**EUROPEAN REGIONAL SCIENCE CONFERENCE**
**Jyväskylä, August 2003**

# An alternative method to the scrambled Halton sequence for removing correlation between standard Halton sequences in high dimensions

Stephane Hess[1] and John Polak[2]
Centre for Transport Studies
Imperial College London
Exhibition Road
London SW7 2AZ

[1] **stephane.hess@imperial.ac.uk**, Tel: +44(0)20-7594-6105, Fax: +44(0)20-7594-6102
[2] **j.polak@imperial.ac.uk**, Tel: +44(0)20-7594-6089, Fax: +44(0)20-7594-6102

---

## ABSTRACT

Halton sequences have become a frequently used alternative to pseudo-random numbers in Monte-Carlo integration and other simulation methods. While the performance of standard Halton sequences is very good in low dimensions, problems with correlation have been observed between sequences generated from higher primes. This can cause serious problems in the estimation of models with high-dimensional integrals. The most widely used solution to this problem is the scrambled Halton sequence, which uses special predetermined permutations of the coefficients used in the construction of the standard sequence. In this paper, we conduct a detailed analysis of the improvements that scrambled Halton sequences offer over standard Halton sequences for high primes. The analysis shows that although the scrambled sequences manage to reduce correlation, for some choices of primes, correlation remains at an unacceptably high level. We compare the performance of the scrambled sequences to that of shuffled Halton sequences, which use randomly shuffled versions of the one-dimensional standard Halton sequences in the construction of multi-dimensional sequences. We show that the shuffled sequences offer significant computational savings and have the potential for providing better coverage than the corresponding scrambled sequences. Unlike the scrambling approach, the shuffling algorithm can, without any changes in the approach, be used for any choice of prime numbers, and hence for number of dimensions.

In view of these advantages, this paper recommends that this new method should be preferred to the scrambling approach when dealing with high correlation between standard Halton sequences.

## 1. INTRODUCTION

With the emergence of ever more powerful mathematical modelling methods, there has arisen a need for advanced simulation methods to estimate the resulting models. Indeed, many of these models are based on the use of multi-dimensional integrals that often do not have a closed-form expression. One example is the use of models for predicting choices of decision-makers; and notably discrete choice models, whose use in diverse areas of research (and many areas of regional science, e.g. models of aspects of spatial choice, such as route or location) has increased steadily over the past few years. A comprehensive review of existing discrete choice models with an emphasis on simulation-based estimation is given by Train (2003).

The simulation methods used to estimate these models were originally based on the use of pseudo-random numbers, until it was noted that important gains in efficiency could be made through using "cleverly crafted" quasi-random number sequences. The aim of using such sequences is to provide better coverage of the area of integration, where good quality of coverage is defined as low discrepancy between the observed distribution of the draws and a multi-dimensional uniform distribution. The main improvements in performance are generally in the number of draws needed to achieve a certain level of accuracy in the simulation process. As an example, in a simulated likelihood maximization process, the aim is to retrieve the true parameters; as this will in general not be possible, the aim becomes to generate parameters that are sufficiently close to the true parameters. The nature of quasi-random number sequences implies that the number of such draws needed to achieve a certain degree of accuracy will in general be inferior to the respective number of pseudo-random draws required.

## 2. THE STANDARD HALTON SEQUENCE

In recent years, the most prominent type of quasi-random number sequence has been the Halton sequence, introduced by Halton (1960).

The one-dimensional Halton sequence takes a prime number $r$, where $r$ ($\geq 2$), and fills the 0-1 space evenly by using cycles of length $r$. There are different ways of quantifying this approach, the most straightforward mathematical implementation is possibly the one given by Bhat (2002).

To generate a sequence of $N$ draws, the $N$ positive integers are rewritten in base $r$, such that integer $g$ is expressed as:

$$g = \sum_{l=0}^{L} b_l(g) r^l \qquad \text{...(1)}$$

with $0 \leq b_l(g) \leq r-1$ and $r^L \leq g < r^{L+1}$. The integer $g$ can now be expressed in digitised form as the r-adic integer string $b_L(g) b_{L-1}(g) \ldots b_1(g) b_0(g)$.

The g$^{\text{th}}$ Halton element is given by the radical inverse of $g$ to the base $r$ by reflection through the radical point. In base r, we get:

$$\varphi_r(g) = 0.b_0(g) b_1(g) \ldots b_L(g) \qquad \text{...(2)}$$

or, as a decimal number:

$$\varphi_r(g) = \sum_{l=0}^{L} b_l(g) r^{-l-1} \qquad \text{...(3)}$$

where $\varphi_r(g)$ is the $g^{\text{th}}$ term in the Halton sequence generated from prime $r$.

Different sequences, constructed from different primes, are used for different dimensions of a multi-dimensional integral. The multidimensional sequence is constructed by combining the one-dimensional sequences. The $g^{\text{th}}$ point of the s-dimensional sequence is:

$$\varphi^{(s)}(g) = \left(\varphi_{r_1}(g), \varphi_{r_2}(g), \ldots, \varphi_{r_s}(g)\right) \qquad \text{...(4)}$$

The standard Halton sequence of length N on the s-dimensional unit cube is then given by:

$$\Psi^{(s,N)} = \left[\varphi^{(s)}(1)', \varphi^{(s)}(2)', \ldots, \varphi^{(s)}(N)'\right]' \qquad \text{...(5)}$$

Transformation from the s-dimensional unit cube (which is the area of the s-dimensional standard uniform distribution) to other domains of integration is straightforward, through the use of the integral transform result. One example is the use of multivariate standard normally distributed points, which are obtained through:

$$X^{(s,N)} = \Phi^{-1}\left(\Psi^{(s,N)}\right) \qquad \text{...(6)}$$

with $\Phi$ being the standard cumulative normal distribution function.

While multi-dimensional Halton sequences do in general provide better coverage than the corresponding pseudo-random number sequences, problems with high correlation can occur between sequences constructed from higher primes, thus sequences used in higher dimensions. This correlation leads to poor multi-dimensional coverage (even though one-dimensional coverage will still be very good due to the use of Halton sequences) and can thus cause serious problems in the estimation of models with high-dimensional integrals.

