GENERATING CORRELATION MATRICES

BY

GEORGE MARSAGLIA and INGRAM OLKIN

TECHNICAL REPORT NO. 186
FEBRUARY 1983

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT

MCS 81-04262

Ingram Olkin, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

GENERATING CORRELATION MATRICES

by

George Marsaglia and Ingram Olkin
Washington State University and Stanford University

ABSTRACT

This paper describes a variety of methods for generating random correlation matrices, with emphasis on choice of random variables and distributions so as to provide matrices with given structure, expected values or eigenvalues.

Keywords: *Random Correlation Matrices, Random Numbers, Monte Carlo, Simulation*

1. INTRODUCTION.

A correlation matrix is a symmetric, positive semi-definite matrix with 1's on the diagonal. Numerous papers have been devoted, in whole or in part, to the problem of generating random correlation matrices, but usually with a particular application in mind. Papers concerned with correlation matrices having a particular structure are those of Tucker, Hoppman and Linn (1969), Herzberg (1969), Dempster, Schatzoff and Wermuth (1977) and Ryan (1978), while Bendel and Afifi (1977), Chalmers (1975), Bendel and Mickey (1978), Johnson and Welch (1980) are concerned with eigenvalues of the generated matrices.

There seems to be need for a discussion of general methods that may be used to generate random correlation matrices which meet various requirements such as structural, distribution of elements, expected values or eigenvalues, with emphasis on possible choices of the random elements at each stage to achieve the desired objective. The following three sections describe methods classified in three ways: random correlation matrices with given expected values; in the form TT'; with given eigenvalues.

2. Random Correlation Matrices with Given Mean.

Let C be a given correlation matrix and let $X = (x_{ij})$ be a random symmetric matrix with zeros on the diagonal and with means $E[x_{ij}] = 0$. Then $C + X$ will be a random correlation matrix with expected value C if, and only if, the eigenvalues of $C + X$ are non-negative. Viewing X as a perturbation of C, we may use well-known results: adding the symmetric matrix X to C cannot change the eigenvalues of C by more than $\|X\|_2$, the 2-norm of X. See, e.g., Stewart (1968). Since X is symmetric, its 2-norm is its spectral radius, which is bounded by both $\|X\|_E = [\Sigma x_{ij}^2]^{\frac{1}{2}}$ and $\|X\|_1 = \|X\|_\infty = \max\limits_i \sum\limits_j |x_{ij}|$. Thus if any one of $\|X\|_2$, $\|X\|_E$ or $\|X\|_1 = \|X\|_\infty$ is less than λ, the least eigenvalue of C, then $C + X$ will be positive definite.

This provides a variety of methods for generating random correlation matries $C + X$ with expected value the given matrix C having least eigenvalue λ:

1. Let $A = (a_{ij})$ be a symmetric matrix with zeros on the diagonal such that $\|A\|_1 = \max\limits_i \sum\limits_j |a_{ij}| < \lambda$. For $j > i$, generate x_{ij} such that the marginal distribution of each x_{ij} is in the interval $|x_{ij}| < a_{ij}$ and such that $E[x_{ij}] = 0$. Then, with $x_{ii} = 0$ and $x_{ji} = x_{ij}$ for $i > j$, $R = C + X$ will be a random correlation matrix with expected value C. A simple way to do this is with x_{ij} independent uniform in $|x| < |a_{ij}|$, $i < j$.

2. For $i < j$ generate $x_{12}, x_{13}, \ldots, x_{n-1,n}$ with a radially symmetric distribution in (or on) the unit $n(n-1)/2$-sphere. Then $R = C + 2^{-\frac{1}{2}}\lambda X$ will be a random correlation matrix with expected value C.

3. Using the method described in Section 4, generate a random correlation matrix R with eigenvalues in the interval $(1-\lambda, 1+\lambda)$, where λ is the least eigenvalue of C. Then $C+R-I$ is a random correlation matrix with expected value C.

3. Random Correlation Matrices of the Form TT'.

In the method of the previous section the matrix C was given, and a random matrix X was then chosen so that $C + X$ was positive definite. An easier way to guarantee definitiveness (but with less control on the distributions) is to form TT' from a random $n \times m$ matrix T. In order that the result be a correlation matrix, the rows of T must have length one, and thus the problem may be put in geometric terms:

TT' is a random correlation matrix if, and only if, the rows of T are random points on the unit m-sphere.

