Correlation Matrices in Cosine Space

By

Alexandria Ree Hadd

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

December, 2016

Nashville, Tennessee

Approved:

Joseph L. Rodgers, Ph.D.

Kristopher J. Preacher, Ph.D.

Andrew J. Tomarken, Ph.D.

To Kirsten with timeless love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

In the introduction of their article, *The Shape of Correlation Matrices*, Rousseeuw and Molenberghs (1994, p.276) asserted that "the correlation coefficient is one of the most frequently used statistical tools." Correlations and correlation matrices are foundational concepts in all disciplines that use statistical analysis, including psychology, genetics, and finance. Diversity in the potential applications of the correlation leads to similar diversity in potential interpretations; the correlation coefficient can be variously interpreted as a mean, a ratio, a cross product, and through several trigonometric functions (Rodgers & Nicewander, 1988). Of particular interest is one of the trigonometric interpretations. Specifically, in person space – where N individuals define the axes of an N-dimensional space, and centered or standardized scores on variable vectors are plotted on these axes – the correlation between two variables X_1 and X_2 can be expressed as

$$r_{12} = \cos \boldsymbol{\theta_{12}} \tag{1}$$

where $\boldsymbol{\theta_{12}}$ is the angle between the centered/standardized variable vectors X_1 and X_2. (Box, 1978, documents Fisher's reliance on this person space in the development of his statistical insights.)

Rousseeuw and Molenberghs (1994) expanded on the geometric literature for the correlation coefficient. The authors demonstrated the three-dimensional closed surface that summarizes the space of true 3x3 correlation matrices – the 3x3 correlation space. This space provides insight for understanding individual correlation matrices, as well as the relationships among the correlations within a correlation matrix. Using the 3x3 correlation space as a starting point, in this thesis I accomplish three things. In Chapter 1, I show how their space can be usefully re-portrayed using the cosine formulation of the correlation coefficient – into the sonamed 3x3 cosine space. This reportrayal carries forward the strengths of the 3x3 correlation space for understanding individual correlation matrices, but also provides insight into the correlation space itself. In Chapter 2, I discuss how the 3x3 cosine space can provide insight into the shape of cosine (and correlation) spaces in higher dimensions. I give particular attention to the cosine space of 4x4 correlation matrices as a case study before considering properties of higher dimension cosine (and correlation) spaces. Third, I give a practical demonstration of the utility of 3x3 cosine space in generating random correlation matrices with a relatively high frequency of extreme correlations – that is, correlation matrices near the boundary of the correlation/cosine space.

Throughout the thesis, I refer to the space originally envisioned by Rousseeuw and Molenberghs as the correlation space, and the transformation of this space by the cosine function as the cosine space. Technically, both spaces are "correlation spaces," as correlations are cosines of angles. I could instead refer to the so-called correlation space as the [-1,1] space, the linear axis space, or the R&M space (or any combination of such descriptors of the space), and I could refer to the so-called cosine space as the angle space, the [0,180] space, the nonlinear axis space, or the transformed space. For ease, I simply refer to these spaces as the correlation spaces and cosine spaces respectively for correlation matrices of given dimension.

CHAPTER 1

THE 3x3 COSINE SPACE OF CORRELATION MATRICES

The Space of Correlation Matrices

Correlations among a set of variables (e.g., $X_1, X_2, \ldots X_p$) are typically summarized in a correlation matrix $\boldsymbol{R}$. Let $\boldsymbol{R}$ be a square matrix of order p; the rows and columns of $\boldsymbol{R}$ indicate the variables being correlated, and the entries r_{ij} in $\boldsymbol{R}$ are correlation coefficients between pairs of variables X_i and X_j with three necessary properties:

(i) $r_{ij} = r_{ji}$ (i.e., $\boldsymbol{R}$ is symmetric)

(ii) $r_{ij} = 1$ if $i = j$ (i.e., the diagonals of $\boldsymbol{R}$ are 1)

(iii) $-1 \leq r_{ij} \leq 1$ if $i \neq j$ (i.e., the off-diagonals of $\boldsymbol{R}$ are correlation coefficients)

These three properties are necessary for $\boldsymbol{R}$ and are simple to check, but they are not sufficient. To be a true correlation matrix, $\boldsymbol{R}$ must also be positive semidefinite (PSD). As such, we add a fourth property to $\boldsymbol{R}$ that ensures it is a PSD matrix:

(iv) $\lambda_1, \lambda_2, \ldots \lambda_p \geq 0$ where $\lambda_i, i = 1, 2, \ldots, p$ are the eigenvalues of $\boldsymbol{R}$.

This fourth property is equivalently satisfied by ensuring that the determinants of $\boldsymbol{R}$ and all principle minor submatrices of $\boldsymbol{R}$ are nonnegative. (Matrices with all positive λ_p and positive determinants for the principle minor submatrices are said to be positive definite, or PD). Matrices that satisfy the first three properties are called pseudo-correlation matrices ($\boldsymbol{R}^*$), with the subset of $\boldsymbol{R}^*$ also satisfying the fourth property being true correlation matrices ($\boldsymbol{R}$). Non-PSD pseudo-correlation matrices cannot occur under typical data circumstances; they may occur through pairwise deletion of variables or use of tetrachoric or polychoric correlations (Knol & Berger, 1991), but matrices constructed from complete, quantitative data must be true correlation matrices. (Note that Rousseeuw and Molenberghs frame the PSD condition in terms of determinants of $\boldsymbol{R}^*$. This framing suffices for the 3x3 case, and aids in providing the equation for the boundary of the correlation space, but I define the PSD condition with eigenvalues of $\boldsymbol{R}^*$ to facilitate expansion of the PSD condition to higher orders of p, where calculation of determinants of the principle minor submatrices of $\boldsymbol{R}^*$ becomes increasingly cumbersome.)

Because of the symmetry and unit diagonal of $\boldsymbol{R}^*$, only the upper-triangular portion of the matrix need be represented. A shorthand for $\boldsymbol{R}^*$ is obtained by half-vectorization of $\boldsymbol{R}^*$ **,** that is, concatenating the rows of the upper-triangular portion of $\boldsymbol{R}^*$ to form an ordered $n = (p(p - 1))/2$-tuple, $\boldsymbol{r}^*$, which uniquely identifies the original pseudo-correlation matrix (also referred to as the vecp or vech operator; Browne & Shapiro, 1986). Note that p is the dimension of $\boldsymbol{R}^*$and n is the dimension of the set of $\boldsymbol{R}^*$ (i.e., $\boldsymbol{r}^* \in [-1,1]^n$). For example, the matrix

$$\boldsymbol{R}^* = \begin{bmatrix} 1 & -.23 & .04 & -.14 \\ -.23 & 1 & .35 & .05 \\ .04 & .35 & 1 & -.06 \\ -.14 & .05 & -.06 & 1 \end{bmatrix}$$

can be uniquely identified by the ordered sextuple $\boldsymbol{r}^* = (-.23,\ .04,\ -.14,\ .35,\ .05,\ -.06) \in [-1,1]^6$. This shorthand will prove useful in depicting the subset of $\boldsymbol{R}$ within the set of all $\boldsymbol{R}^*$ in three dimensions.

