Applied Mathematics E-Notes, 7(2007), 53-59 © ISSN 1607-2510
Available free at mirror sites of http://www.math.nthu.edu.tw/∼amen/

Generating Valid 4×4 Correlation Matrices*

Mark Budden[†‡] , Paul Hadavas[†], Lorrie Hoffman[†], Chris Pretz[§]

Received 10 March 2006

Abstract

In this article, we provide an algorithm for generating valid 4×4 correlation matrices by creating bounds for the remaining three correlations that insure positive semidefiniteness once correlations between one variable and the other three are supplied. This is achieved by attending to the constraints placed on the leading principal minor determinants. Prior work in this area is restricted to 3×3 matrices or provides a means to investigate whether a 4×4 matrix is valid but does not offer a method of construction. We do offer such a method and supply a computer program to deliver a 4×4 positive semidefinite correlation matrix.

1 Introduction

The level of sophistication and complexity of statistical inquiry has placed demands on researchers to increasingly employ simulation techniques. Frequently, distributions that reflect the essence of an application are assumed to be multivariate normal with a general prior knowledge of the correlation structure. More theoretical investigations such as those in the area of Bayesian analysis rely on Markov Chain Monte Carlo methods to investigate posterior distributions. Liechty, Liechty and Muller [4] encounter a challenge when conducting their simulations due to the 'awkward manner in which r_{ij} is embedded in the likelihood' (p. 6) noting that the requirement of positive semidefiniteness is the constraint that imposes an analysis dealing with truncated distributions. This leads them to develop a procedure that generates normalizing constants. This positive semidefinite matrix complication is encountered throughout the applied literature and sometimes leads to erroneous outcomes when simulation results are aggregated while using an invalid correlation matrix.

The need to forecast demand for a group of products in order to realize savings by properly managing inventories is one such application requiring use of correlation matrices. Xu and Evers [10] offer a proof that partial pooling of customers can not lead to fewer resource requirements than complete pooling as was indicated in Tyagi and Das [9]. They further elucidate their position by showing that the error in the examples offered by these authors centered on the use of infeasible correlation matrices.

*Mathematics Subject Classifications: 65C05, 65C60.

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

[‡]The first author was supported in part by AASU Internal Grant #727188

[§]Department of Statistics, University of Wyoming, Dept. 3332, 1000 E. University Ave., Laramie, WY 82071, USA

Xu and Evers reveal that the matrices are not positive semidefinite and do so in the three variable case by revisiting boundary conditions offered by Marsaglia and Olkin [5] and via a computer program they developed to produce eigenvalues for the four variable case. Xu and Evers indicate that such an oversight is understandable because of the complex nature and interrelationships inherent in correlation structures particularly for 'non-statisticians like us ([10], p. 301)'. We contend that even statisticians and mathematicians find the problem of developing feasible correlation matrices to be challenging.

The work done by Xu and Evers is helpful in that it prevents a researcher from offering an infeasible example or conducting a meaningless simulation due to non-positive semidefinite correlation matrices. Their approach, though, is not constructive, as they say 'a correlation matrix is checked, (but) the determination of boundaries necessary for constructing feasible matrices is not performed ([10], p. 305)'. We offer such a method in the four variable case.

The paper proceeds as follows. In the next section, previous results for 3×3 correlation matrices are explained. The generation of valid 4×4 correlation matrices is detailed in section 3. Finally, in section 4, we illustrate the utility of this constructive approach with an example from Xu and Evers [10]. Additionally, in section 4, other examples of valid 4×4 correlation matrices are provided along with directions for further research.

The authors would like to thank Mortaza Jamshidian and Peter Rousseeuw for reading a preliminary draft of this paper and providing several helpful comments.

2 Review of 3×3 correlation matrices

The method described in this section for generating 3×3 correlation matrices has been noted by many sources. Although the bounds described here were known earlier, Stanley and Wang [8] gave the first proof of the bounds in 1969. Other (geometric) proofs were subsequently given by Glass and Collins [1] and Leung and Lam [3]. Olkin [6] investigated the general question of how correlations are restricted when a submatrix of a correlation matrix is fixed. More recently, Rousseeuw and Molenberghs [7] used these bounds to investigate the shape and the volume of the set of valid 3×3 correlation matrices.