These problems are illustrated in figure 1, using sequences of 100 draws. We see that in lower dimensions (using primes 2 and 3) coverage is clearly better than in the corresponding pseudo-random number sequences. Multi-dimensional Halton sequences generated from higher primes however show heavy correlation between the single dimensions, clearly leading to poorer coverage than in the corresponding pseudo-random number sequences (figure 1 shows the combination of primes 29 and 31, and primes 43 and 47).

Although the exact differences between the single sequences can only be analysed once we have introduced a measure for multi-dimensional coverage, the calculation of the correlation between the individual members of the two-dimensional sequences can be used as an initial appraisal of the problem.

The correlation between the Halton sequences for primes 2 and 3 was found to be –0.03, the corresponding correlation measures between the sequences for primes 29 and 31 and between the sequences for primes 43 and 47 were found to be 0.4045 and 0.4363 respectively. The sign of the correlation is in this case of no importance, and the example clearly shows that the correlation between sequences generated from higher primes is significantly higher than that between sequences from lower primes. To compare this to the correlation between pseudo-random number sequences, 500 pairs of pseudo-random number sequences were generated, and the correlation was calculated for each such pair. The absolute values of the results were averaged over the 500 runs, giving a mean absolute correlation of 0.0808 (variance of 0.0036). This reflects the results from figure 1, showing higher correlation (and hence poorer coverage) for the Halton sequences generated from higher primes when compared to pseudo-random number sequences. The mean absolute correlation of 0.0808 is also higher than that between the sequences for primes 2 and 3, again reflecting the results from figure 1.

Such problems with correlation and poor coverage in high dimensions have been well documented in the existing literature, but little effort has been made at describing the reason of the problem. The reason why sequences generated from high primes are especially prone to the problem is however very easy to see. With high primes, the cycles are so long that the different sequences use only a very limited number of cycles in a sequence of say 100 draws. This, on its own, does not lead to correlation. The correlation is caused by the fact that, with high primes, the behaviour of the different sequences becomes very similar, especially with neighbouring primes.

Essentially, problems with correlation and hence discrepancy occur when the ratio of the two prime numbers used is close to an integer value (and especially 1), when the length of cycles used are either very similar (when ratio close to 1) or almost exact multiples of each other (when close to some other integer). This directly leads to correlation as the good coverage quality means that cycles of similar length will use very similar draws. These results are further illustrated in figure 2, showing all possible combinations of Halton sequences generated from prime 19 to 47. These plots also reveal that the behaviour of the Halton sequences can already be quite unsatisfactory even in low dimensions (again, mainly when the ratio of the primes used is close to an integer value). For these choices of primes, the correlation between the sequences may be at an acceptable level, as may the overall coverage, but problems with local area coverage can be quite substantial, as shown in figure 3.

## 3. THE SCRAMBLED HALTON SEQUENCE

Scrambled Halton sequences are constructed by using permutations of the coefficients $b_l(g)$ in the radical inverse function (equations (2) and (3)), such that the scrambled Halton sequence for prime *r* is given by:

$$\varphi_{sr}(g) = \sum_{l=0}^{L} \sigma_r(b_l(g)) r^{-l-1} \qquad \ldots(7)$$

where $\sigma_r$ is the operator of permutations for the possible values of $b_l(g)$ in base *r*.

Different methods for producing the permutations have been proposed; the most popular method is the one given by Braaten and Weller (1979) (hereafter referred to as B&W). The exact approach used by this method has been poorly documented in most of the relevant literature, we will thus give a brief overview of how the permutations are selected.

For prime *r*, the possible digits are $\{0,\ldots,r-1\}$. B&W start by setting $\sigma_r(0)=0$, $\sigma_r(j),\ j=1,\ldots,r-1$ are then set iteratively, choosing first $\sigma_r(1)$, then $\sigma_r(2)$, …, and assigning the last available digit to $\sigma_r(r-1)$. $\sigma_r(j)$ is chosen such as to minimize the one-dimensional discrepancy in the set $\left\{\frac{\sigma_r(1)}{r},\ldots,\frac{\sigma_r(j)}{r}\right\}$, where the division by *r* leads to the set being contained in the one-dimensional 0-1 space.

Tuffin (1997) gives the formula to be used for this one-dimensional discrepancy as:

$$T_N^{(2)^*} = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\left(1 - M^{(n,m)}\right) - \frac{1}{N}\sum_{n=1}^{N}\left(1 - \left(X^{(n)}\right)^2\right) + \frac{1}{3} \qquad \ldots(8)$$

where $N$ is the number of elements in the sequence, $X^{(n)}$ is the $n^{\text{th}}$ element, and $M^{(n,m)} = \max\left(X^{(n)}, X^{(m)}\right)$.

The choice set for $\sigma_r(1)$ contains the digits $\{1,\ldots,r-1\}$, the choice set for $\sigma_r(2)$ is $\{1,\ldots,r-1\}\setminus\sigma_r(1)$, the choice set for $\sigma_r(3)$ is $\{1,\ldots,r-1\}\setminus\{\sigma_r(1),\sigma_r(2)\}$ and the choice set for $\sigma_r(j)$ is $\{1,\ldots,r-1\}\setminus\{\sigma_r(1),\ldots,\sigma_r(j-1)\}$.

B&W (1979) only provide permutations for the first sixteen primes, this is an illustration of the computational cost of generating the permutations. Even without the extra complexity of having to use some kind of decision-rule in the case of ties between the discrepancies in $\left\{\frac{\sigma_r(1)}{r},\ldots,\frac{\sigma_r(j)}{r}\right\}$ for two choices of $\sigma_r(j)$ the number of calculations needed is very substantial. Hess and Polak (2003) found that, to evaluate all the possible discrepancies needed in the generation of the permutations for prime permutations for prime *r,* $\sum_{K=1}^{K=r-2} K\left[2(r-K)^2 + 2(r-K) + 4\right]$ calculations were required. For high *r*, this number becomes very substantial indeed. Although this might still be a small task for modern computers, the number of ties can be expected to increase with the length of the sequences, and hence with *r*, leading to potentially severe problems with the approach. Indeed, it is not known what effect a certain approach for solving ties will have on the final discrepancy (and hence also on the performance of the resulting sequence).