This leads to a variety of methods. Probably the easiest is one that merely generates a matrix T of independent uniform variates, and then forms TT' after normalizing by dividing each row by its root-mean-square. A faster method requires about half as many variates by starting with T lower triangular. There are two easy ways to make an initial row $x_1, \ldots, x_m$ of T a point on the unit m-sphere by normalizing: root-mean-square,

$$x_i \leftarrow \pm x_i / (\Sigma x_j^2)^{\frac{1}{2}}$$

and root-absolute-mean,

$$x_i \leftarrow \pm(|x_i|/\Sigma|x_j|)^{\frac{1}{2}},$$

with the ± optional. For two initial rows $x_1,\ldots,x_m$ and $y_1,\ldots,y_m$ the two kinds of normalization lead to marginal densities in TT' for a ratio or a sum of ratios:

(1) $(\Sigma\pm x_i y_i)/(\Sigma x_j^2 \Sigma x_j^2)^{\frac{1}{2}}$

and

(2) $\Sigma\pm[|x_i y_i|/(\Sigma|x_j|\Sigma|y_j|)]^{\frac{1}{2}}]$.

If the initial elements of T are uniform or exponential, the distribution of (2) is more tractable than that of (1). If the initial elements are standard normal variates then the distribution of (1) may be given explicitly; we will do so below. The central limit theorem is more readily applied to the sum in (2), but since (1) approaches the ratio of normal variates, both cases lead to the elements of TT' being approximately normal for m large. Choosing a random ± when normalizing the rows will center the densities at the origin.

Monte Carlo experiments to compare methods for generating the initial elements of T and normalizing procedures should be worthwhile. In order to have exact distributions against which Monte Carlo results may be compared, we will give explicit densities for the case that the matrix T is initially chosen with non-zero elements independent standard normal, then its rows projected onto the unit m-sphere by the root-mean-square normalization (1). This will include triangular or any other n×m T. For example, if an initial T has this form:

$$\begin{pmatrix} a & 0 & 0 & b & c & 0 \\ r & s & t & u & 0 & 0 \\ 0 & v & 0 & w & 0 & x \end{pmatrix}$$

with non-zero elements standard normal, and if each row is then made a point on the unit 6-sphere by dividing by its root-mean-square, what are the marginal densities of the elements of TT'?

THEOREM. *If*

$$Z = \sum_1^k X_iY_i/[\sum_1^m X_i^2 \sum_1^n Y_i^2]^{\frac{1}{2}} \qquad 0 < k \le m \le n,$$

with the X's and Y's independent standard normal random variables, then Z^2 is distributed as the product of beta variates:

$$Z^2 \sim \beta_{\frac{1}{2},\frac{1}{2}n-\frac{1}{2}}\beta_{\frac{1}{2}k,\frac{1}{2}m-\frac{1}{2}k}$$

and the density of Z is, for $-1 < z < 1$:

$$\frac{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2}m)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2}n-\frac{1}{2})\Gamma(\frac{1}{2}k)\Gamma(\frac{1}{2}m-\frac{1}{2}k)} \int_{z^2}^{1} (y-z^2)^{\frac{1}{2}(n-3)} y^{\frac{1}{2}k-\frac{1}{2}m} (1-y)^{\frac{1}{2}m-\frac{1}{2}k-1} dy.$$

(When $k = m$, Z^2 is the single beta variate $\beta_{\frac{1}{2},\frac{1}{2}n-\frac{1}{2}}$, with corresponding density for Z.)

Proof: We may view Z as an inner product, $\alpha\beta'$, with α and β independent points on the (m+n-k)-sphere. For example, when k = 2, m = 3, n = 4 and

$$\alpha = (X_1,X_2,X_3,0,0)/[\sum_1^3 X_i^2]^{\frac{1}{2}}, \quad \beta = (Y_1,Y_2,0,Y_3,Y_4)/[\sum_1^4 Y_i^2]^{\frac{1}{2}},$$

then the random point α is the projection of $(X_1,X_2,X_3,0,0)$ onto the 5-sphere, so that α is independent of ΣX_i^2 and, similarly, β is independent of ΣY_i^2. Thus, in general, $Z = \alpha\beta'$ is independent of the product $\Sigma X_i^2 \Sigma Y_i^2$. The latter product is distributed as the product of independent gamma variates:

$$\sum_1^m X_i^2 \sum_1^n Y_i^2 \sim 4\gamma_{\frac{1}{2}m}\gamma_{\frac{1}{2}n}.$$