The 3x3 Correlation Space

The correlations among three variables, X_1, X_2, and X_3, produce a correlation matrix $\boldsymbol{R}$ of order $n = p = 3$, which can be represented in the ordered triple $\boldsymbol{r} = (r_{12}, r_{13}, r_{23})$. The set of all possible $\boldsymbol{r}$ within the cube $[-1,1]^3$ is depicted in Figure 1, and is the subject of Rousseeuw and Molenberghs' (1994) article.

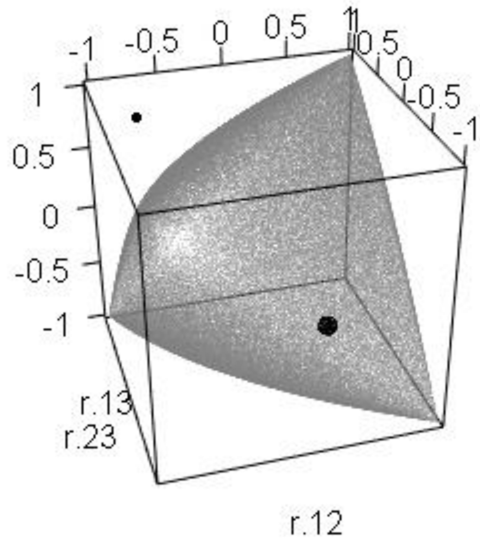


Figure 1. The 3x3 correlation space. The larger point ($r_{12} = .3$, $r_{13} = -.6$, $r_{23} = -.3$) lies within the correlation space and corresponds to a true correlation matrix $\boldsymbol{R}$. The smaller point ($r_{12} = -.7$, $r_{13} = .8$, $r_{23} = .8$) lies outside the correlation space, and corresponds to a non-PSD pseudo-correlation matrix $\boldsymbol{R}^*$.

The convex shape, described as an elliptical tetrahedron or an elliptope (Chai, 2014) meets the edges of the cube at four of the eight corners, and has diagonal lines across the six cube faces. Slicing parallel to any face of the cube, the shape evolves from a diagonal line to ellipses to a perfect circle, before reversing back to ellipses, then a diagonal line along the opposite face from which it started. In the case of 3x3 correlation matrices, the shape of true correlation matrices (i.e., the set of all $\boldsymbol{R}$) occupies approximately 61.7% of the cube. Practically speaking, any $\boldsymbol{r}^*$ generated randomly and uniformly from the $[-1,1]^3$ cube has a 61.7% chance of corresponding to a true correlation matrix. An $\boldsymbol{r}$ closer to the surface of the 3x3 correlation space corresponds to an $\boldsymbol{R}$ that has closer-to-zero eigenvalues or, alternatively stated, has near-linear dependency among the three variables.

Recently, Waller (2016) used the 3x3 correlation space to demonstrate the geometry of fungible correlation matrices. Fungible correlation matrices are $\boldsymbol{R}^*$ that, given a pre-specified set of regression coefficients, will produce identical R-squared values for a linear regression model. The 3x3 correlation space can be used to visualize the set of all 3x3 $\boldsymbol{R}^*$ (or $\boldsymbol{R}$, if desired) that are fungible for a given set of coefficients and R-squared. Such $\boldsymbol{R}^*$ form a cross-sectional plane across the $[-1,1]^3$ cube.

Rousseeuw and Molenberghs (1994) generated Figure 1 by graphing the equation that forms the boundary of the shape (Hubert, 1972); this boundary coincides with the set of all $\boldsymbol{R}$ that have at least one $\lambda_i = 0$, and where $\det(\boldsymbol{R}) = 0$, which implies that

$$r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 \tag{2}$$

Although interesting and informative, the elliptical tetrahedron of the R&M correlation space in Figure 1 is not geometrically intuitive. There are advantages to defining a correlation space with more immediate familiarity and defined mathematical properties. Below I present such a space, first demonstrated by Chai (2014) in a paper that explored methods of sampling from R&M space using various distributions. Although there is overlap in the treatment of the correlation space presented here and Chai's, the current emphasis is on the graphical demonstration of this new space and development and application of several properties of this space and potential insights this space provides for higher dimension spaces.

The 3x3 Cosine Space

I begin with a comment on graphical presentation, and then present a transformation that is advantageous for portraying the correlation space. A useful graphing technique to plot the boundary of the correlation space is to plot enough points $\boldsymbol{r}$ that satisfy the equation for the boundary to visually outline the space. The method employed here involves randomly generating two of the correlations (e.g., r_{12} and r_{13}) uniformly (e.g., from the interval [-1,1], for the correlation space) and solving for the third correlation (r_{23} using Equation (2) to generate a given $\boldsymbol{r}$. Note that, for a given choice of r_{12} and r_{13}, there are two possible r_{23} values (Hubert, 1972): $r_{23} = r_{12}r_{13} + \sqrt{(1 - r_{12}{}^2)(1 - r_{13}{}^2)}$ and $r_{23} = r_{12}r_{13} - \sqrt{(1 - r_{12}{}^2)(1 - r_{13}{}^2)}$. To generate a given $\boldsymbol{r}$, I randomly select one r_{23} value.

The graphical technique aside, it is worth reiterating that, at a very basic level, correlations are cosines on angles. Just as correlations can be generated from the [-1,1] interval, correlations can also be generated as cosines of angles that range from 0 to 180 degrees. (Note that a perfect mapping would transform the interval [-1,1] of correlations to the interval [180,0] of angles. For ease of reading, I use [0,180] and make graphical adjustments where necessary.) Further, the transformation of correlations from the [-1,1] interval to cosines of angles from the [0,180] interval is not a linear transformation. In other words, the shape of 3x3 correlation matrices in cosine space is not expected to match the shape of correlation matrices in 3x3 correlation space.

To portray the cosine space, I developed the following graphical routine, implemented in the software system *R*. Rather than randomly generating two correlations and solving for a third to plot a given point $\boldsymbol{r}$, I implemented the following algorithm:

Step 1. Generate two angles, θ_{12} and θ_{13}, randomly and uniformly from the interval [0, 180].

Step 2. Convert the angles to correlations (r_{12} and r_{13} respectively) and solve for the third correlation (r_{23}) using Equation (2) to generate a given $\boldsymbol{r}$.

Step 3. Convert all three correlations back to angles and plot the ordered triple.[1]

[1] It is not necessary to convert from angles to correlations and then back to angles to produce the cosine space. After generating θ_{12} and θ_{13}, the two possible values for θ_{23} are $\theta_{23} = |\theta_{12} - \theta_{13}|$ and $\theta_{23} = 180 - |\theta_{12} + \theta_{13} - 180|$. To emphasize the relationship between cosine space and correlation space, I convert from angles to correlations and back again, and do so again in Chapter 2.