It should also be noted that due to use of permutations of the $\sigma_r(j)$, the scrambling of the one-dimensional sequences will generally result in changes to the quality of the one-dimensional coverage of the sequences; the points used in the scrambled sequences will generally be different from those used in the original sequences. Our analysis showed that depending on *r* and *N,* the scrambling can result in increases or decreases of the quality of coverage; this could be expected to have an effect on estimation performance.

As mentioned above, the number of dimensions for which B&W have produced permutations is sixteen, most existing computer code is thus also limited to this number of dimensions. Although it would be possible to generate permutations for higher primes, little effort has gone into this process so far, as there has so far been little need for such sequences. However, as the dimensionality of integrals that we use keeps on increasing, the scrambling approach is potentially doomed as it is limited in its scope by the availability of permutations (which are generally included in an input matrix), and as it cannot easily be generalised to any number of dimensions.

In an estimation process it is generally not desirable to use the same draws in different experiments; as an example, we would, in the estimation process of a discrete choice model use different draws for different decision-makers. One problem with scrambled Halton sequences is that it is generally not possible to adapt the permutations used; the draws produced in different runs will thus be exactly the same. Two main approaches have been used to deal with this problem, the first approach uses cycling of primes (association of different primes with different dimensions in different experiments), the second approach uses incremental parts of the sequence in different experiments. The first approach is limited to the case where the number of experiments is smaller than *K!*, where *K* is the number of dimensions used. Another problem is that if all permutations of the dimensions are considered, some decision-makers will have some dimensions associated with the same primes as other decision-makers, potentially leading to problems in estimation. The second approach has no such limitation, but involves the added computational cost of having to generate longer sequences; this can impose a very significant burden when using scrambled sequences. The main disadvantage however could be that the different segments of the sequence can have (significantly) different quality of coverage, as has been observed for example by Bhat (1999) and Train (1999).

There has been a distinct lack of analysis of the performance of the scrambled sequences in terms of coverage and correlation, most authors seem to take the results provided by B&W (1979) for granted. The problem is that B&W used the same measure of discrepancy for the analysis of the scrambled sequence as was used in the selection of permutations, namely the multi-dimensional version of equation (8). Although the two problems (discrepancy in the sequence of permutations and in the actual Halton sequence) can be regarded as being quite separate, the same method should never be used for estimation and validation. Our analysis showed that the discrepancy formula given in equation (8) produces, for certain choices of primes, very counterintuitive results. This led us to conduct a graphical analysis of the coverage provided by the scrambled Halton sequences. Figure 4 clearly shows that although the scrambling seems to improve the performance of the sequences, for certain choices of primes, heavy correlation and poor coverage remain in the scrambled sequences. It seems that the main problem is that for some choices of primes, the scrambling leads to grouping around one of the diagonals, due to high remaining correlation between the sequences, and subsequent low coverage in some areas lying off the diagonal line.

## 4. THE SHUFFLING ALGORITHM

The above discussion has revealed three main weaknesses of the scrambling approach. The performance of the scrambled sequence is very unstable, while it is very good for some choices of primes, there are sequences where the scrambling offers insufficient improvement over the original sequences and where a pseudo-random number sequence could be guaranteed to have better performance. Another weakness is its inability to produce different draws in different runs, such that the different sequences would still exhibit the same quality of one-dimensional coverage. The third weakness is the problem of generalisation, without a high number of prior computations, it is not possible to use the scrambled Halton sequences for any very high number of dimensions (>16).

The aim of this analysis was thus to develop a new method which would improve on at least two of these weaknesses, the performance in terms of correlation and the problems with generalisation to higher dimensions.

It is virtually impossible to produce different sequences that have exactly the same multi-dimensional coverage, it is however possible to produce multi-dimensional sequences that have the same quality of one-dimensional coverage along the same dimensions. This can be achieved by applying a method that uses a different permutation of the same initial Halton sequence (rather than a permutation of the digits used in the radical inverse expression) in different experiments. This will not affect the quality of one-dimensional coverage (as the order of the draws is of no importance), but will lead to different multidimensional draws through the use of different permutations of the single-dimensional sequences (the multi-dimensional sequences are produced through association of the single-dimensional ones).

We propose the use of *shuffled Halton sequences*, generated through association of randomly shuffled standard one-dimensional Halton sequences. This shuffling algorithm takes a standard Halton sequence of length $R$ and simply randomizes the order of the elements in the sequence (by using a procedure based on pseudo-random numbers). Even with relatively low *R*, the number of different possible permutations R! is so high that the probability of applying the same random permutation to both sequences should be sufficiently close to zero to guarantee that the different one-dimensional sequences are shuffled in a different way. The risk of this leading to poorer coverage should also be acceptably low, given the very poor initial quality of coverage.

When used repeatedly on multi-dimensional sequences (through use on their one-dimensional components) the method will produce different multi-dimensional sequences, where the quality of one-dimensional coverage along the different dimensions will be identical for the different sequences. Although the multi-dimensional quality of coverage will differ between sequences, the differences should be less important than when using a method that also changes the one-dimensional coverage; indeed, as has repeatedly been observed in the existing literature, these two measures are closely related. It should also be clear that a combination of sequences with poor one-dimensional coverage can never lead to good multi-dimensional coverage.

Another significant advantage of the shuffled sequences is its computational efficiency. As differently shuffled versions of the same one-dimensional sequences will be used in each experiment, there is no need for longer sequences (as is for example the case when using incremental parts of the scrambled sequences). The multi-dimensional draws will be different across experiments, this should help to prevent problems with correlation in simulation errors, which would occur if the same draws were used repeatedly.

Unlike the computer code for generation of scrambled Halton sequences, the code for the shuffling algorithm is very straightforward and can be used for any number of dimensions; there is no need for prior computation of permutations or for changes in the code, no matter how many dimensions are needed. Also, unlike for the scrambling approach, the method used is the same for any prime used, there are no increases in complexity with higher primes.

