Generating Jointly Distributed Variates by Restricted Random Sampling by William R. Schucany, Lori A. Thombs and Kelly Cunningham

Technical Report No. SMU-DS-TR-197 Southern Methodist University

March 1986

Research sponored by Sandia Laboratories PO #25-7977, Project #5-27565

> Department of Statistics Southern Methodist University Dallas, TX 75275

Generating Jointly Distributed Variates by Restricted Random Sampling

William R. Schucany, Lori A. Thombs and Kelly Cunningham Dept. of Statistics, Southern Methodist University

Summary

This report presents several results pertaining to bias in Monte Carlo estimates of cumulative distribution functions, means and variances when restricted pairing is employed. Some analytical results are derived to yield bounds on the magnitude of the difference between the true cdfs estimated by random sampling and by restricted random sampling. The primary focus of this report is the bivariate distribution with uniform marginals and zero correlation. Several of the findings extend to higher dimensions, other distributions and low to moderate amounts of correlation.

1. Introduction

The restricted pairing technique, Iman and Conover (1982), which transforms an (n×k) matrix, R, whose columns are random permutations of van der Waerden scores, is used to induce specified rank correlation among the k input variables. The resulting score matrix, R*, specifies how the n observations for each input variable X_j should be <u>ordered</u>. Observations for input may be obtained by simple random sampling (RS) or Latin Hypercube Sampling (LHS). Each $(n \times 1)$ vector \underline{X}_{j} is reordered according to the jth column of the R* matrix, which results in an $(n \times k)$ input matrix having rank correlation matrix exactly as specified by the user.

Although Latin Hypercube Sampling is used in practice in uncertainty analysis for large-scale, computer models more often than RS, attention is focused on the latter. In their paper introducing the restricted pairing technique, Iman and Conover (1982) present two methods. It is the "variance reduction technique", the second method, that is studied. A short summary of the technique is as follows:

1. Generate a (n×k) matrix R whose columns are random permutations of $\Phi^{-1}(i/n+1)$. The sample (Pearson) correlation matrix of R is denoted by T.

2. Use Cholesky Factorization to get Q such that T = QQ'. Also use Cholesky Factorization to obtain a matrix P such that C = PP', where C is the desired (rank) correlation matrix for the k input variables.

3. Calculate $R^* = R(PQ^{-1})'$. Each row $\frac{R^*}{-1} = \frac{R}{-1}(PQ^{-1})'$ will have correlation matrix $PQ^{-1}TP' = C$.

4. Independently generate n observations of each of the k-variates and order them as specified by each of the k columns of R*.

The fact that the <u>sample</u> correlation matrix, T, is used rather than the theoretical identity matrix is a matter for concern for small samples only. In other words, the "variance reduction" distinction vanishes as $n \rightarrow \infty$, due to the fact that $T \rightarrow I$. Thus the main concern that one might reasonably have, pertains to the small sample effect of forcing the input correlation matrix to equal the target, C, with no Monte Carlo sampling error in this characteristic.

2

A question of interest is how the output $Z = g(X_1, ..., X_k)$ is affected by the technique. That is, are the estimates of the cdf, mean and variance of Z unbiased? The general concern connected with not allowing sampling variability may be illustrated in the simple univariate situation of iid observations from $N(0,\sigma^2)$. Suppose that analogous to steps 2 and 3, we utilize the sample standard deviation, S, and rescale the data to better conform to the known target value σ . If it happens that the statistic of interest is $Z = S^2$, then we have introduced a severe bias in our estimate of the cdf. That is, instead of learning about the chi-square sample distribution we produce a degenerate cdf.

In the section that follows the bivariate (k=2) case is considered in connection with several transformations $g(X_1,X_2)$. To ease the notational burden on subscripts Z = g(X,Y) will be used. To avoid the ambiguity about the structure of the particular bivariate distribution that is simulated the case of independent X and Y is studied. It follows that the target correlation matrix C=I, the identity. Therefore the subject of this investigation becomes a comparison of Z = g(X,Y) where X,Y are truly independent (RS) with $Z^* = g(X^*,Y^*)$ where X^*,Y^* are generated by restricted pairing (RRS).

2. Exact Distribution for Small Samples

Two settings are considered in some detail:

(i) the input variables are independent and simple random sampling is used to form the $n\times 2$ input matrix of uniform (0,1) variates (RS),

(ii) input variables are all independent and the restricted pairing technique is used to get the input matrix (RRS).

For relatively small n the exact distributions of Spearman's rank correlation in the two cases are easily obtainable. For n=4 pairs of input variables, the distributions (to be examined in more detail later) are given below.