When repeated a sufficiently large number of times, the above algorithm generates enough $\boldsymbol{r}$ on the boundary of the space to produce a visual representation of the 3x3 cosine space. This visual representation is shown in Figure 2. Such a graphical technique is useful for portraying complex surfaces in 3-dimensional space without explicit graphing of (perhaps intractable) equations, and will be utilized similarly in Chapter 2.

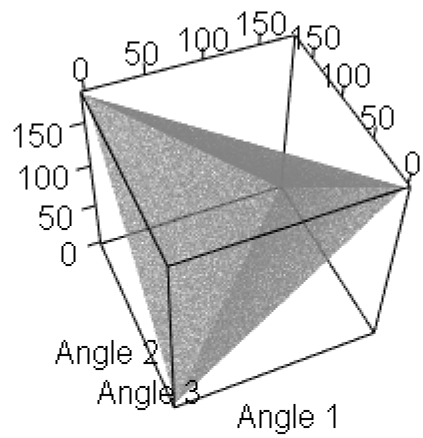


Figure 2. The 3x3 cosine space. The space occupied by 3x3 correlation matrices in cosine space is a regular tetrahedron embedded within a [0,180] cube.

Certain features are immediately apparent from visual inspection of this revised space. First, the axes no longer range from -1 to 1, as they now represent increments on angles, and thus range from 0 to 180 degrees. Second, the shape of this cosine space is a regular 3-simplex, commonly referred to as a tetrahedron. The faces of the tetrahedron are four equilateral triangles, and visually the shape looks like a three-sided pyramid with a triangular base. Third, the shape exactly matches the one independently described by Chai (2014).

Using one of several volume formulas for a regular 3-simplex (e.g., see Stein, 1966), it is easily confirmed that this space takes up exactly one-third of the [0,180] cube. This result compares to the (approximately) 61.7% of the space occupied by the set of all $\boldsymbol{R}$ in the $[-1,1]^3$ cube. The shape is, however, closely related to the shape of the 3x3 correlation space. The shape still meets the surface of the cube at four of the eight corners, and touches each cube face with a single diagonal line. The rounded edges of the correlation shape have become lines, and the "puffy" appearance of each face has flattened. Slices of the space parallel to any cube face would now produce a diagonal line, then rectangles, then eventually a square, before reversing back to rectangles, and then a diagonal line along the opposite face. In terms of correlation coefficients spaced uniformly from -1 to 1, the cosine space involves stretching the correlation space axes nonlinearly, which neutralizes the curvature of the correlation space. As a result, the cosine space compresses the interior points of the cube compared to the 3x3 correlation space. Chai (2014) presented a similar transformation of the 3x3 correlation space, but used the sine function instead of the cosine function, and noted the nonlinear (but one-to-one) mapping from an axis system uniformly defined on [-1,1] to one defined uniformly on angles from [0,180].

Figures 3 and 4 demonstrate in another way the relationship between the 3x3 correlation space and the 3x3 cosine space, i.e., the impact of the nonlinear transformation of the axes. In Figure 3, points are randomly sampled along $[-1,1]^3$ in a light grey scattercloud and are uniformly generated across each of the three axes in dark grey. Once these points are converted to their appropriate $[0,180]^3$ counterparts with the arccosine function, the scattercloud clusters in the middle of the cube and the dark grey points cluster in the middle of their respective axes. In other words, random, uniform sampling in the 3x3 correlation space produces non-uniform sampling in the 3x3 cosine space, and vice versa. This sampling consequence helps explain the seemingly peculiar reduction (by nearly one half) in the volume of the 3x3 correlation space taken up by PSD correlation matrices in 3x3 cosine space. The same infinite set of $\boldsymbol{r}$ occupies both spaces, but the set is more densely packed into the middle within the cosine space.

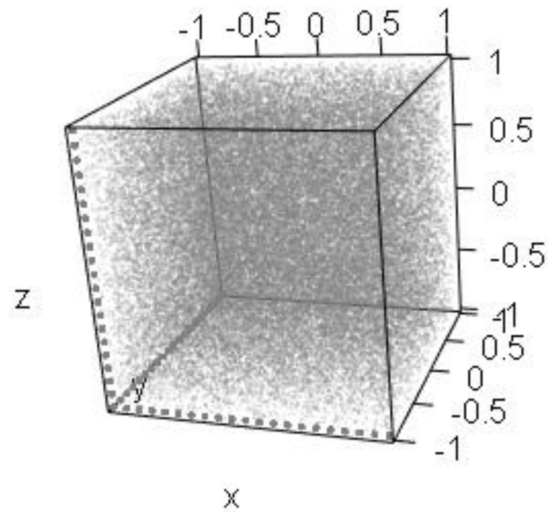


Figure 3. Random sampling from [-1,1]^3.

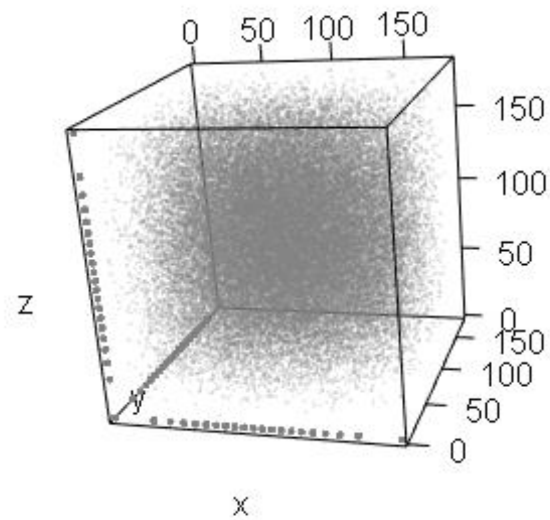


Figure 4. Transformation of the points in [-1,1]^3 to [0,180]^3. Uniform sampling from [-1,1]^3 produces non-uniform sampling from the transformed [0,180]^3.

CHAPTER 2

COSINE SPACES FOR HIGHER DIMENSION CORRELATION MATRICES

In Chapter 1, the 3x3 cosine space for correlation matrices – which takes the form of a regular tetrahedron – was demonstrated graphically, and properties of the cosine space were discussed. In this chapter, I first use graphical methods to demonstrate how higher dimension cosine spaces for correlation matrices, when constrained, can be visualized. I discuss properties of these constrained spaces with particular emphasis on 4x4 correlation matrices. I then discuss some general properties of both correlation and cosine spaces in higher dimensions.

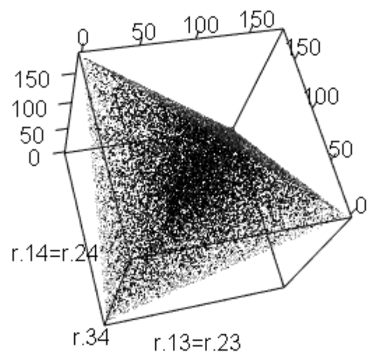
This chapter uses a similar graphical technique to the one employed in Chapter 1 to visual cosine spaces for correlations. However, rather than generating points randomly across the surface of each cosine space, I generate points randomly within the space until the space is saturated sufficiently to visualize its shape.