As a graphical illustration of the coverage provided by the shuffled Halton sequences, plots of the shuffled sequences corresponding to the scrambled sequences shown in figure 4 are shown in figure 5. The shuffling produces different results in different runs; the plots in figure 5 can thus only serve as an illustration. However, the exercise was repeated many times, and the plot repeatedly showed significantly better coverage in the shuffled sequences. These results are reflected on the plots in figure 5, indeed they show none of the major blank areas which are characteristic of the poor quality of coverage shown on the corresponding plots in figure 4.

## 5. QUALITY OF MULTI-DIMENSIONAL COVERAGE

Dobkin et al (1993) propose to measure the discrepancy between the distribution of draws (in the two-dimensional 0-1 space) and a uniform distribution through the use of a large number of differently-sized axis-aligned rectangles.

We let $\mathrm{X}$ be the pattern of *n* points in the unit cube. The method than uses every possible combination of (*x,y*) coordinates and calculates the discrepancy between the proportion of space (out of the area of the rectangle) and the proportion of points (out of the total number of points in the sample) for the rectangle defined by $[0,x]\times[0,y]$. The area of the rectangle is given by *xy*, the discrepancy for the rectangle is measured by:

$$d_{\mathrm{X}}(x,y)=\left|\frac{xy}{A}-\frac{\#([0,x]\times[0,y])}{n}\right| \qquad \ldots(9)$$

where $\#([0,x]\times[0,y])$ gives the number of draws in the area $[0,x]\times[0,y]$, and where in examples using the 0-1 space, *A* will be equal to 1.

The use of all possible combinations of (*x,y*) leads to the use of $n^2$ such rectangles, where $n^2$-*n* of these rectangles do not have a point in $\mathrm{X}$ as their upper-right corner. This high number means that the method measures the discrepancy in a very representative sample of sub-areas of the 0-1 space.

From this set of $n^2$ rectangles, we can calculate the $L^\infty$ discrepancy, which is the maximum discrepancy observed over all rectangles, as well as the $\bar{L}$ discrepancy, the mean discrepancy over all rectangles. An initial analysis showed that the conclusions reached by using the $L^\infty$ discrepancy tend to be more reliable. The $L^\infty$ discrepancy measures the worst observed local area discrepancy, this will in general give a fair representation of the overall quality of coverage. The $\bar{L}$ discrepancy on the other side seems to be biased downwards, probably due to the averaging process. Not enough weight is given to high local discrepancy if discrepancy in the remainder of the sequence is at an acceptable level. The remainder of the analysis was thus only based on the $L^\infty$ discrepancy.

Another possible problem with the method is the use of the coordinates of the actual pattern of points in the testing patterns. It has been observed in the analysis that this can sometimes bias the results. Indeed, one can imagine a two-dimensional pattern where most draws are contained in the lower left quadrant. Most of the rectangles used in formula (9) will then be contained in this area of the 0-1 space. Although the method also measures positive discrepancy, i.e. excessive coverage, and will thus identify the uneven coverage over the entire 0-1 space, it will in this case not be able to precisely analyse coverage in the area that contains a very limited number of draws. An alternative approach would be to use a predetermined vector of coordinates to be used in the test pattern. One way of doing this would be to use *n* evenly placed points along each dimension, the combination of all coordinates would then give a pattern of $n^2$ evenly positioned two-dimensional coordinates in the 0-1 space. This will enable the method to also explicitly measure the discrepancy in areas with relatively low coverage. The formula for calculation of discrepancy as given in equation (9) is not changed, the coordinates at which it is used are simply replaced by these new evenly positioned coordinates.

This measure of coverage can very easily be extended to the case where $K>2$, by setting $c_k$ to be the coordinate in dimensions $k$. For $K$ dimensions, formula (9) then becomes:

$$d_X(c_1,\cdots,c_K)=\left|\frac{\prod_{k=1}^{K} c_k}{A}-\frac{\#([0,c_1]\times\cdots\times[0,c_K])}{n}\right| \qquad \ldots(10)$$

The shuffling algorithm produces different results in different runs; multiple runs must thus be used to obtain a stable estimate of the performance of the draws. Due to this increased computational cost, the detailed analysis was restricted to three different two-dimensional sequences of 100 draws. A more extended analysis, using a larger number of sequences, but fewer runs, was carried out on sequences generated from a wider choice of primes and largely confirmed the results from the detailed analysis.

Although some of the results in table 1 do indeed show superior performance by the shuffled Halton sequences (especially for the sequence based on primes 29 and 31, but also for the sequence based on primes 41 and 43), they are not as convincing as expected, given the results observed in figures 4 and 5. Indeed, according to table 1, some of the scrambled Halton sequences perform better than the corresponding shuffled sequences. As a more detailed graphical analysis showed, these results are not at all representative of the actual differences in coverage, and are generally very counterintuitive. This could indicate some problems with the measure of coverage. Although there thus seem to be some problems with equation (9), some of the results, and all of the graphical analysis, show that the use of the shuffled Halton sequences generally results in more significant improvements in quality of coverage (from the original sequences) than the use of the corresponding scrambled sequences. Table 1 also shows that, according to equation (9), the shuffled sequences do have the potential to outperform the scrambled sequences for all choices of primes used (as shown by the lower confidence interval).

Equation (10) was used in a separate analysis of the quality of coverage of higher dimensional sequences. The increased computational complexity made a large-scale analysis impossible on the computer systems used (the functions were coded in R), but a tentative analysis suggested significant advantages for the shuffled sequences.

Although some of the results produced in this section confirm the advantages of the shuffled Halton sequences, there are clearly some problems with the discrepancy measure of Dobkin et al (1993), at least in the current application. The problem seems to be that the method is not able to adequately measure the proportion of the 0-1 space that has very low or exceedingly high coverage. A closer analysis showed that the method gives too much weight to the lower left corner of the 0-1 area (or with $K>2$, the space around the origin). This is due to the fact that this area is more often included in $([0,c_1]\times\cdots\times[0,c_K])$, no matter what coordinates are used, due to the inclusion of the origin in this space. Hess and Polak (2003) use a different approach, based on using equally sized portions of the 0-1 hypercube, which are not necessarily aligned with any limiting axis. The results from that analysis are far more convincing, showing very constant performance by the shuffled sequences, and performance that is superior to that of the scrambled sequences.