True	Independence:	Restricted	pairing:
r	P(R=r)	r	P(R=r)
-1.	1/24		
8	3/24		
6	1/24		
4	4/24	4	2/22
2	2/24	2	6/22
0	2/24	0	6/22
.2	2/24	.2	6/22
.4	4/24	. 4	2/22
.6	1/24		
.8	3/24		
1.	1/24	4	

Since the restricted pairing does not allow samples having rank correlation coefficient of $\pm 1., \pm .8$ and $\pm .6$, (which have positive probability in the true independence case) the method may be <u>too</u> restrictive. Positive correlation near 1 results when $X_{(1)}$ and $Y_{(1)}$ are paired, or when $X_{(4)}$ and $Y_{(4)}$ are paired. Similarly, the occurrence of pairs $(X_{(4)}, Y_{(1)})$ and $(X_{(1)}, Y_{(4)})$ in the sample yield high negative correlation. Samples with rank correlation near ± 1 are <u>not</u> permitted in the restricted pairing technique. Therefore one might expect scatterplots of the (X^*, Y^*) pairs to exhibit somewhat sparse realizations near the corners of the unit square. As the two plots of 1000 pairs in each of Figures 1a and 1b demonstrate, this is not visually detectable. The points in Figure 1a are 1000 pairs of independent uniform (0,1) random variables. The points in Figure 1b are

4



Figure la

1000 pairs of iid Uniform (0,1): RS





250 sets of 4 pairs of uniforms: RRS

250 sets of 4 pairs (n=4) generated by restricted pairing of uniform (0,1) order statistics. Any differences which may exist, even in this extremely small sample case, are not obvious.

The joint distribution, $f^*(x,y)$ of a randomly selected pair generated by RRS can be derived for comparison with $f(x,y) = I_{(0,1)}(x) \cdot I_{(0,1)}(y)$. There are n!-2 = 22 permissable, equally likely, distinct R matrices of van der Waerden scores. The transformation $R^* = RQ^{-1}$ is applied in each case. This results in ten different outcomes, each having 4 pairs of uniform order statistics.

These final pairings are given below with the associated rank correlation and likelihood.

3/22

3/22

3/22

1/22

1/22

Prob.:

x ₍₁₎	^ү (3)	X ₍₁₎ Y ₍₃₎	x ₍₁₎ Y ₍₁₎	X ₍₁₎ Y ₍₂₎	x ₍₁₎ y ₍₁₎
x ₍₂₎	¥ ₍₁₎	X ₍₂₎ Y ₍₂₎	X ₍₂₎ Y ₍₄₎	X ₍₂₎ Y ₍₃₎	X ₍₂₎ Y ₍₄₎
x ₍₃₎	¥ ₍₄₎	X ₍₃₎ Y ₍₁₎	X ₍₃₎ Y ₍₃₎	X ₍₃₎ Y ₍₁₎	X ₍₃₎ Y ₍₂₎
X(4)	¥ ₍₂₎	X ₍₄₎ Y ₍₄₎	^X (4) ^Y (2)	X ₍₄₎ Y ₍₄₎	X ₍₄₎ Y ₍₃₎
Rank Corr:	0	• 2	.2	.4	.4
Prob.	3/22	3/22	3/22	1/22	1/22

The pdf $f^*(x,y)$ can be calculated as follows. Each pair $(X_{(i)}, Y_{(j)})$ has joint distribution $g_{ij}(x,y) = g_i(x)g_j(y)$, where $g_i(x)$ is the distribution of the ith order statistic of a sample of size 4 from a uniform (0,1), namely a beta distribution with parameters 1 and 5-i. The desired pdf is a mixture of the 10 situations itemized above,

$$f^{*}(x,y) = \frac{1}{22\left\{\frac{1}{4\left[g_{14}(x,y) + g_{21}(x,y) + g_{33}(x,y) + g_{42}(x,y)\right]\right\}}}{+ \frac{1}{22\left\{\frac{1}{4\left[g_{13}(x,y) + g_{22}(x,y) + g_{34}(x,y) + g_{41}(x,y)\right]\right\}}}{+ \dots + \frac{1}{22\left\{\frac{1}{4\left[g_{11}(x,y) + g_{24}(x,y) + g_{32}(x,y) + g_{43}(x,y)\right]\right\}}},$$

which simplifies to $f^{*}(x,y) = \frac{16}{2} w_{ij}g_{ij}(x,y) = \frac{16}{2} w_{ij}g_{ij}(x)g_{j}(y),$

where the weights w ij are

		ţ				
	-	1	2	3	4	
	1	4/88	7/88	7/88	4/88	
i	2	7/88	4/88	4/88	7/88	ł
	3	7/88	4/88	4/88	7/88	1
	4	4/88	7/88	7/88	4/88	<u> </u>