Let x_1, x_2, and x_3 be random variables and let r_{ij} denote the correlation coefficient for the variables x_i and x_j where $i, j \in \{1, 2, 3\}$. A 3×3 correlation matrix is a positive semidefinite matrix of the form

$$A = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}.$$

To produce a valid correlation matrix, we begin by randomly picking the correlation coefficients $-1 < r_{12} < 1$ and $-1 < r_{13} < 1$. Here, and throughout the entire paper, we exclude the possibilities of having correlations of ± 1 since such correlations indicate the redundancy of one variable where it is a perfect linear combination of another and can

thus be eliminated from any analysis. Once these coefficients are selected, the possible range of values for r_{23} can be described by considering the determinant

$$\det A = -r_{23}^2 + (2r_{12}r_{13})r_{23} + (1 - r_{12}^2 - r_{13}^2) \geq 0.$$

This determinant is equal to 0 when

$$r_{23} = r_{12}r_{13} \pm \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}$$

and it obtains its maximum value (for r_{12} and r_{13} fixed) of

$$(1 - r_{12}^2)(1 - r_{13}^2) \leq 1$$

when $r_{23} = r_{12}r_{13}$. Thus, A is a correlation matrix exactly when r_{23} satisfies

$$r_{12}r_{13} - \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)} \ \leq \ r_{23} \ \leq \ r_{12}r_{13} + \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}. \qquad (1)$$

Again, we do not allow $r_{23} = \pm 1$ for the reason described above.

A special case to notice is when $r_{12}r_{13} = 0$. Without loss of generality, assume that $r_{12} = 0$. The range of possible values of r_{23} is given by

$$-\sqrt{1 - r_{13}^2} \ \leq \ r_{23} \ \leq \ \sqrt{1 - r_{13}^2}.$$

From this one sees that if $r_{12} = 0 = r_{13}$, then r_{23} can take on any possible correlation.

3 Algorithmic Generation of 4×4 Correlation Matrices

We now consider a 4×4 matrix of the form

$$A = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{12} & 1 & r_{23} & r_{24} \\ r_{13} & r_{23} & 1 & r_{34} \\ r_{14} & r_{24} & r_{34} & 1 \end{pmatrix}.$$

The correlations r_{12}, r_{13}, and r_{14} can be randomly picked from the interval $(-1, 1)$. To find suitable ranges for the other correlations, we use the fact that a symmetric matrix is positive semidefinite if and only if all of its symmetric submatrices (including itself) have a nonnegative determinant (see. [2]). In other words, A is a correlation matrix if and only if $\det A \geq 0$ and every matrix of the form

$$A_{ijk} = \begin{pmatrix} 1 & r_{ij} & r_{ik} \\ r_{ij} & 1 & r_{jk} \\ r_{ik} & r_{jk} & 1 \end{pmatrix}$$

is a correlation matrix for $i, j, k \in \{1, 2, 3, 4\}$ (no two of i, j, and k are equal).

For each remaining correlation $r_{jk} \in \{r_{23}, r_{24}, r_{34}\}$, there are three limitations on the range of possible values. With the symmetry of the matrix, we use the fact that $r_{ij} = r_{ji}$ for all i, j to simplify the notation used in the bounds for r_{23}, r_{24}, r_{34}. First, the bounds are restricted by $\det(A_{1jk}) \geq 0$. From the previous section, we see that

$$L_{jk}^{(1)} \leq r_{jk} \leq U_{jk}^{(1)},$$

where

$$L_{jk}^{(1)} := r_{1j}r_{1k} - \sqrt{(1-r_{1j}^2)(1-r_{1k}^2)} \text{ and } U_{jk}^{(1)} := r_{1j}r_{1k} + \sqrt{(1-r_{1j}^2)(1-r_{1k}^2)}.$$

Next, we consider the restrictions imposed on the range of r_{jk} by considering $\det(A_{ijk}) \geq 0$. Again, from the previous section, we have

$$L_{jk}^{(2)} \leq r_{jk} \leq U_{jk}^{(2)},$$

where

$$L_{jk}^{(2)} := r_{ij}r_{ik} - \sqrt{(1-r_{ij}^2)(1-r_{ik}^2)} \text{ and } U_{jk}^{(2)} := r_{ij}r_{ik} + \sqrt{(1-r_{ij}^2)(1-r_{ik}^2)}.$$