Finally, it should also be noted that the method shows that when using high primes (≥17), both the scrambled and the shuffled sequences provide better coverage than the original sequence. This reflects the observations made in the graphical analysis.

This analysis was limited to sequences of 100 draws, similar results were also obtained when using a higher number of draws (c.f. Hess and Polak (2003)).

## 7. REDUCTIONS IN CORRELATION

As it was not possible to properly quantify the advantages of the shuffled sequences with the measure of coverage used in section 6, an alternative approach had to be taken. Figure 4 clearly shows how closely the quality of two-dimensional coverage is linked to the correlation between the one-dimensional sequences, with high correlation leading to poor coverage. An alternative to explicitly using a measure of coverage is thus to compare the correlation between the individual components of multi-dimensional sequences when using the different methods.

A very detailed analysis was carried out, using all possible pairs of primes in the first 25 dimensions (prime 2 to prime 97), thus also sequences based on non-neighbouring primes. Multiple runs (500) were used for each pair of primes, to get as stable an estimate of the remaining correlation as possible. The analysis showed a very stable performance by the shuffled sequences, virtually independent of the primes used. The mean absolute correlation over the 500 runs was calculated for each possible pair of primes, the values ranged from 0.0765 to 0.0871, while the variance over runs found in the single experiments ranged from 0.0034 to 0.0039. The results were thus very similar for different choices of primes. They are also very similar to the mean absolute correlation of 0.0802 found when using two pseudo-random number sequences of 100 draws each (where the variance is 0.0036). The behaviour of the shuffled sequences is thus, in terms of correlation, very similar to that of pseudo-random number sequences. This could have been expected, because of the use of a random shuffling process. These results mean that correlation in the shuffled sequences is not significantly higher than that in pseudo-random number sequences; while still providing better coverage in the one-dimensional space, and hence also generally in the multi-dimensional space.

The same analysis was carried out for the scrambled Halton sequences, using all possible pairs of primes with primes ranging from 13 to 47. The correlation matrix between the individual sequences is shown in table 2. This table shows that out of the 45 pairs of sequences, 27 have a correlation that is higher (in absolute terms) than 0.0871, the highest mean absolute correlation observed when using the shuffled sequences. Furthermore, 9 out of the 45 correlations lie above (in absolute terms) the upper 95% confidence limit on the correlation between two shuffled sequences (and 7 lie above the upper 99% confidence limit). These results seem to indicate that the reduction in correlation achieved by using the shuffled sequences is more important than that observed when using the scrambled sequences. Again, very similar results were obtained when using a higher number of draws.

Finally, just as in our analysis of coverage, the analysis of correlation showed that both the shuffled sequences and the scrambled sequences clearly offer an improvement over the original sequences when used on sequences generated from high primes (≥17), although for some choices of primes, the improvements offered by the scrambled sequences are barely significant.

## CONCLUSIONS AND RECOMMENDATIONS

This paper has presented a detailed comparison between the performance of scrambled and shuffled Halton sequences. A graphical analysis has revealed very significant advantages of the shuffled sequences over the scrambled sequences. Indeed, the power of the shuffling algorithm to disrupt the correlation between the one-dimensional Halton sequences seems to be superior to that of the corresponding scrambled Halton sequences. This allows the method to generate sequences that offer a more uniform coverage of the 0-1 space. The analysis has also shown the difficulties in giving a reliable measure of the quality of coverage provided by a pattern of points in a delimited space. Even so, the few results that seemed to actually reflect the true quality of coverage showed clear advantages for the shuffled Halton sequences (notably for the sequence based on primes 29 and 31).

An extended analysis of the effects on correlation confirmed that the power of the shuffling algorithm to reduce correlation is significantly superior to that of the scrambling approach. Indeed, this analysis showed that for 27 of the 45 pairs of primes used in the analysis, the correlation in the scrambled sequences is higher than the mean correlation in the shuffled sequences. For 9 of these pairs, the correlation in the scrambled sequences is higher than the upper 95% limit on the correlation in the corresponding shuffled sequences, and for 7 pairs it is higher than even the 99% limit in the corresponding shuffled sequences.

Another very significant advantage of the shuffled Halton sequences is the simplicity of the approach. It can easily be implemented on any computing platform and can offer considerable runtime savings over other methods due to the fact that the same original sequence can be used repeatedly with a very high guarantee of producing different multi-dimensional sequences. The method can also, unlike the scrambling approach, be generalised to any number of dimensions, without increasing its complexity or requiring any additional coding or calculations.

We would thus strongly recommend that this method be preferred to scrambled Halton sequences when dealing with high correlation and poor coverage in the original multi-dimensional Halton sequences.

Recent experiments (Hess et al, 2003) have confirmed the advantages of the shuffled Halton sequences when used in actual estimation processes. This analysis showed very constant estimation performance when using the shuffled sequences. The performance was superior to that of corresponding standard Halton sequences as well as scrambled Halton sequences and pseudo-random number sequences.

## ACKNOWLEDGMENTS

We would like to thank Andrew Daly for very helpful comments, discussions and suggestions.

# REFERENCES

Bhat, C. R. (1999), "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model", *Transportation Research*, 35B, pp.677-693

Bhat, C. R. (2002), "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences", *Transportation Research*, forthcoming.

Braaten, E. & Weller, G. (1979), "An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration", *Journal of Computational Physics*, 33, pp. 249-258.

Dobkin, D. P., Eppstein, D., & Mitchell, D. P. (1993), "Computing the Discrepancy with Applications to Supersampling Patterns", in *Proc. 9th ACM Symposium for Computational Geometry*, pp. 47-52.

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

Hess, S., Polak, J. and Daly, A. (2003), “On the performance of the shuffled Halton sequence in the estimation of discrete choice models”, accepted for presentation at the *European Transport Conference 2003*.

