous variation in the simulated values of the log-likelihood function may first of all lead to falsely accepted or rejected hypotheses. Secondly it may also falsely indicate that data is not informative enough to support the model, and therefore lead to unnecessary reductions in model complexity. The problem is of course reduced when the number of draws increase, however, this paper shows that the problem remains critical within the span of draws that is feasible today.

This paper examines the particular case of the mixed multinomial logit model, simulated by Quasi-Monte Carlo estimation using Halton draws, and shows that

1) *the problem of inconsistent likelihood values is completely removed when one uses antithetic instead of conventional draws.*

2) *when testing mixing of parameters it is necessary to ensure that the restriction of the mixing distribution is done in the relevant dimension(s), not in arbitrary dimensions.*

The problems and solutions presented in this paper not only apply to mixed logit models, but also to other models estimated by maximum simulated likelihood and the paper therefore provides a valuable contribution to the ongoing struggle to improve the precision of simulation methods. Halton draws are used in this paper, but the properties of the antithetic draws can be generalised to other types of draws.

The structure of the paper is: The above section presented the problems associated with Likelihood Ratio tests performed on simulated log-likelihood values. Section 2 describes the mixed logit model for panels, section 3 explains how the likelihood function of mixed logit can be simulated by Quasi-Monte Carlo integration and section 4 briefly presents the concept of Likelihood Ratio tests. Section 5 investigates the lack of symmetry of the simulated log-likelihood function on artificial data, by comparing results in the true optimum using conventional Halton draws. Section 6 investigates the effect of the lack of symmetry outside the optimum. Section 7 illustrates the problem using real data and section 8 introduces antithetic Halton draws. Section 9 presents the encouraging results of using this type of draws. Section 10 discusses the problems related to restrictions on the number of mixed parameters and section 11 concludes.[1]

2.   Panel Mixed logit

Discrete choice models are based on the assumption that individuals derive different levels of utility from different alternatives. The utility is assumed to depend on the characteristics of the alternative, and perhaps also of the individual making the choice. The utility is assumed to vary randomly from choice to choice. The decision-maker knows the exact utility in each period, but the econometrician can only estimate the non-random part of the utility function. The utility function for individual i at time t can therefore be written as

$$U_{it}\left(j\big|\beta_i, x_{ijt}\right) = x_{ijt}\beta_i + \varepsilon_{ijt} \tag{1}$$

where β_i is an $m \times 1$ vector of parameters giving the utility of the variables in the $1 \times m$ vector of attributes x_{ijt} for individual i. If the error terms are assumed to follow the extreme value distribution, it is possible to estimate the parameters of the utility function using a

---

[1] Appendix A explains how to draw from a distribution, Appendix B presents the artificial data used in this paper and Appendix C illustrates the differences in log-likelihood values in optimum in different quadrants. Appendix D illustrates the degree of symmetry of conventional Haltons by picturing the skewness coefficients for different primes and different numbers of draws.

conventional multinomial logit model (McFadden 1973) which means that the likelihood function is:

$$L_i^{conv}\left(y_i,\beta_i,x_i\right)=p\left(y_i|\beta_i,x_i\right)=\prod_{t=1}^{T_i}p\left(y_{it}|\beta_i,x_{iy_{it}t}\right)=\prod_{t=1}^{T_i}\left(\frac{\exp\left(U_{it}\left(y_{it}|\beta_i,x_{iy_{it}t}\right)\right)}{\sum_{k=1}^{J}\exp\left(U_{it}\left(k|\beta_i,x_{ikt}\right)\right)}\right)\qquad(2)$$

where J is the number of alternatives available, T_i is the number of choices made by individual i, y_i is a $T_i\times1$ vector of choices actually made by individual i, β_i is a $m\times1$ vector containing all of the parameters, x_i is a $T_i\times m\cdot J$ matrix of attributes of the alternatives and $U_{it}\left(k|\beta_i,x_{ikt}\right)$ is the utility for household i from choosing element k from the choice set in period t given the parameters β_i and the attributes x_{ikt}. In a conventional logit it is assumed that all individuals have the same utility function, i.e. that β is the same for all individuals.

In a Mixed MultiNomial Logit (MMNL or MXL) model[2] (McFadden and Train 2000), it is assumed that (part of) the individual utility is drawn from a distribution. This means that the individual utility is known to the individual, but only the distribution is observable to the econometrician. The individual likelihood functions then become the likelihood function of the conventional multinomial logit model integrated over all possible values of β:

$$L\left(\theta\right)=\prod_{i=1}^{I}L_i\left(\theta\right)=\prod_{i=1}^{I}\left(\int p\left(y_i|\beta_i,x_i\right)f\left(\beta_i|\theta\right)d\beta_i\right)\qquad(3)$$

where I is the number of individuals, θ are the parameters determining the distribution of the parameter β, and $f\left(\beta|\theta\right)$ is the density of β given θ. The likelihood function is maximised over θ instead of β.

3. Quasi-Monte Carlo integration

Calculating the likelihood function in (3) is very cumbersome, especially if θ describes a multivariate distribution. The problem can be reduced significantly by using Monte Carlo

---

[2] Also known as Random Parameter Logit (RPL).

integration (see e.g. Morokoff and Caflisch 1995 for asymptotic properties). Monte Carlo integration generally means drawing a set of values of β from the distribution given by θ, calculating the value of the integral for each draw, and taking the mean of these values.[3] Quasi-Monte Carlo integration means that the values of β are drawn quasi randomly from the distribution, instead of randomly. Halton draws are drawn from a Halton sequence and are one out of many types of quasi random draws. Halton sequences were first presented by Hammersley (1960) and Halton (1960). The efficiency of Halton sequences compared to random draws is discussed in detail in both Train (1999) and Baht (2001). Both find that Halton sequences greatly improve accuracy, allowing for far fewer draws and faster computation.

In a panel mixed logit the β's are assumed to be drawn from a common distribution, but to be constant for each individual. The β's therefore vary over individuals, not over observations from the same individual. In a Monte Carlo integration with R draws one must therefore draw R sets of β's for each of the I individuals. For each individual, the value of the likelihood function of the conventional likelihood is then calculated for each of the R randomly drawn β's and the mean is an approximation of the integral in equation (3). This means that the likelihood function for the entire sample is:

$$
\begin{aligned}
L(y,\theta,x) &= \prod_{i=1}^{I} L_i(y_i,\theta,x_i) \\
&= \prod_{i=1}^{I} \left(\int \left(p(y_i|\beta_i,x_i) f(\beta_i|\theta) \right) d\beta_i \right) \qquad (4) \\
&\approx \prod_{i=1}^{I} \left(\frac{1}{R} \sum_{r=1}^{R} \left(p(y_i|\beta_{ir},x_i) \right) \right)
\end{aligned}
$$

where the $R \cdot I$ β_{ir}'s are drawn from the mixing distribution given by θ. According to Hensher and Greene (2003), there is no standard for the number of draws needed, but they find that

---

[3] Appendix A explains how to draw from any given distribution, and illustrates the difference between random and Halton draws.

100 draws appears to be a "good" number. In order to validate the model Hensher and Greene suggest that the models are estimated over a range of draws from 25 to 2,000.