In the case of true independence f(x,y) can also be represented as a weighted sum of $g_{ij}(x,y)$, with all $w_{ij} = 1/16$. It can easily be shown using the binomial formula that the 16-term sum simplifies to $f(x,y) = I_{(0,1)}(x)I_{(0,1)}(y)$. In general, for the independence case, the w_{ij} weights are $w_{ij} = 1/n^2$, while the RRS procedure gives different w_{ij} 's, say w_{ij}^* . It should be possible to write the w_{ij}^* 's as functions of n. This could be of some value since one might consider the quantity $T = \max[w_{ij} - w_{ij}^*]$

or $\Sigma\Sigma(w_{ij}-w_{ij}^*)^2$ and study the rate at which T vanishes as n gets large. At this time no such expression for the explicit dependence of the w_{ij}^* on n is known.

After much algebra (eventually verified by REDUCE) it can be shown that all of these polynomials (products and sums of beta densities) simplify to

$$f^{*}(x,y) = \frac{1}{11}(8 + 18x + 18y - 18x^{2} - 18y^{2} - 108xy + 108xy^{2} + 108x^{2}y - 108x^{2}y^{2}).$$

Four specific transformations were studied to make comparisons of exact moments and derived distributions. The expressions for the densities differ substantially in form. Consider the four transformations

1) $z_1 = g_1(x,y) = xy$ 2) $z_2 = g_2(x,y) = x + y$ 3) $z_3 = g_3(x,y) = Max\{x,y\}$ 4) $z_4 = g_4(x,y) = 2x + y^{1/3}$

and the associated pdf of Z will be denoted by $f_1(z)$ or $f_1^*(z)$ when the X,Y pairs are by RS or RRS, respectively.

1. Product

$$g_1(x,y) = xy$$

True Independence:

$$f_{1}(z) = -\ln z I_{(0,1)} (z)$$

$$E(Z) = \frac{1}{4}$$

$$Var(Z) = \frac{7}{144} = .04861$$

Restricted Pairing:

$$f_{1}^{*}(z) = \frac{1}{11}(18 + 180z - 198z^{2} - 8 \ln z + 108z \ln z + 108z^{2} \ln z) \cdot I_{(0,1)}(z)$$
$$E(Z^{*}) = \frac{1}{4}$$
$$Var(Z^{*}) = \frac{1913}{39600} = .04831$$

2. <u>Sum</u>

$$g_2(x,y) = x+y$$

True Independence:

$$f_{2}(z) = zI_{(0,1)}(z) + (2-z)I_{(1,2)}(z)$$

E(Z) = 1
Var(Z) = $\frac{1}{6}$

Restricted Pairing:

$$f_{2}^{*(z)} = p(z)I_{(0,1)}(z) + [2p(1)-p(z)]I_{[1,2)}(z)$$

$$p(z) = (40z + 90z^{2} - 150z^{3} + 90z^{4} - 18z^{5})/55$$

$$E(2^{*}) = 1$$

$$Var(2^{*}) = \frac{1}{6}$$

3. <u>Maximum</u>

$$g_3(x,y) = Max\{x,y\}$$

True Independence:

$$f_{3}(z) = 2zI_{(0,1)}(z)$$

$$E(Z) = \frac{2}{3}$$

$$Var(Z) = \frac{1}{18} = .0555.$$

Restricted Pairing:

$$f_{3}^{*}(z) = (16z + 54z^{2} - 156z^{3} + 180z^{4} - 72z^{5})/11 \cdot I_{(0,1)}(z)$$
$$E(Z^{*}) = \frac{1543}{2310} = .6679$$
$$Var(Z^{*}) = \frac{29413}{533610} = .05512$$

4. Asymmetric

$$g_4(x,y) \approx 2x + \sqrt[3]{y}$$

True Independence:

$$f_{4}(z) = \begin{cases} \frac{1}{2} z^{3} , z \in [0,1) \\ \frac{1}{2} , z \in [1,2) \\ \frac{1}{2} [1-(z-2)^{3}] , z \in [2,3] \end{cases}$$

