**Applied Mathematics E-Notes, 8(2008), 279-282 ©** ISSN 1607-2510
Available free at mirror sites of http://www.math.nthu.edu.tw/∼amen/

# On The Generation Of Correlation Matrices*

Mark Budden, Paul Hadavas and Lorrie Hoffman†

Received 30 August 2007

**Abstract**

The authors previously provided an algorithm and program for generating $4 \times 4$ correlation matrices. In this note, the algorithm is refined and extended to the generation of $n \times n$ correlation matrices for $n > 4$.

## 1 Introduction

A common problem encountered by many researchers in statistics is the generation of valid correlation matrices for use in Monte Carlo studies. While the generation of $3 \times 3$ correlation matrices is rather straight-forward, statisticians are often left to construct possible $n \times n$ correlation matrices (when $n > 3$) that must then be checked for positive semidefiniteness. In the case of $4 \times 4$ correlation matrices, approximately 18.3% of the possible "guesses" turn out to be valid and the likelihood decreases as the size of the matrices increases. This percentage was noted by Rousseeuw and Molenberghs (see [4], § 3) and was determined by randomly generating millions of possible $4 \times 4$ correlation matrices that were checked for positive semidefiniteness. Recent studies [2, 3] by Mishra have used a global optimization method known as differential evolution to create valid correlation matrices. Mishra has been able to complete correlation matrices of arbitrary size when faced with an incomplete matrix.

The authors [1] have provided an algorithm for generating valid $4 \times 4$ correlation matrices that allows one to randomly pick three of the correlations, and then provides successive bounds for each of the remaining correlations that insure positive semidefiniteness. In this article, we explain how the algorithm of [1] can be simplified and extended to the generation of $n \times n$ correlation matrices, relying solely on bounds provided by the determinants of the matrices.

## 2 The Structure of Correlation Matrices

Let $x_1, x_2, \ldots, x_n$ be random variables with $r_{ij}$ denoting the correlation coefficient between $x_i$ and $x_j$. Then a correlation matrix is a matrix of the form $(r_{ij})$. Such a matrix is clearly symmetric, has 1's along the diagonal, and all of its entries lie within the interval $[-1, 1]$. However, not every matrix satisfying these three properties is a

[*] Mathematics Subject Classifications: 65C05, 65C60.

[†] Department of Mathematics, Armstrong Atlantic State University, Savannah, GA 31419

correlation matrix. It is well-known that a necessary and sufficient condition for such a matrix to be a correlation matrix is the positive semidefiniteness of the matrix. This is a property that is relatively simple to verify, but not easily constructed.

For the remainder of this article, we will assume that all correlations between distinct random variables lie within $(-1, 1)$ since the values $\pm 1$ indicate a redundancy in the data and hence, a potential reduction in the number of random variables. In the case of 2 random variables, every matrix of the form

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}, \qquad \text{where} \;\; r_{12} \in (-1, 1),$$

is positive definite (and therefore, positive semidefinite).

The construction of $3 \times 3$ correlation matrices has been considered by many authors (for example, see Stanley and Wang [5]). If $r_{12}$ and $r_{13}$ are randomly chosen from the interval $(-1, 1)$ and $r_{23}$ is chosen so that

$$r_{12}r_{13} - \sqrt{(1-r_{12}^2)(1-r_{13}^2)} \le r_{23} \le r_{12}r_{13} + \sqrt{(1-r_{12}^2)(1-r_{13}^2)}, \tag{1}$$

then

$$A_{123} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}$$

is a valid correlation matrix. Any value of $r_{23}$ chosen outside of the given range results in a matrix that is not positive semidefinite.

Until recently, there were no necessary and sufficient conditions on the bounds of the correlations in an $n \times n$ correlation matrix that would insure positive semidefiniteness when $n \geq 4$. In [1], the authors provided an algorithm for the generation of $4 \times 4$ correlation matrices that provided both necessary and sufficient bounds. This result utilized the fact that a symmetric matrix is positive semidefinite if and only if the determinants of the matrix and all of its principal minor matrices are nonnegative.

In the next section, we provide an algorithm for the generation of $n \times n$ matrices for $n \geq 4$. The algorithm not only extends the work in [1], but simplifies the algorithm in the $4 \times 4$ case. The basis of the algorithm is the observation that given an $n \times n$ correlation matrix, removing the last column and the last row results in an $(n-1) \times (n-1)$ correlation matrix and that every $(n-1) \times (n-1)$ correlation matrix can be extended to form an $n \times n$ correlation matrix. The first part of this statement is clear as it follows from positive semidefiniteness. The second part follows since one only needs to include the correlations related to the introduction of an $n^{th}$ random variable in a manner that assures that the determinant of the $n \times n$ matrix is greater than or equal to zero.

## 3 The Algorithm

Suppose that $A_{12\cdots(n-1)} = (r_{ij})$ is an $(n-1) \times (n-1)$ correlation matrix that we wish to extend to an $n \times n$ correlation matrix $A_{12\cdots n}$. We begin by picking $r_{1n} \in (-1, 1)$.

This range is justified as none of the correlations in $A_{12\cdots(n-1)}$ involve the random variable $x_n$. To determine the possible ranges for the remaining correlations $r_{jn}$, where $j \in \{2, 3, \ldots, n-1\}$, we note that all of the proper principal minor matrices have nonnegative determinant. Thus, we only need to ensure that the determinant of $A_{12\cdots n}$ is nonnegative.