4.　Likelihood Ratio tests on simulated log-likelihood values

In many cases the purpose of estimating a likelihood function is twofold: Maximising the likelihood function leads to the set of parameters which fit the data best, and comparing the best likelihood values of different models makes it possible to determine whether the models are significantly different. The latter is done by Likelihood Ratio (LR) tests based on the difference between the restricted (L_R) and the unrestricted (L_U) likelihood values (see e.g. Greene 1997).

Table 1 shows the critical difference between log-likelihood values, given the degrees of freedom and significance level.

Table 1 Critical differences in log-likelihood values (ln(L_U)-ln(L_R))

	log-likelihoods significantly different at:		
Degrees of freedom	**1% level**	**5% level**	**10% level**
1	3.32	1.92	1.35
2	4.61	3.00	2.30
3	5.67	3.91	3.13
4	6.64	4.74	3.89
5	7.54	5.54	4.62
6	8.41	6.30	5.32

If the difference between the restricted and the unrestricted log-likelihood function is 3.32 – and only one parameter is restricted – the values of the LR test becomes $\chi_1^2(2 \cdot 3.32) = \chi^2(6.63) = 1\%$ which means that the probability that the models have the same explanatory power is one per cent, which usually leads to the conclusion that the two models are significantly different. If the difference is only 1.35 the probability of equal explanatory power is ten per cent, since $\chi_1^2(2 \cdot 1.35) = \chi_1^2(2.71) = 10\%$.

From Table 1 it is clear that even rather small variations in the log-likelihood value can have a significant impact on the results. The absolute level of the log-likelihood function is of no interest, but if the difference in the value of the simulated log-likelihood given different starting values is above e.g. two, testing hypotheses may easily lead to false conclusions. Sometimes the LR value becomes too small, other times too big, and in very unfortunate cases it may have the wrong sign if the log-likelihood value of the restricted model becomes higher than the log-likelihood value of the unrestricted model. The varying values of the log-likelihood function may first of all lead to falsely accepted or rejected hypotheses. Secondly,

it may also falsely indicate that data are not informative enough to support the model, and therefore lead to unnecessary reductions in model complexity.

5.   Symmetry of simulated log-likelihood, artificial data

As mentioned in the introduction, the mixed distribution will be symmetric around zero for parameters describing standard deviations of the mixing distribution. If a logit model is mixed with e.g. the normal distribution, each mixed parameter leads to two mixing parameters, a mean and a standard deviation. Both of these parameters are maximised over the entire real axis $\square$. The distribution actually depends on the mean and the *variance*, and since the standard deviation is the square of the variance there is no mathematical problem in a negative standard deviation. As illustrated in Figure 1, the true mixed likelihood function will be symmetric around zero, when focusing on one dimension relating to a parameter for a standard deviation.
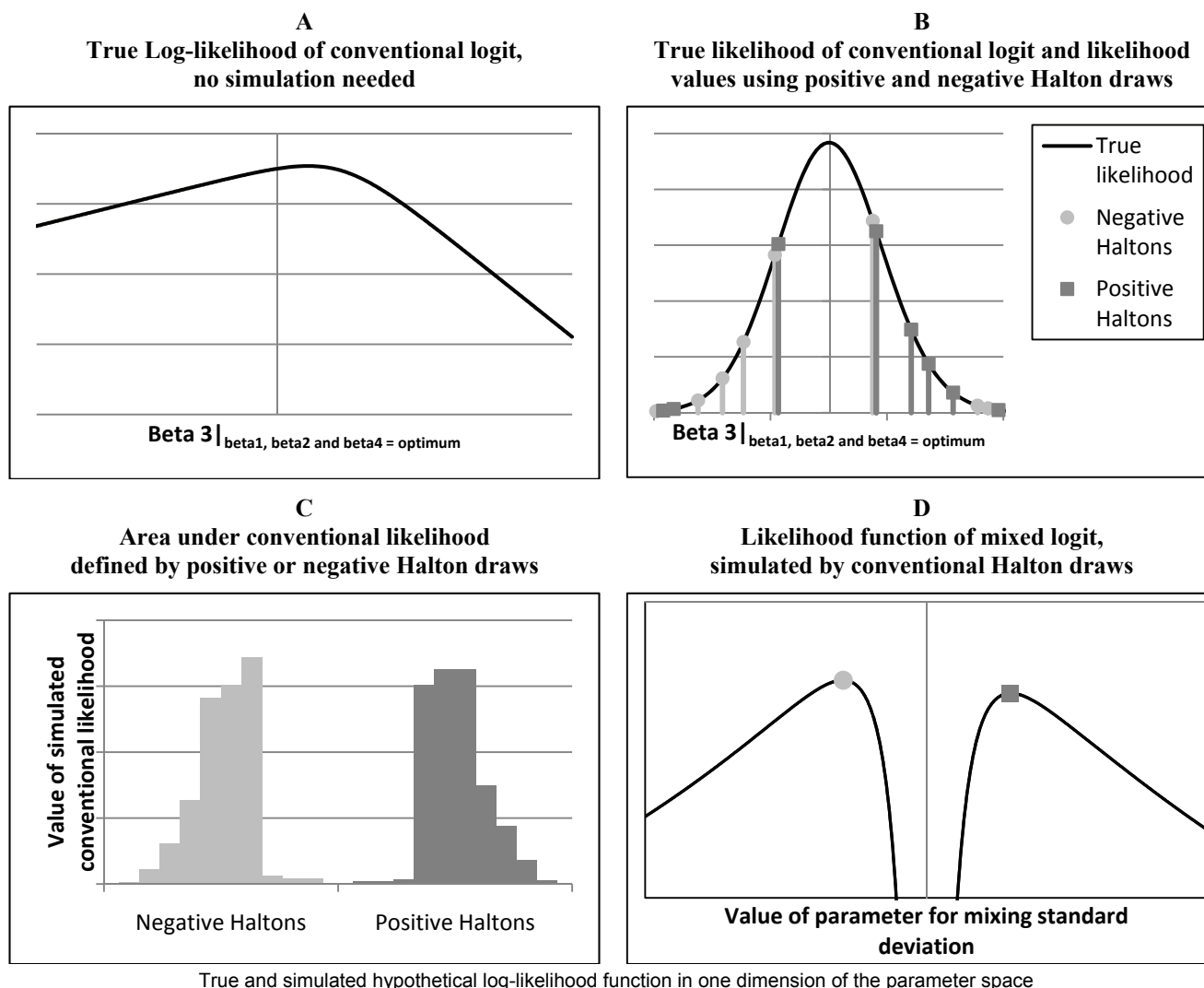
Figure 1 True symmetric mixed likelihood function, one dimension