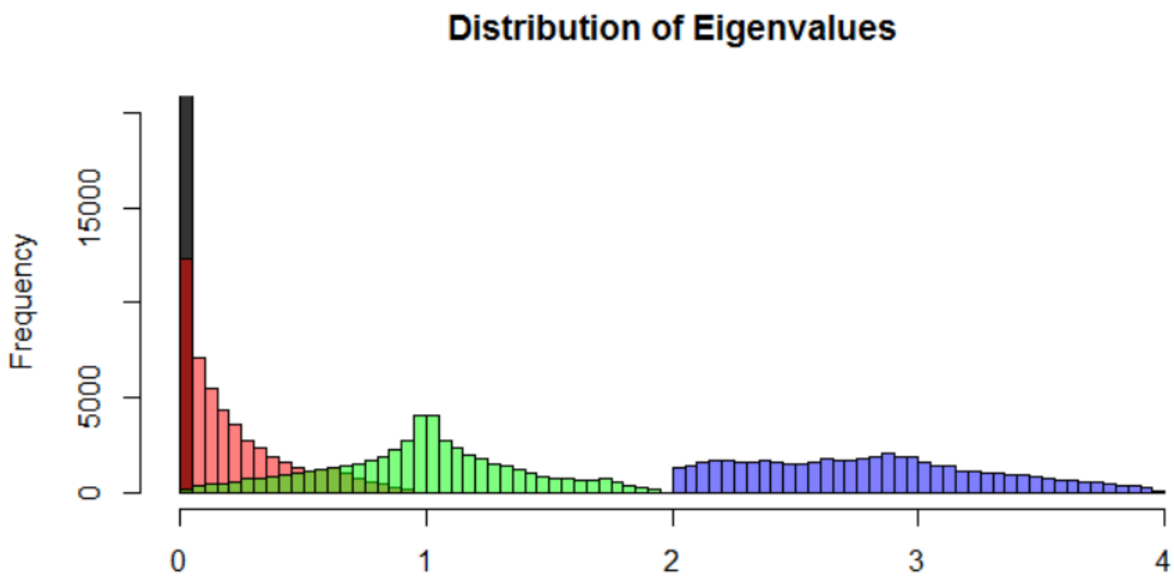
Reproducing the 3x3 Cosine Space with 4x4 Correlation Matrices

Because we can only render visuals up to three dimensions, I begin by projecting from the 4x4 cosine space (which occupies a $[0,180]^6$ hypercube, see Table 1) into the 3x3 cosine space. There are two ways to induce the 3x3 cosine space as a subspace within the n = 6 cosine space occupied by $\boldsymbol{R}$ of order p = 4. The first way is to impose perfect correlation between two of the four variables in $\boldsymbol{R}$. Without loss of generality, the cosine space of 4x4 $\boldsymbol{R}$ of the form specified in Figure 5A will occupy a regular tetrahedron, as in the unrestricted 3x3 cosine space (Figure 5B). While $\boldsymbol{r}$ generated from Figure 5A pervade the entire regular tetrahedron, it should be noted that all $\boldsymbol{R}$ constructed in this way will have at least one zero eigenvalue because of the perfect linear dependency between the first and second variables (Figure 5C). Points on the surface of the tetrahedron in Figure 5B will have at least one additional zero eigenvalue.

A. $$\begin{bmatrix} 1 & & & \\ 1 & 1 & & \\ r_1 & r_1 & 1 & \\ r_2 & r_2 & r_3 & 1 \end{bmatrix}$$

B.





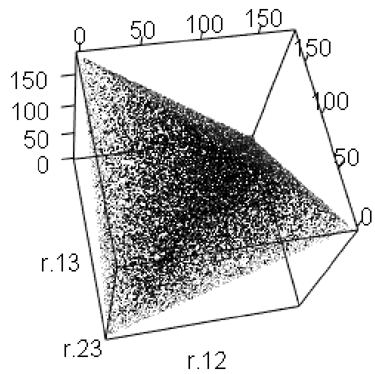
C.

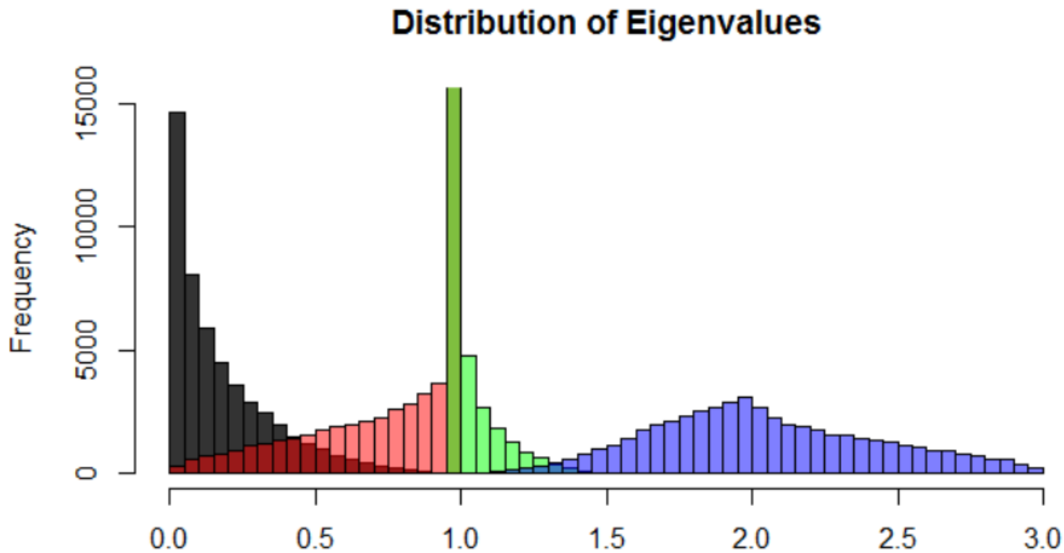
Figure 5. 4x4 correlation matrices with two variables constrained to equality, up to linear transformation. In (A), the first two variables are perfectly correlated ($r_{21} = 1$); this constrains the remaining entries in $\boldsymbol{R}$, such that only three of the six correlations are unrestricted. (B) demonstrates points generated across the cosine space occupied by matrices of the form in (A). (C) demonstrates the first (blue), second (green), third (red), and fourth (black) eigenvalues of 50,000 $\boldsymbol{R}$ generated using (A). The fourth eigenvalue is always zero for matrices of this construction.