Restricted Pairing:

$$f_{4}^{*}(z) = (z^{3}(-504z^{8} + 5544z^{7} - 18480z^{6} + 1485z^{5} - 11880z^{4} + 27720z^{3} - 1386z^{2} + 6930z + 24640))/67760, \quad 0 \le z < 1$$

$$= (-594z + 34663)/67760, \quad 1 \le z < 2$$

$$= (504z^{11} - 5544z^{10} + 18480z^{9} - 1485z^{8} + 11880z^{7} - 1269576z^{6} + 7452522z^{5} - 21295890z^{4} + 35900480z^{3} - 36608880z^{2} + 20983182z - 5152185)/67760, \quad 2 \le z \le 3.$$

Figures 2 through 5 display two pdfs for transformations 1 through 4, respectively. In each case the graphs of the two densities exhibit small differences between the distributions under RS and RRS. Changing to the cdf domain, a comparison of upper tail probabilities for $g_1(X,Y) = XY$ is given below.

	RS	RRS		
k	P(Z > k)	$P(Z^* > k)$	difference	
.900	.00518	.00427	.000910	
.925	.00289	.00231	.000580	
.950	.00127	.00099	.000280	
.975	.000315	.000237	.000078	
.990	.0000502	.0000370	.000047	
.995	.0000125	.0000092	.000003	

At any particular point k, $P(Z^* > k)$ is biased for P(Z > k) but the problem is not severe. It is not easy to obtain results for general n to investigate the bias as n grows. Moments of the distributions differ very little, with $Var(Y^*) \leq Var$ (Y) in all cases examined.

In light of these results the natural line of investigation would seem to be to seek an upper bound on

$$\Delta_{g} = \sup_{z} |F(z) - F^{*}(z)|,$$

the maximum difference between the cdfs of the output Z = g(X,Y) produced by RS and RRS. For a single specific function, g, such a bound is not difficult to derive. To be of practical importance the value of Δ , that holds for a broad class of functions desired. In other words, if

 $\Delta = \sup_{g \in \Gamma} \Delta_g$

Pdf of $Z = X \cdot Y$



Figure 2

Pdf of Z = X + Y



Figure 3

Pdf of $Z = Max{X,Y}$



Figure 4

Pdf of Z = $2X + Y^{1/3}$



Figure 5

is small and Γ contains a rich collection of transformations, then this could allay most concern about RRS as a methodology for estimating cdfs.

The class of transformations examined is z = g(X,Y) = X + Y/c. This transformation is motivated by the following: consider the class of transformations that are well approximated by a first order Taylor series expansions, i.e.,

$$g(X,Y) = a_0 + a_1^X + a_2^Y$$
.

Here we have

$$|F_{Z}(z) - F_{Z}^{*}(z)| \stackrel{:}{=} |P(a_{0} + a_{1}X + a_{2}Y \leq z) - P^{*}(a_{0} + a_{1}X + a_{2}Y \leq z)|$$

= $|P(X + \frac{a_{2}}{a_{1}}Y \leq \frac{z - a_{0}}{a_{1}}) - P^{*}(X + \frac{a_{2}}{a_{1}}Y \leq \frac{z - a_{0}}{a_{1}})|.$

The value of a_0 is immaterial to the calculation of $\sup_{Z} |F_Z(z) - z|$ $F_Z^*(z)|$, and a_1, a_2 enter into the calculation only through their ratio. They affect the location only; not the magnitude.

Four distinct cases arise depending on the size of |l/c| and the sign of c. These are:

Let

$$G_{1}^{\star}(z) = \frac{|c|}{11} (-4z^{2} - 3(c+1)z^{3} + \frac{3}{2}(c^{2} + 3c + 1)z^{4} - \frac{9}{5}c(c+1)z^{5} + \frac{3}{5}c^{2}z^{6})$$

$$G_{2}^{\star}(z) = \frac{|c|}{11}(\frac{3}{5c^{3}} - \frac{11}{c})z$$

$$G_{3}^{\star}(z) = \frac{|c|}{11}(11 - \frac{3}{5}c^{2})z + G_{1}^{\star}(z)$$

and

$$G_{1}(z) = \frac{-|c|}{2}z^{2}$$

$$G_{2}(z) = \frac{-|c|}{c}z$$

$$G_{3}(z) = |c|(y-y^{2}/2)$$

then the distribution function of Z = X + 1/cY is

Case (1), c ϵ (- ∞ ,-1),

$$F(z) = \begin{cases} 0 & z \in (-\infty, \frac{1}{c}) \\ G_2(z) - G_1(z) - G_2(\frac{1}{c}) + G_1(\frac{1}{c}) & z \in [\frac{1}{c}, 0) \\ G_2(z) - G_2(\frac{1}{c}) + G_1(\frac{1}{c}) & z \in [0, 1 + \frac{1}{c}) \\ G_2(1 + \frac{1}{c}) + G_3(z) - G_2(\frac{1}{c}) + G_1(\frac{1}{c}) - G_3(1 + \frac{1}{c}) & z \in [1 + \frac{1}{c}, 1) \\ 1 & z \in [1, \infty) \end{cases}$$