It follows that, with $V = \sum_1^k X_i Y_i$,

$$E[Z^{2r}]E[\gamma^r_{\frac{1}{2}m}]E[\gamma^r_{\frac{1}{2}n}] = E[Z^{2r}\gamma^r_{\frac{1}{2}m}\gamma^r_{\frac{1}{2}n}] = E[V^{2r}].$$

Thus $E[Z^{2r}] = E[V^{2r}]/E[\gamma^r_{\frac{1}{2}m}\gamma^r_{\frac{1}{2}n}]$ and, using the moments of V^2 from the Lemma below,

$$E[Z^{2r}] = \frac{\Gamma(\frac{1}{2}k+r)\Gamma(\frac{1}{2}+r)}{\Gamma(\frac{1}{2}k)\Gamma(\frac{1}{2})} \frac{\Gamma(\frac{1}{2}n)\Gamma(\frac{1}{2}m)}{\Gamma(\frac{1}{2}n+r)\Gamma(\frac{1}{2}m+r)} .$$

If $\beta_{a,b}$ is a random variable having a beta distribution with parameters a and b, then $E\beta^r_{a,b} = B(a+r,\ b)/B(a,b)$. Since Z^2 is a bounded random variable, its distribution is determined by its moments, namely, $EZ^{2r} = E\beta^r_{\frac{1}{2},\,(n-1)/2}\ \beta^r_{k/2,(m-k)/2}$. The density of Z^2 is obtained from the density of a product and of a square root. ‖

LEMMA. *Let $V = \Sigma_1^k X_i Y_i$, with the X's and Y's independent standard normal. Then*

$$E[V^{2r}] = 4\Gamma(\tfrac{1}{2}k+r)\Gamma(\tfrac{1}{2}+r)/[\Gamma(\tfrac{1}{2}k)\Gamma(\tfrac{1}{2})].$$

Proof: Because of the radial symmetry of the distributions of $(X_1, X_2,\ldots,X_k)$ and $(Y_1,Y_2,\ldots,Y_k)$, we may assume that $(X_1,X_2,\ldots,X_k)$ has the form $(W,0,\ldots,0)$, with W distributed as $(X_1^2+\ldots+X_k^2)^{\frac{1}{2}}$, so that $W^2/2$ is distributed as a gamma variate with parameter $k/2$, i.e. $W^2 \sim 2\gamma_{k/2}$. Then $V^2 \sim 4\gamma_{\frac{1}{2}k}\gamma_{\frac{1}{2}}$, where $\gamma_{k/2}$ and $\gamma_{\frac{1}{2}}$ are independent. Then

$$E[V^{2r}] = 4E[\gamma^r_{\frac{1}{2}k}]E[\gamma^r_{\frac{1}{2}}] = 4\Gamma(\tfrac{1}{2}k+r)\Gamma(\tfrac{1}{2}+r)/[\Gamma(\tfrac{1}{2}k)\Gamma(\tfrac{1}{2})]. \quad \|$$

Using root-mean-square normalization on a set of k independent normal variates produces a point on the surface of the unit k-sphere and the resulting spherically symmetric distribution seems the most desirable for many applications; it is one of the few methods for which the resulting distributions of elements in the correlation matrix can be given explicitly, as above. If a fast method for generating normal variates is not available, there are other efficient methods for generating uniform points on a k-sphere—see, e.g., Marsaglia (1972,1980).

4. Generating a Correlation Matrix with Given Eigenvalues.

Methods for generating a correlation matrix whose eigenvalues are close to a given set may be based on classical perturbation theory: If the ordered eigenvalues of C are $\lambda_1 \le \cdots \le \lambda_n$ and those of C + X are $\mu_1 \le \cdots \le \mu_n$ then, Hoffman and Weilandt (1953):

$$\Sigma(\lambda_i - \mu_i)^2 \le \|X\|_E^2.$$

Thus choosing a random symmetric matrix X with zeros on the diagonal and $\|X\|_E$ small will ensure that C + X is a correlation matrix with eigenvalues close to C. A disadvantage of this method is that $\|X\|_E$ may be so small that all the random matrices C + X will look like C.