Hess, S. & Polak, J. (2003), “The Shuffled Halton sequence”, submitted to *Mathematical and Computer Modelling*.

Train, K. (1999), Halton sequences for mixed logit, technical paper, Department of Economics, University of California, Berkeley .

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

Tuffin, B. (1997), “Simulation accélérée par les méthodes de Monte Carlo et quasi-Monte Carlo: théorie et applications.”, PhD dissertation (in French language), Université de Rennes, France.

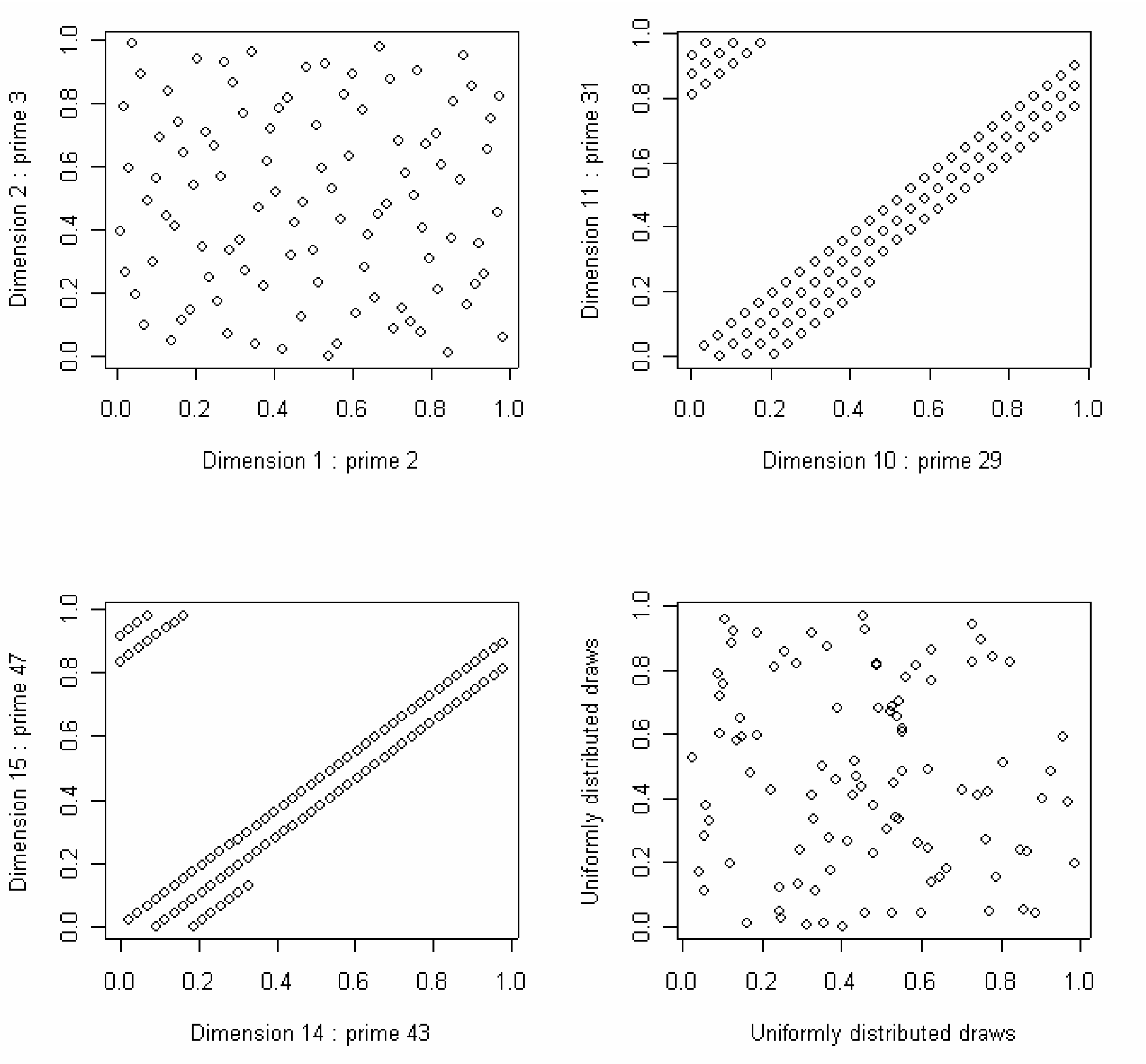