Case (2),

$$F(z) = \begin{cases} c \in [-1,0) \\ 0 & z \in (-\infty,\frac{1}{c}) \\ 0 & z \in (-\infty,\frac{1}{c}) \\ 0 & z \in [\frac{1}{c},1+\frac{1}{c}) \\ 0 & z \in [1+\frac{1}{c},0) \\ 0 & z \in [1,\infty) \end{cases}$$

Case 3,
$$c \in (0,1)$$

$$\begin{array}{c} 0 \\ -G_1(z) \end{array} z \varepsilon (-\infty, 0) \\ z \varepsilon [0, 1) \end{array}$$

$$F(z) = \begin{cases} G_3(z) - G_1(z) - G_3(1) & z \in [1, \frac{1}{c}) \end{cases}$$

$$\begin{bmatrix} -G_{1}(\frac{1}{c}) - G_{3}(1) + G_{3}(z) - G_{2}(z) + G_{2}(\frac{1}{c}) , & z \in [\frac{1}{c}, 1 + \frac{1}{c}) \\ 1 & z \in [1 + \frac{1}{c}, \infty) \end{bmatrix}$$

Case 4
$$c \in [1,\infty]$$

$$F(z) = \begin{cases} 0 & z \in (-\infty,0) \\ -G_1(z) & z \in [0,\frac{1}{c}) \\ G_2(\frac{1}{c}) - G_1(\frac{1}{c}) - G_2(z) & z \in [\frac{1}{c}, 1) \\ -G_1(\frac{1}{c}) + G_2(\frac{1}{c}) + G_3(z) - G_2(z) - G_3(1), & z \in [1, 1 + \frac{1}{c}) \\ 1 & , & z \in [1 + \frac{1}{c}, \infty) \end{cases}$$

Replace G with G_{i}^{*} for the distribution function of Z*.

A plot of $\Delta_c = \sup_{z} |F^*(z,c)-F(z,c)|$ was obtained by numerical methods for various values of c. Figure 6 indicates that $\sup_{c} \sup_{z} |F^*(z;c) - F(z;c)|$ occurs at |c| = 1 and that $\Delta_c \rightarrow 0$ as $c \rightarrow \pm \infty$ or $0 \pm .$ Since $Z \rightarrow X$ and $Z^* \rightarrow X$ as $|c| \rightarrow \infty$ it follows immediately that $\Delta_c \rightarrow 0$. For $c \rightarrow 0^+$, $Z^* \rightarrow \infty$ and $Z \rightarrow \infty$ so $F^*(z) \rightarrow 0$ and $F(z) \rightarrow 0$ and $\sup_{Z} |F^*(z) - F(z)| \rightarrow 0$. Finally for $c \rightarrow 0^-$, $Z^* \rightarrow -\infty$, $Z \rightarrow -\infty$ so $F^*(z) \rightarrow 1$ and $F(z) \rightarrow 1$, thus $\sup_{Z} |F^*(z) - F(z)| \rightarrow 0$. At $c = 1 \sup_{Z} |F^*(z) - F(z)| = 5.71 \times 10^{-3}$ occurs at z = .78 and 1.22. Maximum over all z of $|F^{*}(z;c)-F(z;c)|$ times 10^{3} as a function of c



Figure 6

In view of the analytical difficulties encountered with the first order Taylor series class it is not conjectured that the broader class of all quadratic response functions would lend itself to similar analysis. Comparable complexities arise in attempts to derive the exact distribution of any summary statistic for the entire RRS Monte Carlo run. This result would allow an examination of the issue of whether RRS truly produces a variance reduction uniformly over all types of statistics.

Conclusion

In the smallest of samples the maximum deviation of the cdf produced by RRS from that by RS is not large. Therefore in low correlation or independent cases, the biases for moderate size n appear to be inconsequential. At the same time it is difficult to demonstrate that the variance reduction that one obtains from RRS is worth the introduction of some bias.

Reference

Iman, R. L. and Conover, W. J. (1982) "A distribution-free approach to inducing rank correlation among input variables", <u>Commun. Statist.</u>, <u>B</u>, 11, 311-334.