The second way to generate the regular tetrahedron as a subspace within the 6-dimensional space occupied by 4x4 $\boldsymbol{R}$ is to allow one variable to be uncorrelated with the remaining three variables (Figure 6B). Without loss of generality, these matrices will be of the form in Figure 6A. Unlike matrices of the form in Figure 5A, 4x4 $\boldsymbol{R}$ constructed by Figure 6A can be PSD or strictly PD, but necessarily have at least one eigenvalue of one (Figure 6C).

A. $$\begin{bmatrix} 1 & & & \\ r_{12} & 1 & & \\ r_{13} & r_{23} & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

B.





C.

Figure 6. 4x4 correlation matrices with one variable uncorrelated with other variables. In (A), the fourth variable is uncorrelated with the other three variables ($r_{41} = r_{42} = r_{43} = 0$), which produces the tetrahedron similarly formed in the unrestricted 3x3 cosine space. (B) demonstrates points generated across the cosine space occupied by $\boldsymbol{R}$ of the form in (A). (C) shows the first (blue), second (green), third (red), and fourth (black) eigenvalues of 50,000 $\boldsymbol{R}$ generated using (A). Each $\boldsymbol{R}$ has at least one eigenvalue of one, as either the second or third eigenvalue.

Similarly, the regular tetrahedron can be imposed from higher dimension correlation matrices by setting all but three variables to be uncorrelated or by constraining some variables to equality such that only three unique variables are present in $\boldsymbol{R}$ of order p. Just as in the 4x4 case, these higher dimension $\boldsymbol{R}$ will have multiple zero eigenvalues (if $\boldsymbol{R}$ is a higher-dimension extension of Figure 5A) or multiple one eigenvalues (if $\boldsymbol{R}$ is a higher-dimension extension of Figure 6A).

4x4 Cosine Spaces Within the Tetrahedron

In the previous section, I demonstrated that, if a fourth variable is added to the 3x3 $\boldsymbol{R}$, such that this fourth variable is uncorrelated with any of the other three variables, the resulting cosine space is exactly the regular tetrahedron observed in the 3x3 cosine space. However, if the fourth variable is correlated with the other three variables by magnitudes greater than zero, the resulting 3-dimensional cosine space will be a subset of the tetrahedral space. Figure 7 shows $\boldsymbol{r}$ generated across the resulting 3-dimensional cosine spaces when the fourth variable is restrained to correlations of r = .3, .6, and .9 with the other three variables in assorted combinations. As the magnitudes of the correlations in the fourth row of $\boldsymbol{R}$ increase, the cosine space for respective set of possible r reduces. This phenomenon is similar to that observed by Hubert (1972) and Olkin (1981) in the range restriction literature: as the magnitudes of correlations r_{13} and r_{23} increase, the potential values for r_{12} are necessarily limited. Likewise, we see the same phenomenon when inspecting horizontal and vertical cross-sections of the 3x3 cosine space: slices at the extremes of the cube produces rectangles with less area than slices towards the center of the cube. Finally, the convexity of the produced cosine spaces is preserved in the slices of these higher dimensions, but the spaces are no longer polyhedrons; that is, curvature is introduced in these constrained 4x4 cosine spaces.

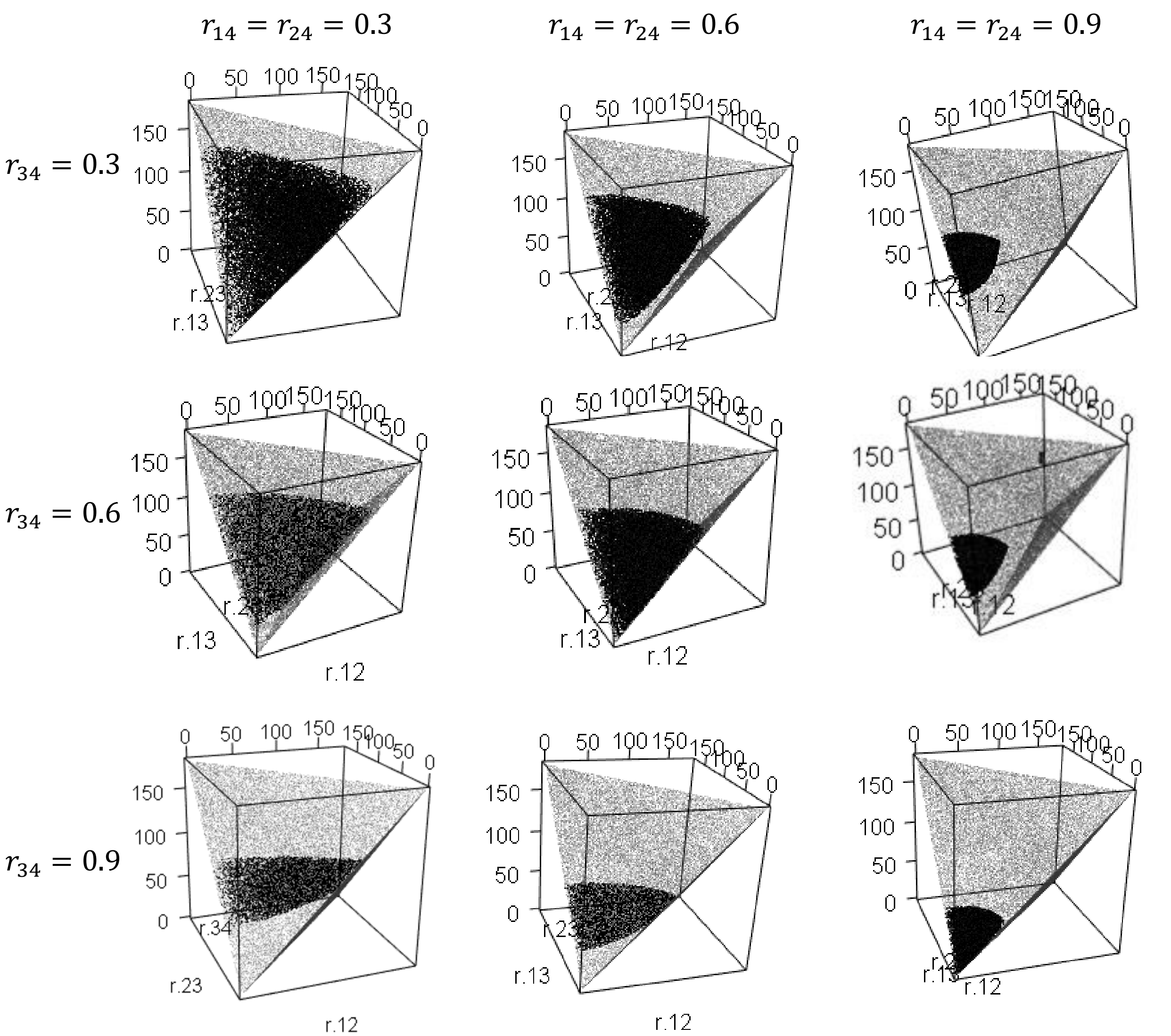


Figure 7. Subsets of the 4x4 cosine space that are subsets of the regular tetrahedron. Points randomly generated across the cosine spaces for 4x4 correlation matrices such that $r_{14} = r_{24}$ = .3, .6, and .9 across the three columns, respectively, and r_{34} = .3, .6, and .9 across the three rows, respectively. The 3x3 cosine space is superimposed in grey. Axes represent increments on angles (in degrees) corresponding to the correlations shown.