**Fig. 1. Problems with correlation in two-dimensional Halton sequences**

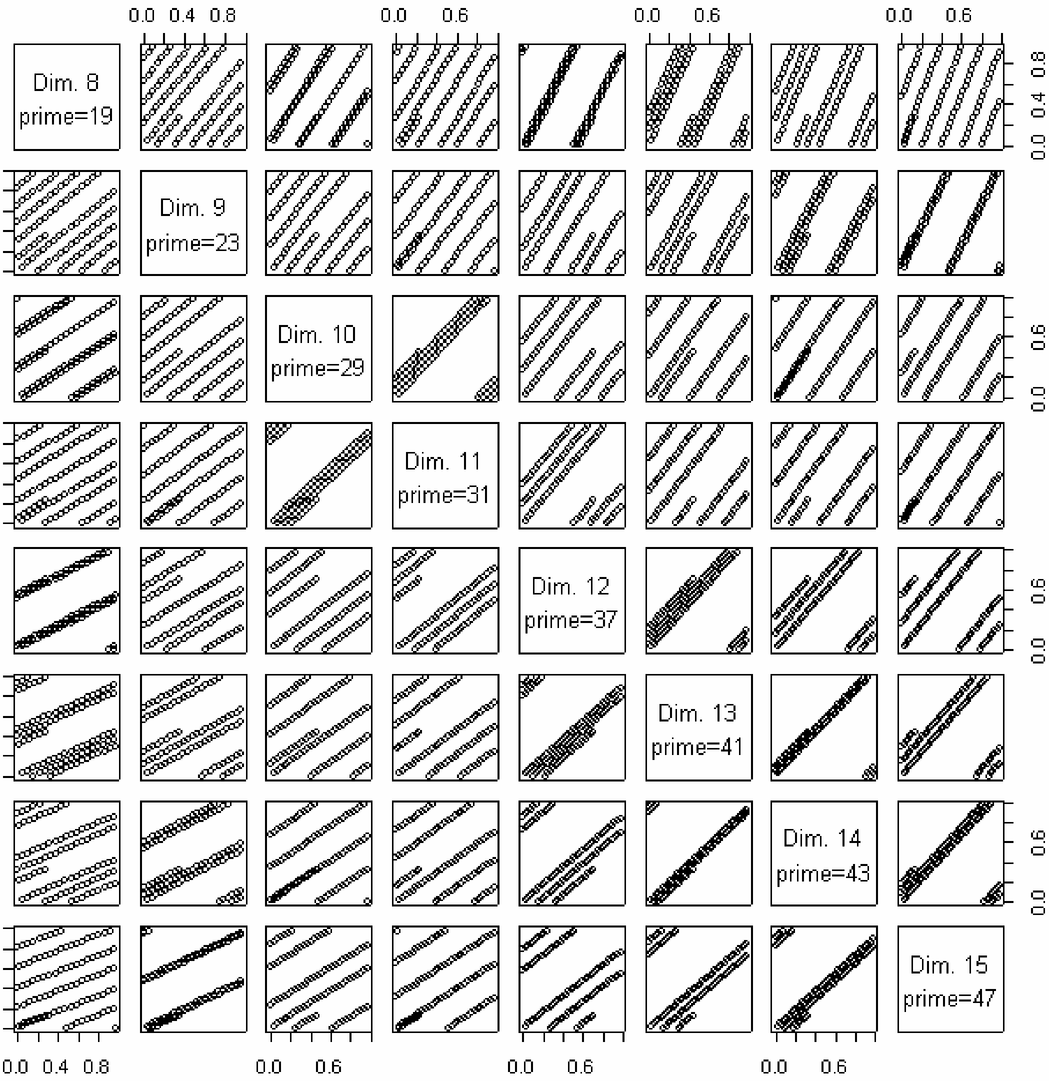

**Fig. 2. Two-dimensional Halton sequences generated from combination of primes between dimension 8 and dimension 15**

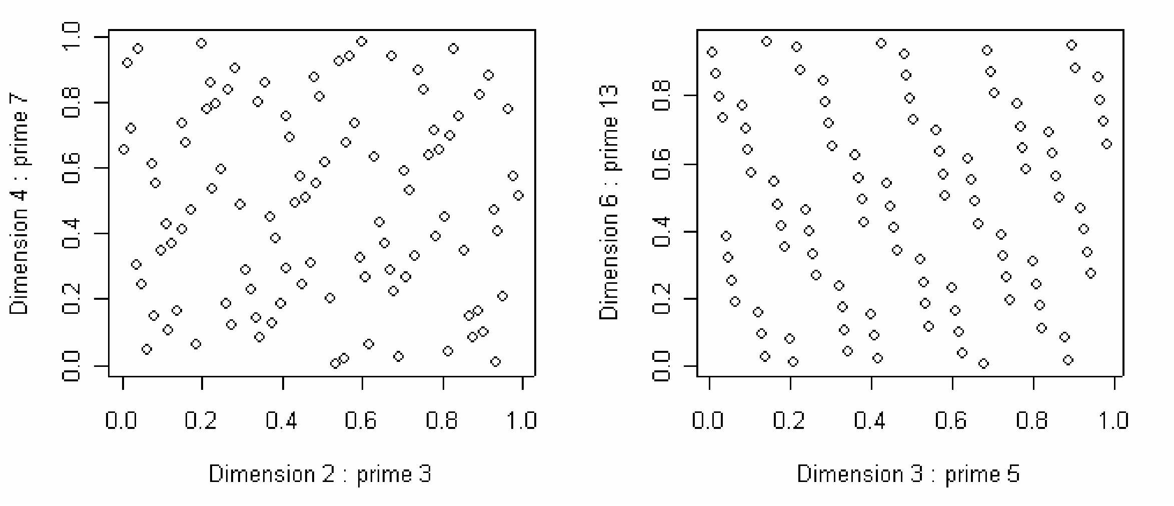

Fig. 3. Problems with correlation between Halton sequences generated from low primes

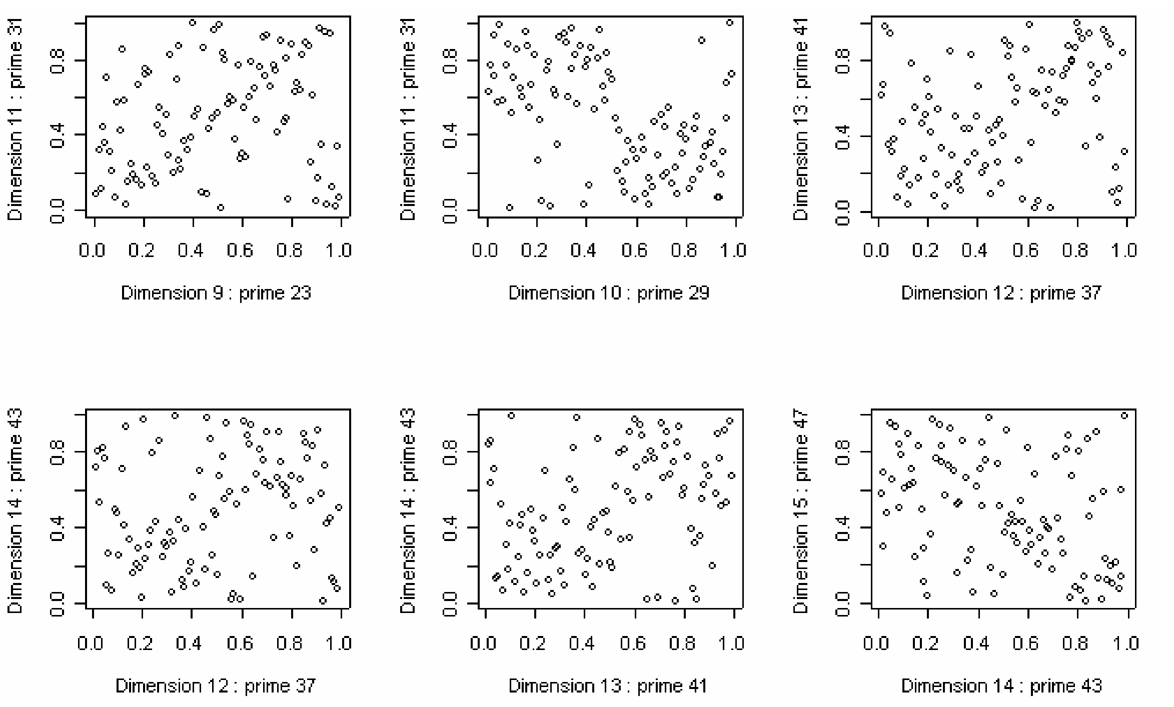

Fig. 4. Problems with high correlation and poor coverage in scrambled Halton sequences