Value of parameter for mixing standard deviation

Mixed Log-likelihood function in one dimension of parameter space
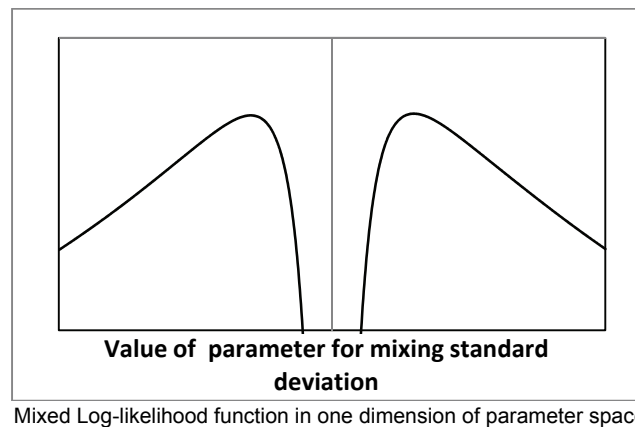
In a mixed multinomial logit model the likelihood function is an integral over the likelihood function of a conventional logit model (see equation (4)). This means that each point on the likelihood function of the mixed logit in Figure 1 corresponds to the area under the likelihood function of a conventional logit. For each value of the mixing standard deviation, the area under the conventional likelihood function is calculated by Quasi-Monte Carlo integration, where the betas are drawn from a distribution defined by the parameters in the mixing

distribution at this specific point. As illustrated in Figure 2 A and B, the likelihood of a conventional logit is not symmetric. The asymmetry is clear in figure 2 A, but since 2 A is the log of 2 B, the likelihood in 2 B is also asymmetric even though it is hard to tell from the picture. The lines in Figure 2 B indicate the values of the conventional likelihood function which enters the Quasi-Monte Carlo integration if one uses conventional Halton draws and a *negative* value of the mixing standard deviation (light grey, labelled 'negative Haltons') and a mirrored version which is the result of the same conventional Halton draws combined with a *positive* value of the mixing standard deviation (dark grey, labelled 'positive Haltons'). Figure 2 C shows the two resulting numerical integrals which are not identical in value because the likelihood of the logit is not symmetric. If the likelihood in 2 B had been symmetric, the numerical integrals in 2 C would have been mirror images of each other, and the area of the numerical integrals would therefore have been identical. When these Quasi-Monte Carlo integrals are calculated for many different values of the mixing standard deviation it is possible to draw the simulated likelihood function of the mixed logit pictured in Figure 2 D, where the values of the individual points calculated in Figure 2 C are also highlighted. The absolute value of the optimum is the same on the negative and the positive axis, but the value of the mixed likelihood differs between the negative and the positive optimum. If the conventional likelihood function had been symmetric, the asymmetry of the conventional Halton draws would not have created problems, but when asymmetric likelihoods are mixed, asymmetric draws poses a problem. As shall be illustrated later, this problem can in some cases invalidate the estimation results.

Figure 2 Lack of symmetry of simulated likelihood function, one dimension

A
True Log-likelihood of conventional logit,
no simulation needed

B
True likelihood of conventional logit and likelihood
values using positive and negative Halton draws

C
Area under conventional likelihood
defined by positive or negative Halton draws

D
Likelihood function of mixed logit,
simulated by conventional Halton draws

True and simulated hypothetical log-likelihood function in one dimension of the parameter space

In order to investigate the magnitude of the problem under controlled conditions, an artificial data set has been created. The data are panel data with 1,000 individuals each choosing 20 times between 4 alternatives. The utility of the alternative specific constant is zero for the alternative which is used as base, the utility of the remaining alternatives follows a three-dimensional normal distribution with no correlation. For more on the definition of the artificial data, se Appendix B. One of the virtues of artificial data is that the true parameters are known. Note that the true parameters are not the parameters used to create the artificial data but the empirical means and standard deviations of the simulated β's (see appendix B for the difference).

In the case of three mixings, the parameters for the standard deviation are estimated in $\square^3$, which means that the number of different quadrants is $2^3 = 8$, and the likelihood function must therefore be symmetric in all eight quadrants. In the rest of this section the likelihood values calculated from the true parameters of the artificial data will be compared between the different quadrants. In these calculations, the probability of finding an optimum in a given quadrant is treated as equal for all quadrants. In actual estimations the probability of ending up in a given quadrant may well vary, and the results in this section therefore only illustrate the *magnitude* of the problems that may *potentially* arise from actual estimations. As will be illustrated in section 7, estimations on actual data lead to results in all eight quadrants, so the problem also exists when parameters are optimised rather than known a priory.

Table 2 compares the results of the eight quadrants for increasing numbers of draws. For 100 draws, the highest difference between the log-likelihood values of different quadrants is 9.26 which is definitely not zero as it should theoretically be. As illustrated in Table 1, differences of this magnitude can ruin Likelihood Ratio tests completely. The difference between the quadrants decreases as the number of draws increases, simply because the distance between draws is reduced, but it does not disappear within a feasible span of draws, and certainly not for the low number of draws recommended in Hensher and Greene (2003).

Table 2 Variation in simulated log-likelihood, by number of draws, artificial data

|	Number of draws per individual					
	100	500	1,000	1,500	5,000	7,500
Highest absolute difference in simulated log-likelih.	9.26	4.44	1.09	0.88	0.47	0.19

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws.

As mentioned above, one of the problems caused by the difference between the values of the log-likelihood function evaluated at different quadrants is that it influences the results of Likelihood Ratio tests. The rest of this section investigates the effect on a Likelihood Ratio test where the restricted model assumes that the mean utility of alternative B is zero, and places no bounds on the standard deviation. The restricted model should be accepted. Note

that the base alternative A has both mean *and* standard deviation equal to zero. Testing away the mean is therefore *not* the same as testing whether the utility of alternative B is the same as the utility of the base alternative A.

The large variation in the value of the log-likelihood function means that the value of the restricted model in one quadrant may be higher than the value of the unrestricted model in another quadrant, but never within quadrants. Figure 3 shows the log-likelihood values for the unrestricted and the restricted model using 100 Halton draws. **I** shows the relationship between the two models in each quadrant and **II** ignores the quadrants and sorts the values by size. Especially from **II** it is evident that the value of the restricted model will sometimes be higher than the value of the unrestricted model, leading to negative values of the LR test statistic.

Figure 3 Differences between likelihood values of unrestricted and restricted model
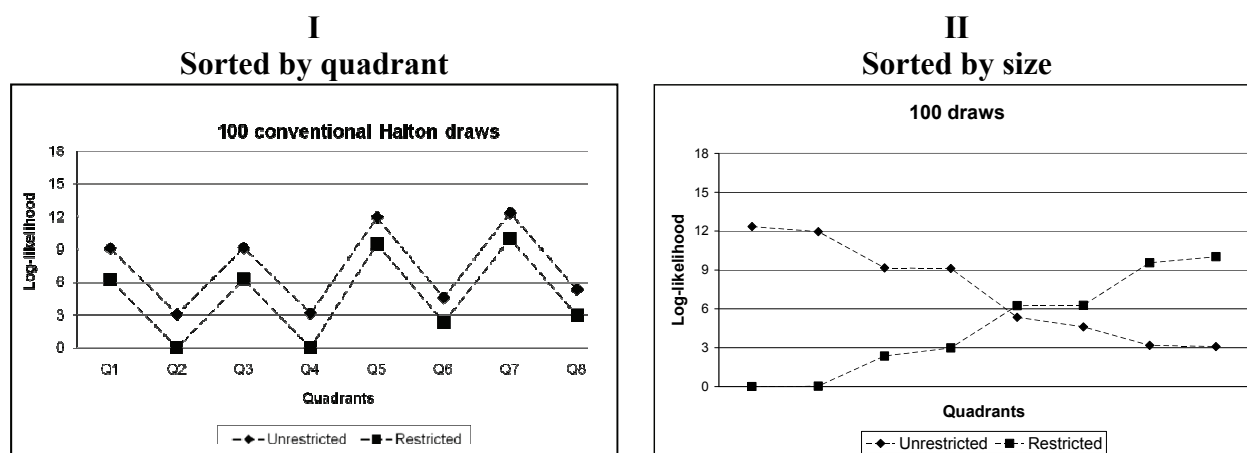
I	**II**
Sorted by quadrant	**Sorted by size**

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. The lowest estimated value (in Q2) is subtracted from all the estimated values. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Table 3 shows that for 100 conventional Halton draws, the LR test statistic will become negative in 20 of the 64 different combinations of restricted and unrestricted log-likelihood values, corresponding to 31 per cent of the cases. The problem decreases with the number of draws, but is still present at 1,000 draws. Appendix C repeats Figure 3 for all the different numbers of draws presented in Table 3.