Finally, we look at the restrictions imposed by $\det(A) \geq 0$. Noting that

$$\begin{aligned}\det A = {} & 1 - r_{12}^2 - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2 + r_{12}^2 r_{34}^2 + r_{13}^2 r_{24}^2 + r_{14}^2 r_{23}^2 \\ & + 2r_{23}r_{24}r_{34} + 2r_{12}r_{13}r_{23} + 2r_{12}r_{14}r_{24} + 2r_{13}r_{14}r_{34} \\ & - 2r_{12}r_{14}r_{23}r_{34} - 2r_{13}r_{14}r_{23}r_{24} - 2r_{12}r_{13}r_{24}r_{34},\end{aligned}$$

can be factored as a quadratic with respect to each $r_{jk} \in \{r_{23}, r_{24}, r_{34}\}$, we find the bounds

$$L_{jk}^{(3)} \leq r_{jk} \leq U_{jk}^{(3)},$$

where

$$L_{jk}^{(3)} := \frac{r_{ji}r_{ki} + r_{1j}r_{1k} - r_{1j}r_{1i}r_{ki} - r_{1k}r_{1i}r_{ji} - \sqrt{\det(A_{1ji}) \cdot \det(A_{1ki})}}{1 - r_{1i}^2}$$

and

$$U_{jk}^{(3)} := \frac{r_{ji}r_{ki} + r_{1j}r_{1k} - r_{1j}r_{1i}r_{ki} - r_{1k}r_{1i}r_{ji} + \sqrt{\det(A_{1ji}) \cdot \det(A_{1ki})}}{1 - r_{1i}^2}$$

Now we describe the algorithm using the bounds from above. Once r_{12}, r_{13}, and r_{14} are fixed, we choose r_{23} from the range

$$\max\{L_{23}^{(1)}, L_{23}^{(2)}, L_{23}^{(3)}\} \leq r_{23} \leq \min\{U_{23}^{(1)}, U_{23}^{(2)}, U_{23}^{(3)}\},$$

where $L_{23}^{(2)}$ and $L_{23}^{(3)}$ are minimized and $U_{23}^{(2)}$ and $U_{23}^{(3)}$ are maximized over the regions

$$L_{24}^{(1)} \leq r_{24} \leq U_{24}^{(1)} \text{ and } L_{34}^{(1)} \leq r_{34} \leq U_{34}^{(1)}.$$

This guarantees the existence of some valid correlation matrix with the chosen values r_{12}, r_{13}, r_{14}, and r_{23} since all of the conditions of positive semidefinite are met.

A similar process is repeated for r_{24}. We choose r_{24} from the range

$$\max\{L_{24}^{(1)}, L_{24}^{(2)}, L_{24}^{(3)}\} \leq r_{24} \leq \min\{U_{24}^{(1)}, U_{24}^{(2)}, U_{24}^{(3)}\},$$

where $L_{24}^{(2)}$ and $L_{24}^{(3)}$ are minimized and $U_{24}^{(2)}$ and $U_{24}^{(3)}$ are maximized over the region

$$L_{34}^{(1)} \leq r_{34} \leq U_{34}^{(1)}.$$

Finally, there is only one correlation left to bound. We choose r_{34} from the range

$$\max\{L_{34}^{(1)}, L_{34}^{(2)}, L_{34}^{(3)}\} \leq r_{34} \leq \min\{U_{34}^{(1)}, U_{34}^{(2)}, U_{34}^{(3)}\}.$$

The constructive approach of this algorithm guarantees that any matrix formed in this way will be a valid correlation matrix. Furthermore, the ranges of possible values of each of the correlations r_{23}, r_{24}, and r_{34}, are the largest possible ranges so that there will exist some valid correlation matrix with any chosen value. In other words, any value chosen outside of the determined range will fail one of the necessary conditions for being valid. Thus, our algorithm provides both necessary and sufficient conditions for producing valid 4×4 correlation matrices.

4 Applications

Having a method to create valid correlation matrices, we can now discuss the application presented by the resource allocation problem and present other examples that illustrate the difficulty of being able to guess proper correlation values. In offering a proof related to resource allocation, Xu and Evers [10] refute a prior claim by Tyagi and Das [9]. Xu and Evers explain that the error in logic was fostered by errors in the examples that inadvertently used invalid correlation matrices - ones that failed to be positive semidefinite. In the three variable case, Xu and Evers rely on boundary conditions offered by Marsaglia and Olkin [5] to establish that the matrix used by Tyagi and Das was not positive semidefinite. In the four variable case, they employ their own computer program developed to produce eigenvalues.