Another approach may be used: choose $D = \text{diag}\{d_1,\ldots,d_n\}$ so that $\Sigma(\lambda_i - d_i)^2 < \varepsilon$; then choose a random orthogonal matrix P so that PDP' is a correlation matrix. Putting $\varepsilon = 0$ would then handle the case where the eigenvalues must be exactly a given set.

The trace of an n×n correlation matrix must be n, so that we may put the problem in general form: given an n×n positive semi-definite matrix A whose trace is n, choose a random orthogonal matrix P so that PAP' is a correlatiom matrix, i.e., has 1's on the diagonal. It is elementary to prove by induction that there are such P's: Choose any point α on the n-sphere $\alpha\alpha' = 1$ so that it also satisfies $\alpha A\alpha' = 1$, (the Rayleigh quotient guarantees the range of $\alpha A\alpha'$ will include 1 if the trace of A is n) and any orthogonal matrix $P = \binom{\alpha}{B}$ whose first row is α. Then

$$\binom{\alpha}{B}A(\alpha' B') \;=\; \begin{pmatrix}1 & \alpha AB' \\ BA\alpha' & BAB'\end{pmatrix}$$

and BAB' is an $(n-1)\times(n-1)$ positive definite matrix with trace equal to $(n-1)$. This is backward induction so that we need only consider the case $n = 2$, for which the solution is explicit.

The most difficult problem in implementing this algorithm is in choosing a random point α from the n-sphere so that $\alpha A\alpha' = 1$. There seems to be no easy, explicit way to do this, but if one has a way, the following scheme may be used to generate all of the rows of the required orthogonal matrix P:

Start with the symmetric idempotent matrix $E = I$.

Choose α_1 from the row space of E, subject to $\alpha_1\alpha_1' = \alpha_1 A\alpha_1' = 1$, and replace E by $E - \alpha_1'\alpha_1$.

Choose α_2 from the row space of E, subject to $\alpha_2\alpha_2' = \alpha_2 A\alpha_2' = 1$, and replace E by $E - \alpha_2'\alpha_2$.

. . .
. . .
. . .

Choose α_n from the row space of E, subject to $\alpha_n\alpha_n' = \alpha_n A\alpha_n' = 1$, and replace E by $E - \alpha_n'\alpha_n$.

The resulting $\alpha_1,\alpha_2,\ldots,\alpha_n$ will be orthonormal with $\alpha_i A\alpha_i' = 1$, so that the matrix P with rows $\alpha_1,\ldots,\alpha_n$ will be orthogonal and PAP' will have unit diagonal elements. The α's will be orthogonal because α_2 is in the row space of $I - \alpha_1'\alpha_1$, α_3 is in the row space of $I - \alpha_1'\alpha_1 - \alpha_2'\alpha_2$, and so on.

This provides an easy-to-follow algorithm, but for one aspect: How do we "choose α from the row space of E, subject to $\alpha\alpha' = \alpha A\alpha' = 1$?" Assume $A = D$, a diagonal matrix of the given eigenvalues. Spherical symmetry in the choice of α's will make the result invariant under choice of the initial symmetric matrix A, provided it has the given eigenvalues. The set $\alpha\alpha' = \alpha D\alpha' = 1$ is a subset of the surface of the unit n-sphere, a pair of "spherical ellipses"– one the reflection of the other through the origin. A random slice through the sphere may not hit the two "ellipses," but if it does, it provides a nice way, by choosing one of the resulting four points of intersection.

This then suggests a rejection procedure for sampling α from the row space of a matrix E, subject to $\alpha\alpha' = \alpha D\alpha' = 1$: Choose two points ξ and η from the row space of E, each in the form ζE, with $\zeta = (z_1,\ldots,z_n)$ having independent normal coordinates. If the plane determined by ξ, η and the origin cuts the set $\alpha\alpha' = \alpha D\alpha' = 1$, take one of the four points of intersection and follow the algorithm. If the plane doesn't cut the set, choose a new plane. The plane determined by ξ and η may be represented as $r\xi + \eta$, rather than the conventional $r_1\xi + r_2\eta$, in view of the fact that we project onto the unit n-sphere. The condition that the plane cut the set $\alpha\alpha' = \alpha D\alpha' = 1$ now becomes: $b^2 < ac$, where $a = \xi(I-D)\xi'$, $b = \xi(I-D)\eta'$, $c = \eta(I-D)\eta'$.