Table 3 Testing away one mean using conventional Halton draws

	Number of draws per individual:					
	100	**500**	**1,000**	**1,500**	**5,000**	**7,500**
Share of negative LR values	.31	.31	.05	.00	.00	.00
Results of positive LR values						
Lowest test probability	.00	.00	.03	.05	.08	.12
Highest test probability	.64	.66	.77	.65	.27	.20
Standard deviation of test prob.	.13	.16	.15	.13	.05	.02

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Table 3 also summarises the results of the Likelihood Ratio tests that can be performed on the positive LR tests statistics. The test probabilities vary from zero to 64 per cent for 100 draws leading to a standard deviation of 13 per cent. This is of course deeply problematic. The problem decreases as the number of draws increases but even for 5,000 draws the restricted model will sometimes be accepted at the 10 per cent level, and other times rejected. Note that these tests are all performed on the *same* data set. Had the test been performed on different realisations of data with identical parameters the test should have been accepted on 10 per cent of the data sets at the ten percent level, but when the tests are performed at the same dataset the results should all be identical. The differences are caused by simulation errors, not by statistical properties of the test.

Table 4 shows the probability of rejecting the restricted model at different significance levels.

Table 4 Probability of rejecting the restricted model

	Number of draws per individual:					
	100	**500**	**1,000**	**1,500**	**5,000**	**7,500**
At the 1 per cent level	.55	.25	.00	.00	.00	.00
At the 5 per cent level	.89	.45	.07	.05	.00	.00
At the 10 per cent level	.91	.59	.28	.30	.17	.00

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Using 100 draws, 55 per cent of the 64 combinations of unrestricted and restricted log-likelihood values reject the restricted model at the 1 per cent significance level. In 91 per cent of the cases the restricted model is rejected at the 10 per cent level. Using 7,500 draws, the model is never rejected.

6.   Comparing likelihood values outside the optimum

In the discussion of symmetry above, focus has been on differences in optimum, i.e. in the true parameters of the artificial data. During the search for optimum the optimisation routine has to perform outside the optimum, and it is therefore important to know what happens there.

The difference between quadrants is generally higher outside the optimum. This problem decreases to some extent with the number of draws, but as illustrated in Andersen (2008), the difference still does not disappear even with 7,500 draws, and the problem is sometimes smaller for 5,000 draws than for 7,500 draws. This has to do with the symmetry of the Halton draws. As illustrated in Appendix D, the degree of symmetry increases as the number of draws increases, but not monotonically. Increasing the number of draws by a few thousand may therefore lead to set of draws with a lower degree of symmetry, and thereby a bigger difference between the likelihood values in different quadrants.

7.   An example using real data

The problems described above have also been experienced on real data. The example below is based on 10,971 observations from 848 individuals, choosing between four different alternatives. The utility of the non-base alternatives is assumed to follow a tree-dimensional normal distribution with correlation. In this example the true parameter values are not known, and the model is therefore *optimised* using different sets of starting values.

Table 5 shows the optimised values of the Log-Likelihood function in the eight different quadrants, along with the probability of finding a maximum in each quadrant. It is clear that increasing the number of draws reduces the scale of the problem, but it does not solve the problem. The maximal log-likelihood values of 52 different sets of starting values have been sorted into quadrants by the sign of the Choleski parameters.[4] The estimated log-likelihood

---

[4] The Choleski factorisation (Q) is a triangular matrix with the property $QQ' = \Omega$, where Ω is the covariance matrix (Train 2003). The Choleski matrix is therefore the 'square root' of the covariance matrix, and if the covariance matrix is diagonal (i.e. no correlations) the Choleski matrix is merely a diagonal matrix of standard deviations.

values differ significantly between the eight quadrants, but not within quadrants for a given number of draws (not shown). Comparing the quadrants of the starting values and of the optimised results shows that there is apparently no connection between the quadrant of the starting point and quadrant of the final result (not shown).

Table 5 Maximum simulated log-likelihood values by quadrant. Real data, estimated optima

	Number of draws per individual:							
	1,000		1,500		5,000		10,000	
	Prob.	Log-L*	Prob.	Log-L*	Prob.	Log-L*	Prob.	Log-L*
Q1	19%	10.69	23%	7.83	17%	11.88	15%	10.23
Q2	8%	0.00**	12%	8.64	8%	11.98	15%	12.58
Q3	25%	6.64	6%	11.27	17%	11.79	13%	11.45
Q4	19%	8.72	8%	11.32	23%	10.55	21%	11.19
Q5	15%	5.24	19%	11.90	12%	12.80	13%	11.64
Q5	6%	7.05	15%	13.05	6%	12.03	6%	10.99
Q7	4%	4.29	8%	13.76	10%	12.24	13%	12.51
Q8	4%	10.88	10%	13.85	8%	11.37	2%	11.43
Total***	100%		100%		100%		100%	
Min		0.00		7.83		10.55		10.23
Max		10.88		13.85		12.80		12.58
Difference		*10.88*		*6.02*		*2.26*		*2.35*

* The lowest estimated log-likelihood value (-8,388.20, from the estimation with 1,000 draws in Q2) is subtracted from all the estimated values.
** One of the four results in this quadrant differs from the others by 0.274.
*** 52 sets of starting values.
Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants with correlation. 52 sets of starting values. Real data, unbalanced panel, 848 individuals, 10,971 observations. Estimations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws.

It is also clear from Table 5 that the level of the maximum simulated log-likelihood generally is higher for estimations with 5,000 draws than for estimations with 1,500 draws, and that the difference between the quadrants is smaller for 5,000 draws than for 1,500 draws. The problem thus decreases with the number of draws, but even with 10,000 draws (which is in most cases too time consuming) the problem is still there. The problems described above using artificial data therefore also appear on real data. In this example, the variation in likelihood values is sufficiently large to disturb LR tests, even for high numbers of draws. It is important to note that whereas the variation is large (and highly problematic) for the values of the log-likelihood function, the estimated parameters are identical when numerical values of parameters for standard deviations are used in the comparison.

8. Antithetic Haltons

If the model includes more than one mixing parameter, symmetry in one dimension is not enough. If the number of mixed parameters is n – and if perfect symmetry is the goal – for each point in a given quadrant a corresponding point must be present in all of the other 2^n-1 quadrants. Even with a two dimensional mixing based on 2 and 3, this is never the case for conventional Halton draws. The problem is solved by creating antithetic Halton draws. As in Train (2003), the draws are created so that each point is "mirrored" into the 2^n-1 other dimensions.