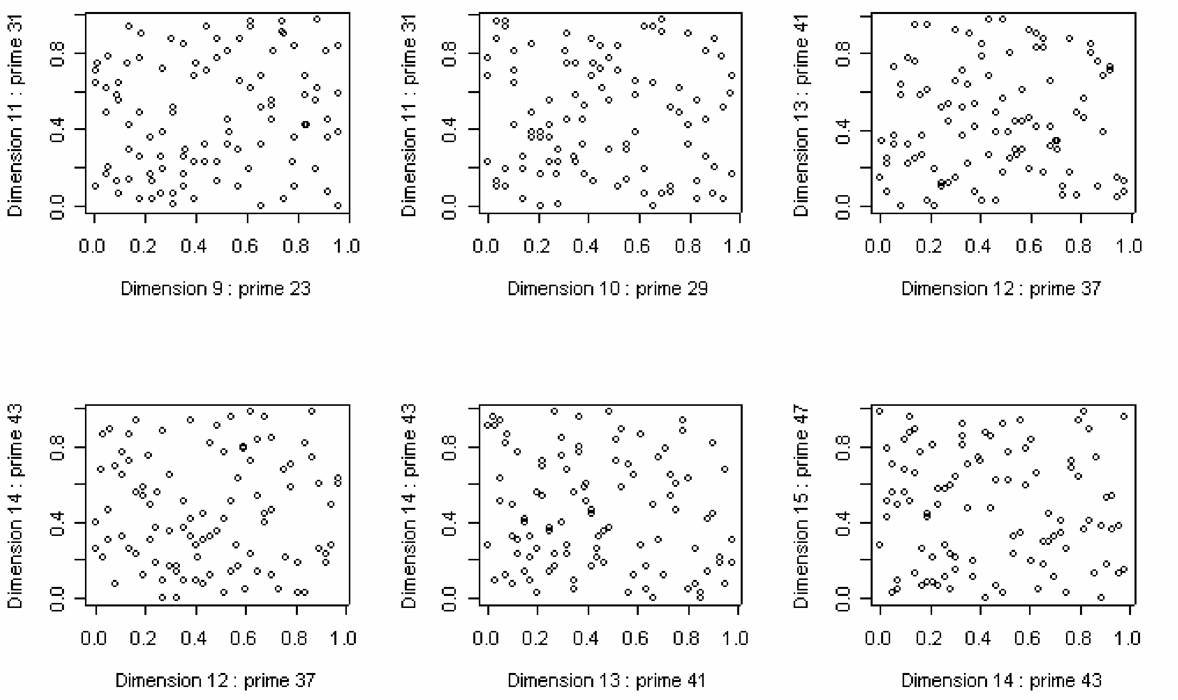

**Fig. 5. Coverage provide by shuffled Halton sequences**

**Table 1 : Discrepancy in two-dimensional Halton sequences of 100 draws**

| | | Primes 23 & 31 Dimensions 9 &11 | Primes 29 & 31 Dimensions 10 &11 | Primes 37 & 41 Dimensions 12 &13 | Primes 37 & 43 Dimensions 12 &14 | Primes 41 & 43 Dimensions 13 &14 | Primes 43 & 47 Dimensions 14 &15 |
|---|---|---|---|---|---|---|---|
| | Scrambled sequence | 0.0957 | 0.1803 | 0.1295 | 0.105 | 0.1348 | 0.1115 |
| Shuffled sequences | Lower 95% limit | 0.0755 | 0.0938 | 0.1173 | 0.0989 | 0.1056 | 0.0963 |
| | Mean | 0.0975 | 0.1142 | 0.1422 | 0.1195 | 0.1284 | 0.1139 |
| | Upper 95% limit | 0.1195 | 0.1346 | 0.1671 | 0.1401 | 0.1512 | 0.1315 |

**Table 2 : Correlation between scrambled Halton sequences of 100 draws**

| | **Prime 13** | **Prime 17** | **Prime 19** | **Prime 23** | **Prime 29** | **Prime 31** | **Prime 37** | **Prime 41** | **Prime 43** |
|---|---|---|---|---|---|---|---|---|---|
| **Prime 17** | -0.01846 | | | | | | | | |
| **Prime 19** | 0.063063 | 0.33435 | | | | | | | |
| **Prime 23** | -0.01103 | -0.15156 | -0.00853 | | | | | | |
| **Prime 29** | -0.09241 | -0.14344 | -0.02671 | -0.27799 | | | | | |
| **Prime 31** | 0.039695 | 0.162972 | 0.092292 | 0.221403 | -0.4492 | | | | |
| **Prime 37** | 0.181811 | 0.060718 | 0.092974 | 0.001876 | 0.073196 | 0.122347 | | | |
| **Prime 41** | -0.11697 | 0.087703 | 0.09592 | 0.009998 | -0.06049 | 0.061755 | 0.285909 | | |
| **Prime 43** | 0.106949 | -0.04095 | -0.00665 | 0.076421 | -0.02143 | 0.200931 | 0.138587 | 0.31837 | |
| **Prime 47** | 0.111372 | -0.13978 | -0.18581 | -0.16359 | 0.042235 | -0.08658 | -0.19614 | -0.31577 | -0.34897 |