Banded Cosine Spaces and Those Spaces That Do Not Subset 3x3 Cosine Space

The previous section explored subsets of the 4x4 cosine space that were also subsets of the 3x3 cosine space. Indeed, the cosine spaces of all 4x4 ***R*** such that a given variable is fixed to pre-specified correlations with the other three variables will produce a cosine space, in that particular slice, that is a subset of the 3x3 cosine space. However, not all possible subsets of the 4x4 cosine space will produce subsets of 3x3 cosine space. A simple example is that of the 4x4 banded cosine space.

Banded correlation matrices are $\boldsymbol{R}$ such that only consecutive variables are allowed to correlate, as in Figure 8A, and all other correlations are fixed to 0. Rousseeuw and Molenberghs (1994) studied 4x4 banded correlation matrices because they can be visually portrayed in 3-dimensional space. The 3x3 correlation space (i.e., the set of all $\boldsymbol{r}$) for banded 4x4 correlation matrices is shown in Figure 8B. The "puffy" figure when sliced across its length and width produces ellipses. The shape touches the top and bottom of the cube directly in the center at a single point; a horizontal slice halfway through the cube (where $r_{32} = 0$) produces a square. Between the endpoints and the midway square, horizontal slices produce squares with rounded edges, with sharper edges observed closer to the center.

A. $\begin{bmatrix} 1 & & & \\ r_{21} & 1 & & \\ 0 & r_{32} & 1 & \\ 0 & 0 & r_{43} & 1 \end{bmatrix}$

B.

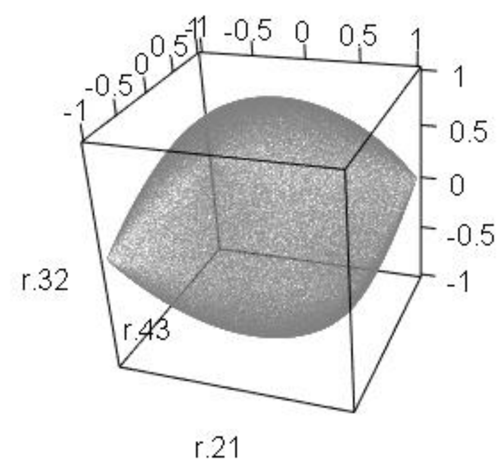


Figure 8. 4x4 banded correlation matrices and correlation space. (A) The form of 4x4 banded correlation matrices. (B) The correlation space occupied by 4x4 correlation matrices of the form in (A).

Similar to their methodology for producing the 3x3 correlation space, Rousseeuw and Molenberghs (1994) plotted the boundary of the shape by graphing the equation corresponding to a zero determinant for the 4x4 banded correlation matrix:

$$(1 - r_{21}^2)(1 - r_{43}^2) = r_{32}^2 \tag{3}$$

As in Chapter 1, I departed slightly from Rousseeuw and Molenberghs (1994) by generating r_{21} and r_{43} randomly and uniformly from the interval [-1,1] and solving for r_{32} using (3) to produce points $\boldsymbol{r}$ on the surface of the shape shown in Figure 8B. From here, I implemented the algorithm described in Chapter 1 to develop the cosine space by generating θ_{21} and θ_{43} and solving for the r_{32} using Equation (3). Sampling a great many times from the boundary, I produced the 4x4 banded cosine space (Figure 9).

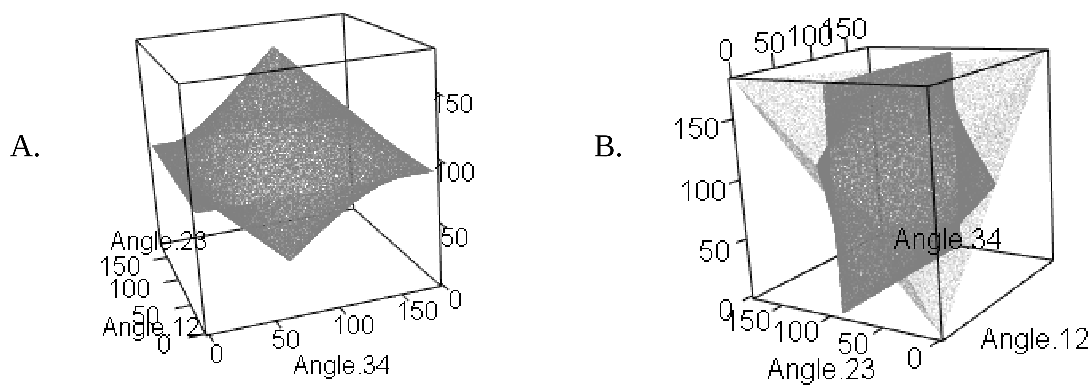


Figure 9 4x4 banded cosine space. (A) and (B). The banded 4x4 cosine space viewed from two different rotations.

The cosine space for 4x4 banded correlation matrices shares many features in common with its counterpart in correlation space. The convex shape touches the top and bottom of cube at precisely the center of the cube faces and retains the square at the midway horizontal slice. The shape also shares similarities with the 3x3 cosine space tetrahedron (Figure 2); the cosine transformation flattens some of the “puffiness” of the banded space. However, unlike the 3x3 cosine space, the 4x4 banded cosine space does not produce as tidy a geometric figure, and some curvature remains in the shape. Finally, from Figure 9 it is clear that the 4x4 banded cosine space is not a subset of the 3x3 cosine space.

When the non-consecutive variables are fixed to values other than zero, the cosine space occupied by such 4x4 ***R*** decreases in volume and increases in curvature as the magnitude of the correlations increases (see Figure 10). Again, these spaces are not subsets of the regular tetrahedron, but neither are they subsets of the 4x4 banded cosine space.

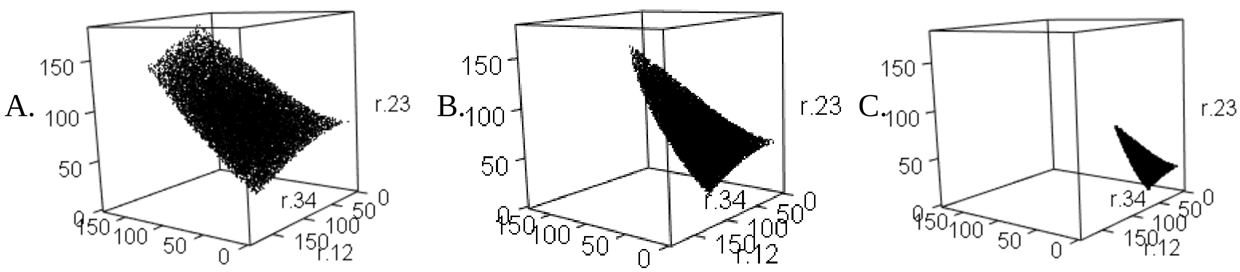


Figure 10. Subsets of the 4x4 cosine space that are not subsets of the tetrahedron. Points randomly generated across the cosine spaces for 4x4 banded ***R*** with correlations between non-consecutive variables set to r = .3, .6, and .9 for (A), (B), and (C) respectively. These cosine spaces are neither subsets of the 3x3 cosine space nor the 4x4 banded cosine space. Axes represent increments on angles (in degrees) corresponding to the correlations shown.