These details are now incorporated into an explicit algorithm.

ALGORITHM

Given the diagonal matrix $D = \mathrm{diag}\{d_1,\ldots,d_n\}$ *with* $d_i \geq 0$ *and* $\Sigma d_i = n$, *this algorithm produces a random orthogonal matrix P such that PDP' is a correlation matrix (1's on the diagonal).*

1. [Initial state] Start with the $n \times n$ *matrix* E *and the index* k: $E \leftarrow I$, $k \leftarrow 1$.

2. Generate a random vector $\xi = (x_1,\ldots,x_n)$ *in the row space of* E. *[Let* $\xi = (z_1,\ldots,z_n)E$, *with the z's independent normal.] Compute* $a = \Sigma(1-d_i)x_i^2$.

3. Generate another random vector $\eta = (y_1,\ldots,y_n)$ *in the row space of* E.

4. Compute $b = \Sigma(1-d_i)x_iy_i$ *and* $c = \Sigma(1-d_i)y_i^2$. *If* $d^2 = b^2 - ac \leq 0$ *go to step 3.*

5. Put $r = (b \pm d)/a$, *with the ± chosen at random. Then the vector* $\zeta = r\xi + \eta$, *when normalized with a random sign:* $\zeta \leftarrow \pm\zeta/(\zeta\zeta')^{\frac{1}{2}}$ *is a random choice for the* k^{th} *row of* P. *Replace:* $E \leftarrow E - \zeta'\zeta$, *increment* k *and go to step 2, unless* $k > n$, *in which case the* n *rows of* P *have been generated, and, as a check,* E *is now* $I - P'P$ *and should be* 0.

The above algorithm may be compared to two others: Bendel and Mickey (1978) choose a random orthogonal matrix P, form PDP', then use 2×2 rotations (never reflections) to make the diagonal elements unity, one at a time. Chalmers (1975) provides an algorithm that chooses the rows of P sequentially to produce the required form, with rejection sampling from an n-cube at each stage. Our approach is similar, but with the intent to describe more fully the available choices for each row, providing possibly simpler algorithms or better control on the resulting distributions or expectations. Choosing the available elements from spherically symmetric distributions makes the resulting PAP' invariant under choice of the initial A with given eigenvalues.

REFERENCES

Bendel, R.B. and Afifi, A. A. (1977), "Comparison of Stopping Rules in Forward Stepwise Regression," *J. Amer. Stat. Assoc., 72*, 46-53.

Bendel, R. B. and Mickey, M. R. (1978), "Population Correlation Matrices for Sampling Experiments," *Commun. Statist-Simul. Comput.,* B7(2), 163-182.

Chalmers,C. P. (1975), "Generation of Correlation Matrices with Given Eigen-Structure," *J. Statist. Comput. Simul.,* 4, 133-139.

Dempster,A. P., Schatzoff, Martin and Wermuth,N. (1977), "A Simulation Study of Alternatives to Least Squares," *J. Amer. Stat. Assoc.,* 72, 77-90.

Herzberg, P. A. (1969), "The Parameter of Cross-Validation," *Psychometrika Monograph Supplement No. 16,* 34 part 2.

Hoffman A. J. and Weilandt, H. W. (1953), "The Variation of the Spectrum of a Normal Matrix," *Duke Math. Journal,* 20, 37-39

Johnson, D. G. and Welch, W. J. (1980), "The Generation of Pseudo-Random Correlation Matrices," *J. Statist. Comput. Simul.*, **11**, 55-69.

Marsaglia,George (1972), "Choosing a Point From the Surface of a Sphere," *Annals Math. Statist.*, **43**, 645-646

Marsaglia, George (1980), "Generating a Normal Sample with a Given Mean and Variance," *J. Statist. Comput. Simul.*, **11**, 71-73.

Ryan, T. P. (1980), "A New Method of Generating Correlation Matrices," *J. Statist. Comput. Simul.*, **11**, 79-85.

Stewart, G. W. (1973), *Intoduction to Matrix Computation*. New York: Academic Press

Tucker, L. R., Koopman, R. F. and Linn, R. F. (1969), "Evaluation of Factor Analytic Research Procedures by Means of Simulated Correlation Matrices," *Psychometrika*, 34, 421-459.