We complete the generation of an $n \times n$ correlation matrix by successively picking correlations $r_{2n}, r_{3n}, \ldots, r_{(n-1)n}$ based upon the correlations that have already been chosen. This is done by first noting that each correlation $r_{jn}$ is subject to the bounds

$$L_{jn}^{(1)} \le r_{jn} \le U_{jn}^{(1)},$$

where

$$L_{jn}^{(1)} = r_{1j}r_{1n} - \sqrt{(1-r_{1j}^2)(1-r_{1n}^2)}$$

and

$$U_{jn}^{(1)} = r_{1j}r_{1n} + \sqrt{(1-r_{1j}^2)(1-r_{1n}^2)}.$$

The second set of conditions on each correlation is not quite as easy to describe.

Assume that $r_{2n}, r_{3n}, \ldots, r_{(j-1)n}$ have already been chosen. To determine the range of values for $r_{jn}$, we must be sure that a correlation chosen from the range can occur in some $n \times n$ correlation matrix that has all of the previously chosen correlations. This leads to the determinant of $A_{12\cdots n}$, which we consider as a quadratic in $r_{jn}$. Of course, the coefficients of this quadratic contain the "variables" $r_{(j+1)n}, r_{(j+2)n}, \ldots, r_{(n-1)n}$. For each choice of these variables, the quadratic $\det(A_{12\cdots n})$ opens downward and has real roots, which are easily determined. The smaller of the two roots (which we denote by $R_1(r_{(j+1)n}, r_{(j+2)n}, \ldots, r_{(n-1)n})$) gives a lower bound for $r_{jn}$ that ensures positive semidefiniteness, while the larger root (denoted $R_2(r_{(j+1)n}, r_{(j+2)n}, \ldots, r_{(n-1)n})$) gives an upper bound.

Let $L_{jn}^{(2)}$ be given by

$$\min\{R_1(r_{(j+1)n}, r_{(j+2)n}, \ldots, r_{(n-1)n}) \mid L_{kn}^{(1)} \le r_{kn} \le U_{kn}^{(1)} \text{ for each } j+1 \le k \le n-1\}$$

and $U_{jn}^{(2)}$ be given by

$$\max\{R_2(r_{(j+1)n}, r_{(j+2)n}, \ldots, r_{(n-1)n}) \mid L_{kn}^{(1)} \le r_{kn} \le U_{kn}^{(1)} \text{ for each } j+1 \le k \le n-1\}.$$

Optimization of multivariable functions over closed regions is achieved using commands such as the `Maximize` and `Minimize` commands new to version 5 of Mathematica. Picking $r_{jn}$ from the interval

$$\max\{L_{jn}^{(1)}, L_{jn}^{(2)}\} \;\le\; r_{jn} \;\le\; \min\{U_{jn}^{(1)}, U_{jn}^{(2)}\}$$

guarantees the existence of an $n \times n$ correlation matrix with the chosen correlations.

In the pseudocode below, we provide a formal description of the process. In this manner, any arbitrarily-sized correlation matrix can be generated. This algorithm uses the matrix size as the only input, a random number generator for selecting correlations, and Mathematica for the optimization in the second set of bounds. In the absence of such software, a grid search over the appropriate hypercube would be necessary.

```
algorithm validcor;
begin
    input size, $n$;
    randomly select $r_{12}$ from $(-1,1)$;
    for $k = 3$ to $n$ do
    begin
        randomly select $r_{1k}$ from $(-1,1)$;
        for $j = 2$ to $k-1$ do
        begin
            lower $= \max\{L_{jk}^{(1)}, L_{jk}^{(2)}\}$;
            upper $= \min\{U_{jk}^{(1)}, U_{jk}^{(2)}\}$;
            randomly select $r_{jk}$ from (lower, upper);
        end;
    end;
end;
```

In conclusion, we have shown that it is possible to generate any correlation matrix by successively adding variables up to the size of the desired matrix. We note here that this program is flexible as to the end-user's needs, as it could be extended to generate a specified number of correlation matrices for a particular size or, if a certain matrix was required, the random selection could be replaced with the user prompted to enter a correlation within the specified range. Overall, this method simplifies previous efforts in the $4 \times 4$ case and provides a straight-forward approach to randomly generate matrices of arbitrary size.

## References

[1] M. Budden, P. Hadavas, L. Hoffman and C. Pretz, Generating valid $4 \times 4$ correlation matrices, Appl. Math. E-Notes 7(2007), 53–59.

[2] S. K. Mishra, Completing correlation matrices of arbitrary order by differential evolution method of global optimization: A Fortran program, (March 5, 2007). Available at SSRN: http://ssrn.com/abstract=968373

[3] S. K. Mishra, The nearest correlation matrix problem: Solution by differential evolution method of global optimization, (April 14, 2007). Available at SSRN: http://ssrn.com/abstract=980403

[4] P. Rousseeuw and G. Molenberghs, The shape of correlation matrices, The American Statistician, 48 (1994), 276-279.

[5] J. Stanley and M. Wang, Restrictions on the possible values of $r_{12}$, given $r_{13}$ and $r_{23}$, Educational and Psychological Measurement, 29 (1969), 579-581.