When only three variables in the 4x4 correlation matrix are allowed to freely correlate, and the correlations with a fourth variables are fixed, the cosine spaces are subsets of the tetrahedron in 3x3 cosine space. However, when all four variables are allowed to freely correlate in some way – in other words, when the constraints on the 4x4 correlation matrix do not all occur in the same row and column – then the resulting space will not be a subset of the 3x3 cosine space. This observation extends to higher dimensions: if only three variables are allowed to freely correlate – with the rest of $\boldsymbol{R}$ having numerical constraints – then the resulting 3-dimensional cosine space will be a subset of the regular tetrahedron. If other patterns of fixed correlations are used to reduce the space of the $p\ x\ p$ matrix from n dimensions to three dimensions, then the resulting space does not have to be a subset of the tetrahedron.

As previously stated, horizontal and vertical slices of the 3x3 cosine space will produce rectangles, with the largest-area rectangle (in fact, a square) observed when slicing the figure at 90 degrees on any axis. However, no slice of the tetrahedron on any axis will produce a 2-dimensional subset that spans $[0,180]^2$. Practically, this means that no value can be chosen for one correlation in a 3x3 $\boldsymbol{R}$ such that any possible combination of values for the other two correlations in $\boldsymbol{R}$ will produce a PSD matrix. This phenomenon can occur, however, for spaces that are not subsets of the tetrahedron, such as the 4x4 banded cosine space. Figure 9C shows that the 4x4 banded cosine space does span $[0,180]^2$ if sliced at θ_{32} = 90 degrees (i.e., r_{32} = 0). Conceptually, this makes sense: if r_{32} is zero, then there is no dependency between the first two variables and the last two variables in the 4x4 banded $\boldsymbol{R}$. Similarly, if a 6x6 $\boldsymbol{R}$ were constrained such that X_1 and X_2, X_3 and X_4, X_5 and X_6 were each allowed to pairwise correlate, and all other correlations were fixed to 0, the resulting 3-dimensional cosine space would span $[0,180]^3$ (Figure 11); that is, the three correlations would be independent, and any values for the three correlations would produce a PSD $\boldsymbol{R}$.

A. $\begin{bmatrix} 1 & & & & & \\ r_{12} & 1 & & & & \\ 0 & 0 & 1 & & & \\ 0 & 0 & r_{34} & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & r_{56} & 1 \end{bmatrix}$

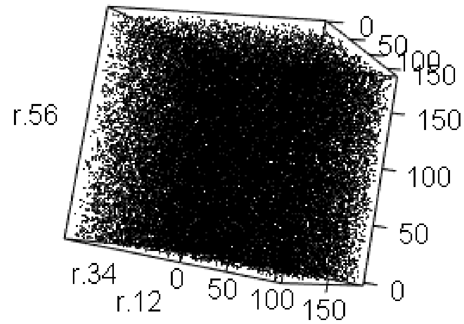
B. 

Figure 11. (A) A 6x6 $\boldsymbol{R}$ such that the separate pairs of variables may correlate and all other correlations are fixed to 0. (B) The cosine space occupied by all $\boldsymbol{R}$ of the form in (A) spans $[0,180]^3$; all possible values for r_{12}, r_{34}, and r_{56} will produce a 6x6 PSD $\boldsymbol{R}$. Axes represent increments on angles (in degrees) corresponding to the correlations shown.

Cosine Space in Higher Dimensions

The simplex structure of the 3x3 cosine space may be informative for understanding the space of correlation matrices in higher dimensions, although the cosine space in higher dimensions is not an n-simplex (Chai, 2014). The set of all p-dimensional correlation matrices – plotted using either correlation or angle units – occupies a convex shape in an $n = p(p-1)/2$-hypercube (Chai, 2014). Further, from the pattern of vertices exhibited by correlation spaces for 3x3 and 4x4 correlation matrices, a general pattern for the number of vertices for a $p \; x \; p$ matrix can be deduced (Table 1). Note that, if the cosine space in higher dimensions were an n-simplex, the number of vertices in n dimension of the space would be $n+1$, rather than 2^{p-1}, which, above three dimensions, will be larger than $n+1$. The ratio between the number of vertices of the correlation space and the number of vertices of the hypercube in which it resides is $\frac{1}{2}^{(p-1)(p-2)/2}$; this relationship results in a quick reduction of the volume of the correlation and cosine spaces as the size of the correlation matrix increases (Bohm & Hornik, 2014; Table 2). In all dimensions, the percent of the hypercube volume occupied by the cosine space is less than the percent occupied by the corresponding correlation space.

p	n	# of vertices (n-hypercube)	# of vertices (correlation space)
1	0	1	1
2	1	2	2
3	3	8	4
4	6	64	8
5	10	1,024	16
⋮	⋮	⋮	⋮
p	$n = p(p-1)/2$	2^n	2^{p-1}

Table 1. General properties of correlation (and cosine) spaces for $p \times p$ correlation matrices.

p	Proportion of hypercube	
	Correlation	Cosine
2	1.00000000	1.000000
3	0.61685027	0.333333
4	0.18277045	0.033323
5	0.02200445	0.000845
6	0.00094952	0.000005
7	0.00001328	<.000001

Table 2. The proportion of the hypercube occupied by correlation and cosine spaces. Proportions for correlation spaces are reproduced from Bohm and Hornik (2014). Proportions for cosine spaces are approximated using rejection sampling in *R* for 1,000,000 pseudo-correlation matrices per matrix size.

Further, the space of correlation matrices in higher dimensions must have a diagonal edge along each hypercube face, corresponding to the degenerate cases of perfect (negative or positive) association among variables. Simply put, the face of a hypercube is the potential combination of two correlations (say, r_{xw} and r_{xv}) after the other correlations in the matrix have been defined as either -1 or 1 (in the case of correlation space), or 180 or 0 (in the case of cosine space). Given the specified values of the other correlations, either $r_{xw} = r_{xv}$ or $r_{xw} = -r_{xv}$, depending on the pattern in the correlation matrix, which produces the diagonal line observed in the hypercube.