For a case with three mixed parameters a Halton draw $d_1 = \begin{bmatrix} d_1^1 & d_1^2 & d_1^3 \end{bmatrix}$ (between zero and one) is drawn, and then paired with 7 mirrors in the following way:

$$
\begin{bmatrix}
d_{11} \\
d_{12} \\
d_{13} \\
d_{14} \\
d_{15} \\
d_{16} \\
d_{17} \\
d_{18}
\end{bmatrix}
=
\begin{bmatrix}
d_1^1 & d_1^2 & d_1^3 \\
1-d_1^1 & d_1^2 & d_1^3 \\
d_1^1 & 1-d_1^2 & d_1^3 \\
d_1^1 & d_1^2 & 1-d_1^3 \\
1-d_1^1 & 1-d_1^2 & d_1^3 \\
1-d_1^1 & d_1^2 & 1-d_1^3 \\
d_1^1 & 1-d_1^2 & 1-d_1^3 \\
1-d_1^1 & 1-d_1^2 & 1-d_1^3
\end{bmatrix}
\tag{5}
$$

The Haltons must be symmetric for each individual in the panel, and it is therefore important that each 'set' of symmetric draws is assigned to one individual only, and not distributed over different individuals. The number of draws per individual in a model with n-dimensional mixing must therefore be a multiple of 2^n. In the case of 1,500 draws and three mixings this means that the number of draws must be e.g. $63 \cdot 2^3 = 1,504$ instead of 1,500 to ensure symmetry. Antithetic draws always have perfect symmetry, and therefore always skewness coefficient equal to zero. See Appendix D for more on skewness.

9.   Results of antithetic Halton draws

When the simulated likelihood function for the artificial data is calculated using antithetic Halton draws the difference between the log-likelihood over positive and negative standard deviations is always zero as desired and the Likelihood Ratio test of the restricted model therefore no longer varies. However, the result still changes as the number of draws increases. Table 6 presents the test probabilities already presented in Table 3 above, combined with the results of the antithetic draws. In this example, the variation in likelihood values using conventional Halton draws is sufficiently large to disturb LR tests, even for high numbers of draws.

Table 6 Testing away one mean using conventional or antithetic Halton draws

	Number of draws per individual:					
	100	**500**	**1,000**	**1,500**	**5,000**	**7,500**
Conventional Halton draws *(as in Table 3)*						
Lowest test probability	.00	.00	.03	.05	.08	.12
Highest test probability	.64	.66	.77	.65	.27	.20
Antithetic Halton draws:						
Highest and lowest test probability	.01	.08	.11	.09	.15	.15

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using standard or antithetic Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Table 7 compares the differences in log-likelihood values presented in Table 5 with the results of antithetic draws, and clearly demonstrates the effect of the antithetic Halton draws. The precision of the optimisation is set to 10^{-4}, and the highest difference between two results using antithetic draws is now lower than twice this level, and thereby completely acceptable.[5] Differences of this magnitude will have absolutely no effect on Likelihood Ratio tests, and the antithetic Halton draws therefore solve the problem of instability in the simulated likelihood of the mixed logit. At least the part caused by lack of symmetry of the likelihood function.

---

[5] The precision of the optimisation indicates how close to zero the gradient of the log-likelihood function must be to be perceived as a maximum.

**Table 7 Difference in simulated log-likelihood between quadrants, by number of draws, real data**

	Number of draws per individual			
	1,000	**1,500**	**5,000**	**10,000**
Conventional Haltons (as in Table 5)				
Highest absolute difference in simulated log-likelihood	10.88	6.02	2.26	2.35
Antithetic Haltons				
Highest absolute difference in simulated log-likelihood	0.000196			

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants with correlation. 52 sets of starting values. Real data, unbalanced panel, 848 individuals, 10,971 observations. Estimations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using either conventional Halton draws or antithetic Halton draws.

10. Testing away mixing dimensions

Even when the problem of symmetry is solved, the problem of comparing log-likelihood values of models with different dimensions still remains. In a model with two mixed parameters (β_1 and β_2) the Halton draws will be based on two primes, e.g. 2 and 3 (2 representing β_1 and 3 representing β_2). If one of the mixed parameters (e.g. β_1) is restricted to be fixed (standard deviation restricted to zero), the dimension of the log-likelihood function is decreased by one, and the Halton draws will be based on only one prime. The standard choice would be the first prime, i.e. 2, independent of which dimension is restricted. Figure 4 illustrates the simulated conventional likelihood function which is to be integrated to form the likelihood function of the mixed logit. The heavy black line shows the likelihood function when one of the parameters is restricted to zero, and the dots on this line show the points in which the one-dimensional log-likelihood function would be evaluated for the given grid.

Figure 4 Simulated log-likelihood function in one and two dimensions
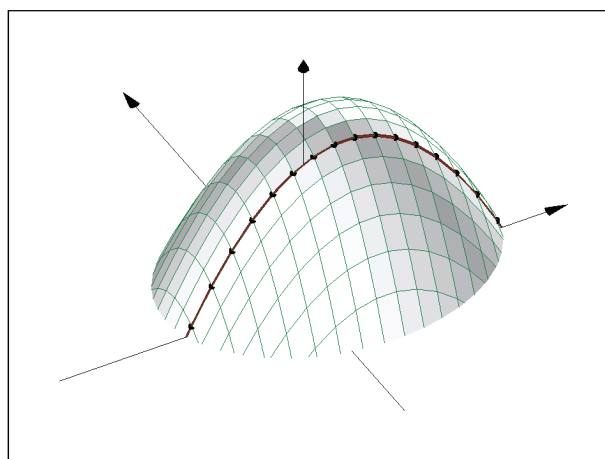
The figure describes a hypothetical log-likelihood function on a two dimensional parameter space. The heavy black line shows the likelihood function when one of the parameters is restricted to zero.

The symmetry of antithetic Haltons is needed to ensure that the log-likelihood functions of the different quadrants are identical, but as illustrated in Figure 5, the choice of prime may also matter. Figure 5 describes the same hypothetical log-likelihood function as Figure 4. The dots show the points in which the one-dimensional log-likelihood function would be evaluated for different draws. The dots in **I** illustrate a case where the one-dimensional draws correspond with the two-dimensional grid, and **II** illustrates a case where the one-dimensional draws are not part of the two-dimensional grid. The area under the one-dimensional likelihood function is clearly not the same in **I** and **II**.