The previous section contains a constructive algorithm that generates valid 4×4 correlation matrices. We have developed a computer program called "validcor" (2005) to accomplish this task. Our program delivers the remaining correlations after specifying those relating one variable to the other three. The program can be found in the general archive at Carnegie Mellon University's Statlab.

http://lib.stat.cmu.edu

To illustrate the utility of this constructive approach we use Tyagi and Das' matrix

$$\begin{pmatrix} 1 & -.5 & .5 & -.5 \\ -.5 & 1 & -.5 & .5 \\ .5 & -.5 & 1 & .5 \\ -.5 & .5 & .5 & 1 \end{pmatrix}$$

and show that it would not be generated by our program. For this example we set $r_{12} = -.5$, $r_{13} = .5$, and $r_{14} = -.5$. Feasible bounds found by our program for r_{23} are $(-1, .5]$. We chose $r_{23} = -.5$. Then, feasible bounds found for r_{24} are $[-.5, 1)$. We chose $r_{24} = .5$. Finally, feasible bounds found for r_{34} are $(-1, .3333]$. The r_{34} value selected for the Tyagi and Das matrix was .5 and is outside the range that assures a positive semidefinite matrix.

In addition to the use of the program showing how certain matrices are not valid correlation matrices, we provide the table below where each of the five rows correspond to generations of valid correlation matrices. The first three columns give the input parameters.The following columns give the bounds for the remaining correlations and the choices for r_{23} and r_{24} that led to the successive feasible bounds found by the computer program. Here, r^*_{23} and r^*_{24} represent those choices, respectively. No choice for r_{34} is given as any value within the bounds will complete a valid matrix.

Table 1: Examples of Valid Correlation Matrix Generations

r_{12}	r_{13}	r_{14}	r_{23} bounds	r^*_{23}	r_{24} bounds	r^*_{24}	r_{34} bounds
.777	.39	.472	$(-.2766, .8827)$	.88	$(-.1852, .9217)$	.92	$(.9835, .9957)$
.981	.397	.961	$(.2114, .5675)$	.567	$(.8895, .9964)$	.996	$(.6304, .6351)$
.027	.495	.986	$(-.8552, .8819)$	.881	$(-.1381, .1933)$	$-.138$	$(.3441, .3462)$
.801	.913	$-.6$	$(.4871, .9755)$	.975	$(-.9553, -.0017)$	$-.002$	$(-.2232, -.2216)$
.04	.008	.207	$(-.9988, .9995)$	.754	$(-.9692, .9858)$	$-.96$	$(-.8716, .641)$

One interesting outcome to note from the table are the final bounds for the r_{34} value. When both selections for r_{23} and r_{24} occur near their respective bounds, the interval for r_{34} is severely limited. However, in the last example, a selection of r_{23} not close to either bound allows r_{34} to take on a wide range of possible values. This phenomenon along with its potential application to finding the hypervolume of possible 4×4 correlation matrices in 6-space will be explored in future research. Another direction will include the generalization of the algorithm to $n \times n$ correlation matrices where n is greater than 4. Such a generalization is not trivial as the complexity of the bounds seems to grow exponentially with the size of the matrix. Most likely, the above algorithm will require a significant simplification before it can be extended to correlation matrices of arbitrary size. We save such an analysis for future study.

References

[1] G. Glass and J. Collins, Geometric proof of the restriction on the possible values of r_{xy} when r_{xz} and r_{yx} are fixed, Educational and Psychological Measurement, 30(1970), 37-39.

[2] R. Harris, A Primer of Multivariate Statistics, Lawrence Erlbaum Associates, Mahwah, NJ, 2001.

[3] C.-K. Leung and K. Lam, A Note on the geometric representation of the correlation coefficients, The American Statistician, 29(1975), 128-130.

[4] J.C. Liechty, M.W. Liechty, P. and Muller, Bayesian correlation estimation, Biometrika, 9(2004), 1-14.

[5] G. Marsaglia and I. Olkin, Generating correlation-matrices, SIAM Journal on Scientific and Statistical Computing, 5(2)(1984), 470-475.

[6] I. Olkin, Range restrictions for product-moment correlation matrices, Psychometrika, 46(1981), 469-472.

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

[8] 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.

[9] R. Tyagi and C. Das, Grouping customers for better allocation of resources to serve correlated demands, Computers and Operations Research, 26(1999), 1041-1058.

[10] K. Xu and P. Evers, Managing single echelon inventories through demand aggregation and the feasibility of a correlation matrix, Computers and Operations Research, 30(2003), 297-308.