Figure 5 Different one-dimensional likelihood functions given by different draws

I	**II**
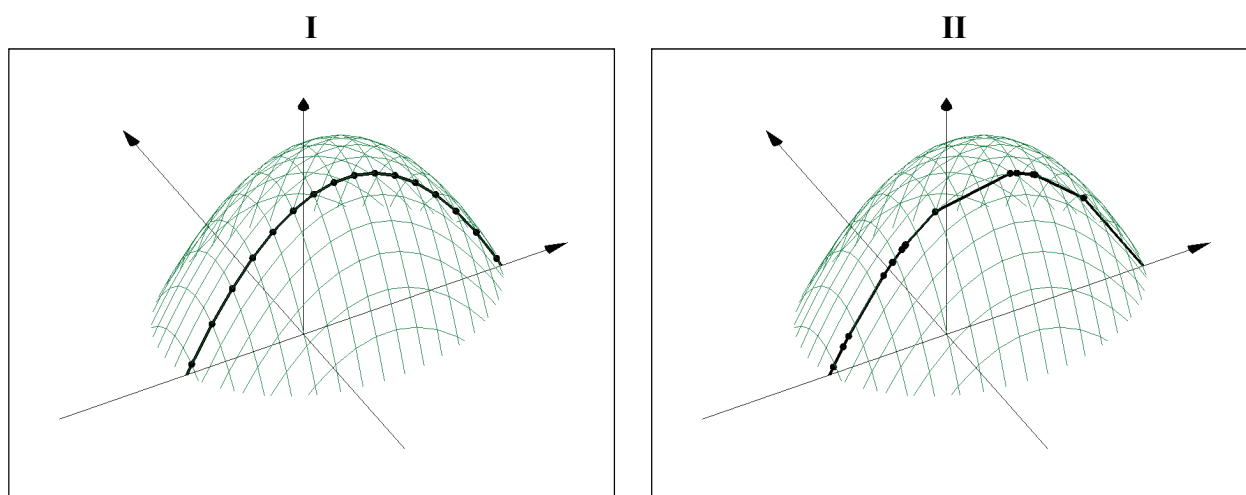
The figure describes the same hypothetical log-likelihood function as in Figure 4. **I** illustrates a case where the one-dimensional draws correspond with the two-dimensional grid, **II** illustrates a case where the one-dimensional draws are not part of the two-dimensional grid.

To investigate the size of the problem we return to the three dimensional mixing on the artificial data used above (1,000 individuals and 20 observations per individual, defined in Appendix B). The restriction is now placed on the utility of alternative C instead of the utility of alternative B, which was restricted in the mean-restriction case above. The utility of alternative C has a mean of 0.9981 and a standard deviation of 0.0984. The restricted model assumes that the standard deviation is zero, but places no bounds on the mean. This means that the restricted model does not assume that the utility of alternative C is the same as the utility of the base alternative A. Table 8 shows the results of evaluating the log-likelihood

function in the true parameters of the restricted model, using different primes for the antithetic Haltons. The difference is substantial.

Table 8 Log-likelihood values in the optimum of the restricted model using antithetic Haltons[6]

	Number of draws per individual:					
	100	**500**	**1,000**	**1,500**	**5,000**	**7,500**
Antith. Haltons based on 2 and 3	-18,466	-18,413	-18,412	-18,410	-18,409	-18,408
Antith. Haltons based on 2 and 5	-18,475	-18,415	-18,410	-18,411	-18,409	-18,408
Difference between results based on different primes	8.37	2.98	-1.64	0.92	-0.70	-0.85

Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the standard deviation of the utility of alternative C is zero.

As illustrated in the previous sections, the use of asymmetric draws can easily lead to wrong conclusions when testing away the mean of a mixed parameter. However, the effect of restricting the standard deviation of a mixed parameter to be zero is very different from the effect of restricting the mean of a mixed parameter to be zero. The true value of the restricted mean in Table 6 is 0.0834 (see Appendix B) and the true value of the restricted standard deviation in Table 8 is 0.0984, i.e. the absolute values of the restricted parameters in the two models are almost identical. Yet the restricted mean-model is accepted as the number of draws increases, but the restricted variance-model leads to differences in the log-likelihood value above 1,000 even for 7,500 draws, so it is unequivocally rejected. The difference between the log-likelihoods based on different primes in the example in Table 8 is therefore of no importance. But for LR tests which are 'close' to being accepted it will be important to keep track of the relationships between primes and mixing dimensions.

11. Conclusion

In mixed models involving standard deviations it is assumed that the signs of parameters for these standard deviations of the mixing distribution have no influence on the value of the likelihood function. When the Monte Carlo integration of the likelihood function is done by

---

[6] Looking at the differences in likelihood function outside the optima leads to results which are very similar to the ones presented in section 1.6. For more details, see Andersen (2008).

conventional Halton draws, this assumption breaks down because the Halton draws differ from quadrant to quadrant. This means that sets of optimal parameters with different signs of the standard deviation can lead to a number of different values of the log-likelihood function, even though the estimated variance is the same. If the solution to an unrestricted and a restricted model is found in different quadrants the Likelihood Ratio test will make no sense. This paper demonstrates that using antithetic Halton draws eliminates this problem. The paper also illustrates that when testing restrictions on the number of mixed parameters, the relationship between primes and mixed parameters must be maintained in the restricted model.

Note, however, that local maxima may still occur if the model cannot be empirically identified by the data. The stability of the simulated log-likelihood should therefore still be investigated by estimating with different sets of starting values.

References:

- Andersen, L. M. (2008): Antithetic Halton draws, in *Information Provision to Consumers as an Instrument of Environmental Regulation*, PhD Series No. 130 -2008, Department of Economics, University of Copenhagen.

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

- Ben-Akiva, M., D. Bolduc and M. Bradley (1993): Estimation of travel choice models with randomly distributed values of time. *Transportation Research Record*, **1413**: 88-97.

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

- Berry, S., J. Levinsohn and A. Pakes (1995): Automobile Prices in Market Equilibrium. *Econometrica*, **63** (4): 841-890.

- Bhat, C. (1996): Accommodating variations in responsiveness to level-of-service measures in travel model choice modelling. *Working paper, Department of Civil Engineering, University of Massachusetts at Amherst.*

- Brownstone, D. and K. Train (1999): Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics*, **89:** 109-129.

- Chiou and Walker (2007): Masking identification of discrete choice models under simulation methods. *Journal of Econometrics*, 141: 683–703.

- Greene, W. H. (1997): *Econometric analysis*. Third edition. Prentice Hall, Upper Saddle River, New Jersey, USA.

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

- Hammersley, J. M. (1960): Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of Sciences*, **86**: 844-874.

- Hensher, D. A. and W. H. Greene (2003): The Mixed Logit model: The state of practice, *Transportation,* **30**: 133–176.

- McFadden, D. (1973): Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (ed.): *Frontiers in econometrics*. Academic Press, New York. USA, 105-142.

- McFadden, D. and K. Train (2000): Mixed MNL models for discrete response. *Journal of Applied Econometrics*, **15**: 447-470.

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

- Train, K. (1999): Halton sequences for mixed logit. *Working paper*, Department of Economics, University of California, Berkeley.

- Train, K. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press.

- Walker (2001): *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables*. Ph.D. dissertation, Massachusetts Institute of Technology.

- Walker, J. L., M. Ben-Akiva and D. Bolduc (2007): Identification of Parameters in Normal Error Component Logit-Mixture (NECLM) Models. *Journal of Applied Econometrics*, **22**: 1095-125.

Appendix A: Drawing from a distribution

Draws from any given distribution can be created by drawing from the uniform distribution (greater than zero and smaller than one) and taking the inverse of the desired cumulative distribution of these draws (Train 2003). The results will follow the desired distribution. The draws from the uniform distribution can be random draws, Halton draws or other types of quasi-random draws. The efficiency of Halton draws is discussed in detail in both Train (1999) and Baht (2001). Both find that Halton draws greatly improve accuracy with far fewer draws and faster computation. Halton draws are used in this paper, but the properties of the antithetic draws can be generalised to other types of draws.

Figure 6 illustrates how normally distributed draws can be created from random or Halton draws from the uniform distribution.

Figure 6 Drawing from a normal distribution, random and Halton draws
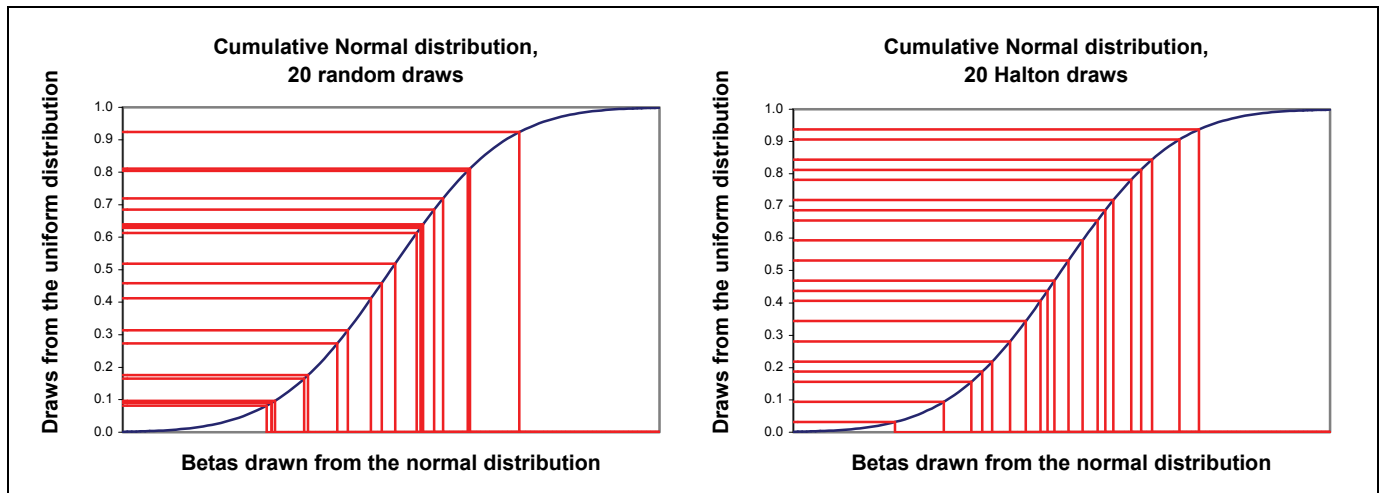
This principle is also formulated in equation (6):

$$
\left.
\begin{array}{l}
\gamma_r \sim uniform \,]0,1[\\
\beta_r = F^{-1}(\gamma_r) \\
F = cdf \ \text{of mixing distribution} f
\end{array}
\right\}
\quad \Rightarrow \quad \beta_r \sim f
\tag{6}
$$

Appendix B: Artificial data

The number of individuals is 1,000 and each individual makes 20 choices. The number of alternatives is 4.

The artificial utility of the four alternatives is defined as:

$$
mean\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.1 \\ 1 \\ 2 \end{bmatrix}, \quad std.dev\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0.1 \\ 3 \end{bmatrix}, \quad corr\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix}
\tag{7}
$$

which means that the artificial covariance is defined as:

$$
cov\begin{bmatrix} u_A \\ u_B \\ u_C \\ u_D \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 0.01 & 0 \\ & & & 9 \end{bmatrix}
\tag{8}
$$

The realisation of the utility is

$$
mean\begin{bmatrix} u_A \\ u_B \\ u_C \\ u_D \end{bmatrix} = \begin{bmatrix} 0 \\ 0.0834 \\ 0.9981 \\ 2.0682 \end{bmatrix}, \quad std\begin{bmatrix} u_A \\ u_B \\ u_C \\ u_D \end{bmatrix} = \begin{bmatrix} 0 \\ 0.9934 \\ 0.0984 \\ 3.0343 \end{bmatrix},
$$

$$
cov\begin{bmatrix} u_A \\ u_B \\ u_C \\ u_D \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ & 0.9868 & 0.0021 & -0.0633 \\ & & 0.0097 & -0.0004 \\ & & & 9.2070 \end{bmatrix}
\tag{9}
$$

Alternative number 1 is used as the base. The mean and the standard error of the base alternative is zero and since differences of normally distributed parameters are also normally distributed, the result is:

$$
mean\begin{bmatrix} u_B - u_A \\ u_C - u_A \\ u_D - u_A \end{bmatrix} = \begin{bmatrix} 0.0834 \\ 0.9981 \\ 2.0682 \end{bmatrix}, \quad cov\begin{bmatrix} u_B - u_A \\ u_C - u_A \\ u_D - u_A \end{bmatrix} = \begin{bmatrix} 0.9868 & 0.0021 & -0.0633 \\ & 0.0097 & -0.0004 \\ & & 9.2070 \end{bmatrix}
\tag{10}
$$

The values in (10) are used as the *true* values in the paper.

Appendix C: Illustration of differences in log-likelihood values in optimum from different quadrants

Table 9 Log-likelihood function evaluated in the true parameters, by quadrant
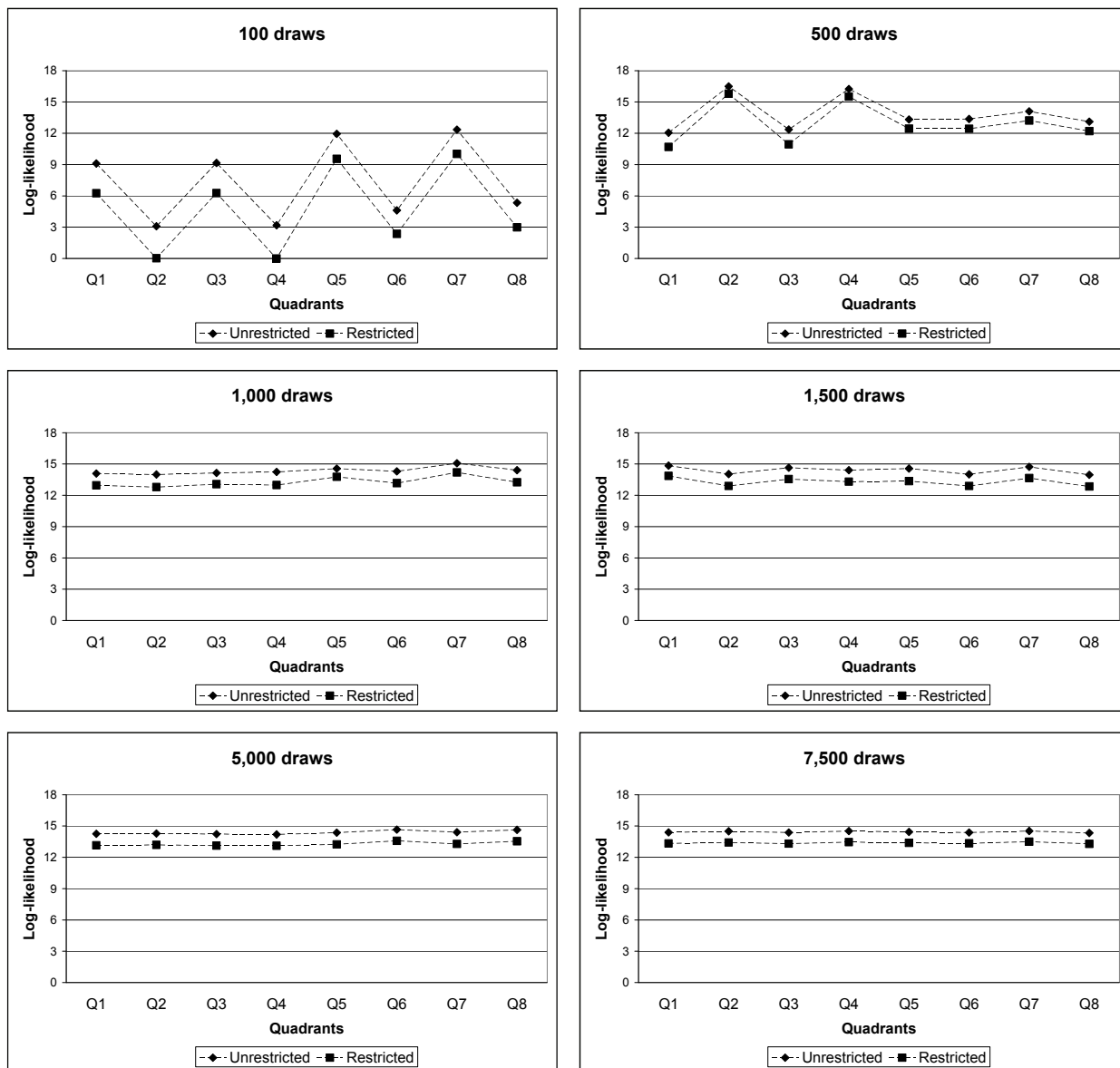
Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. The lowest estimated value (in Q2) is subtracted from all the estimated values. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Table 10 Log-likelihood function evaluated in the true parameters, by size
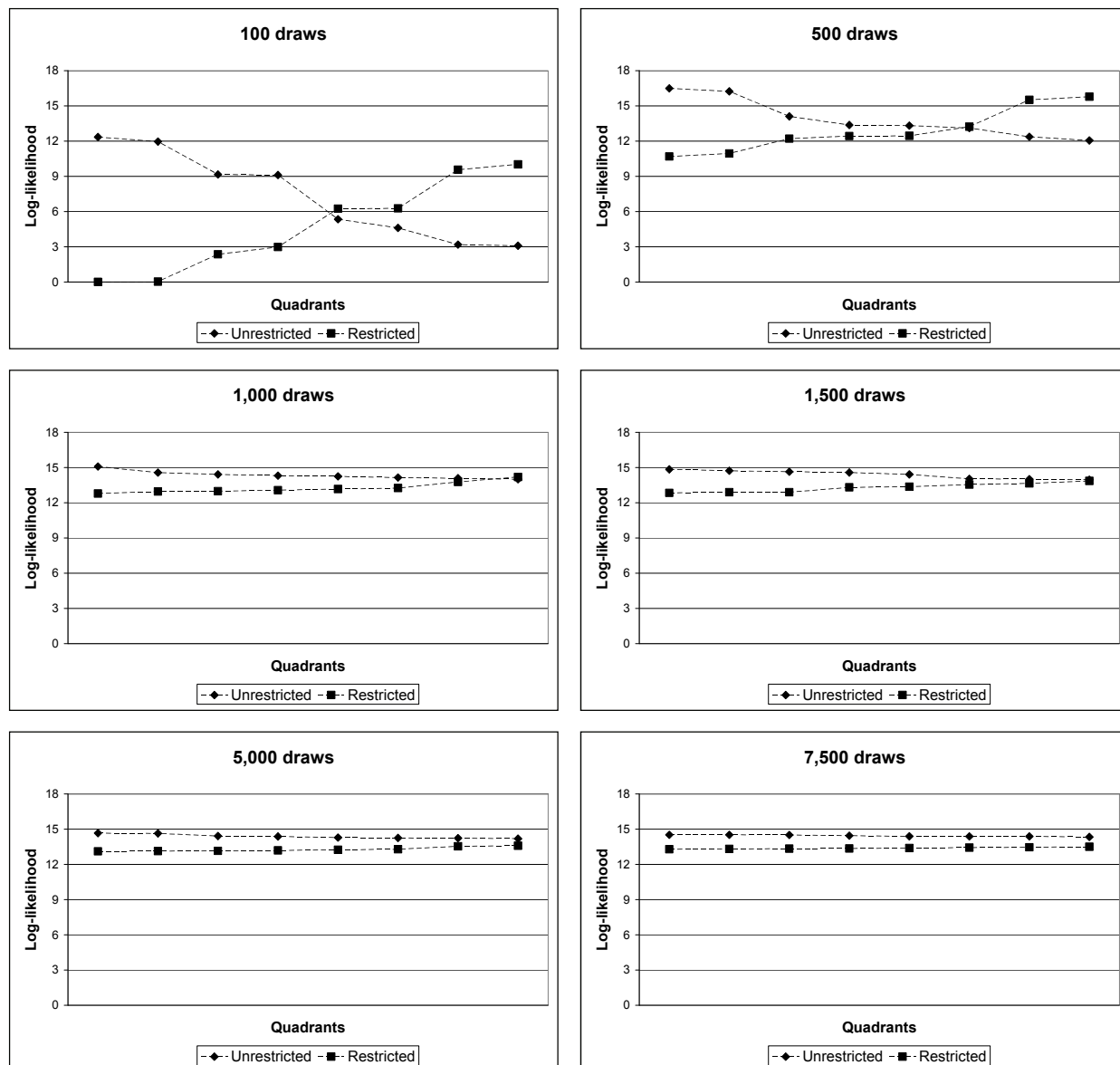
Simulated log-likelihood in a model with 4 alternatives and 3 mixed alternative specific constants. The lowest estimated value (in Q2) is subtracted from all the estimated values. Artificial data, 1,000 individuals, 20 observations per individual. Data is defined in Appendix B. Calculations conducted in the MMNL GAUSS programme developed by Train, Revelt and Ruud, using conventional Halton draws. The restricted model assumes that the mean utility of alternative B is zero.

Appendix D: Skewness coefficient of random and Halton draws

The skewness coefficient is a measure of symmetry and can be calculated for any distribution. As in Greene (1997), the skewness coefficient is calculated as:

$$skewness\ coefficient = \frac{E\left(\left(x-E(x)\right)^3\right)}{\left(E\left(\left(x-E(x)\right)^2\right)\right)^{3/2}} \tag{11}$$

The skewness coefficient is only measured in one dimension at a time, which means that the symmetry of the multivariate distribution may be far smaller than for each of the single dimensions.

Figure 7 shows the skewness coefficients of random draws and Haltons based on the primes 2, 3 and 5. First of all, it is clear that Halton draws are far more symmetric than random draws. Secondly, it is clear that the symmetry of the Halton draws increases with the number of draws, but that it keeps fluctuating, even for very high numbers of draws. Increasing the number of draws by a few thousand may therefore lead to set of draws with a lower degree of symmetry.

The antithethic Haltons presented in this paper always have skewness coefficient zero.

Figure 7 Skewness coefficients of random draws and Halton draws based on the primes 2, 3 and 5