CHAPTER 3

APPLICATION OF THE 3X3 COSINE SPACE TO CORRELATION MATRIX GENERATION

The tetrahedral shape of the cosine space in three dimensions adds to a rich history of geometric interpretations of correlation coefficients and their derivatives (see, e.g., Glass & Collins, 1970; Thomas & O'Quigley, 1993; Trosset, 2005). Further, the simple geometric shape has advantageous properties for sampling. Specifically, Rocchini & Cignoni (2000) developed an algorithm for sampling randomly from a tetrahedron within a cube. Because random, uniform sampling across angles in cosine space will produce correlation matrices whose elements tend to be more extreme (i.e., a larger proportion of correlations near -1 and 1, and hence minimum eigenvalues closer to zero) and less homogenous, sampling from the tetrahedron in 3x3 cosine space may be advantageous for methods of generating random correlation matrices.

Methods of Generating Random Correlation Matrices

Monte Carlo studies in economics, genetics, and quantitative methods often require the generation of a large number of random correlation matrices (Hardin, Garcia, & Golan, 2013). Because of the PSD constraint on correlation matrices, the task of generating random $\boldsymbol{R}$ is nontrivial. The simplest method to implement – the rejection method, which involves generating a random $\boldsymbol{R}^*$ and "rejecting" it if it is not PSD – quickly becomes infeasible as p becomes large (Bohm & Hornik, 2014; Numpacharoen & Atsawarungruangkit, 2012).

Since the late 1960s, methods for generating random $\boldsymbol{R}$ with given eigenvalues (e.g., Chalmer, 1975), expected means (Marsaglia & Olkin, 1984), or factor structure (e.g., Tucker, Koopman, & Linn, 1969) have been proposed. Following these early developments, emphasis has shifted to the generation of random $\boldsymbol{R}$ without specified structure. In effect, these newer methods attempt to generate $\boldsymbol{R}$ uniformly across the uniform [-1,1] correlation space of order p.

However, these methods can fail to generate correlation matrices that uniformly cover the correlation space or the range of correlations: methods may produce homogenous $\boldsymbol{R}$ or $\boldsymbol{R}$ with too many near-zero off-diagonal elements (i.e., matrices with small magnitude correlations). Although such generated $\boldsymbol{R}$ have properties that suit some research goals, it may be advantageous to generate matrices more uniformly across the correlation space – or even with an increased density of correlations near -1 and 1. Most recently, generation methods have begun to take advantage of the significant advances in computing power of recent decades, allowing researchers to generate more extreme and more variable correlation matrices, and provide users with more control over the distribution of correlations in generated $\boldsymbol{R}$ (e.g., see Numpacharoen & Atsawarungruangkit, 2012; Lewandowski, Kurowicka, & Joe, 2009). However, these methods can still be computationally inefficient – or worse, produce non-PSD $\boldsymbol{R}^*$ on occasional iterations due to high dimensionality and numerical instability. To this end, sampling randomly from the tetrahedron in 3x3 cosine space can be advantageous because such sampling is fast and does not require optimization or other mathematical techniques which may become unstable.

Demonstration Simulation

To demonstrate how sampling from 3x3 cosine space can be used to increase the frequency of extreme correlation matrices in practice, $\boldsymbol{R}$ of order $p = 3$ and $p = 6$ were explored. To generate matrices of both sizes, four different methods were used: 1) the rejection method, which generates $\boldsymbol{R}^*$ and tests if it is PSD; 2) the *rcorrmatrix* function in *R*, based on the generation method in Joe (2006) using partial correlations; 3) the TT' method, which generates a matrix $\mathbf{T}$ whose p rows are points on the surface of a unit hypersphere of arbitrary dimension; and 4) the TT' method in which some correlations are generated first randomly across the tetrahedron in 3x3 cosine space (hereafter referred to as the "cosine method") using the Rocchini and Cignoni (2000) tetrahedron sampling routine. For $p = 3$ $(n = 3)$, all three off-diagonal elements could be generated from one point from 3x3 cosine space. For $p = 6$ $(n = 15)$, six correlations were generated from two independent draws from 3x3 cosine space. The first draw determined how the first three variables correlated with each other, and the second draw determined both how the last three variables correlated with each other as well as constrained the potential values of the remaining nine correlations (those correlations between the first three and last three variables). Both the TT' and cosine methods employed here assumed $\mathbf{T}$ to be a square matrix. Because the rejection method samples uniformly from the $[-1,1]^n$ hypercube, and keeps only those $\boldsymbol{R}^*$ which are elements of the correlation space, this generation method should most closely sample randomly and uniformly from the correlation space, and thus can serve as a control of sorts for the generation methods. For each method of order p, the eigenvalues and off-diagonal elements of 10,000 randomly generated $\boldsymbol{R}$ were compared.

Results of Simulation Study

The results for the $p = 3$ condition (Figures 12 and 13) show the distributions of correlations and eigenvalues respectively for each of the four generation methods. The cosine method and the TT' method performed similarly, with both methods generating correlations more uniformly across the [-1,1] range of correlations than either the *rcorrmatrix* method or the rejection method. The *rcorrmatrix* method performed similarly to the rejection method, suggesting the *rcorrmatrix* method samples randomly and uniformly from the 3x3 correlation space. Alternatively, the presence of a higher frequency of correlations near -1 and 1 for the cosine and TT' methods indicates non-uniform sampling from the 3x3 correlation space, despite the more uniform distribution of correlation coefficients; instead, the cosine and TT' methods are oversampling random $\boldsymbol{R}$ with greater linear dependency among the three variables, and hence such $\boldsymbol{R}$ correspond to $\boldsymbol{r}$ nearer to the boundary of the correlation space.

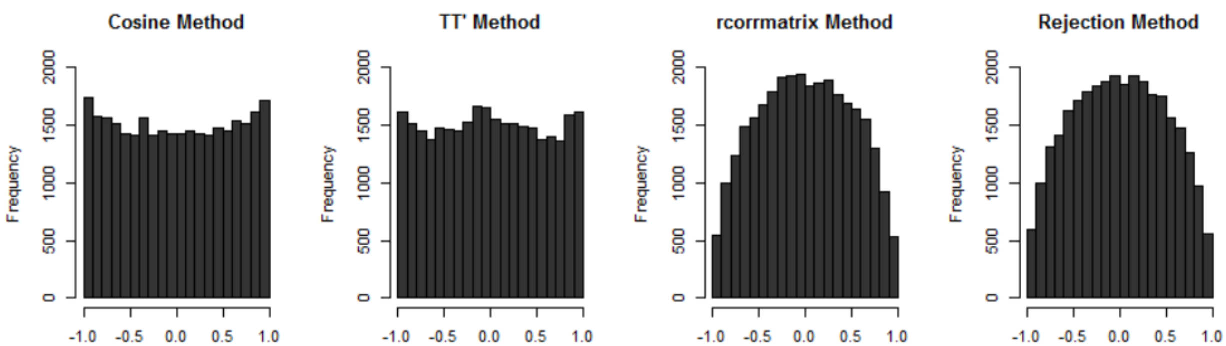


Figure 12. Distribution of correlations across the four generation methods for $\boldsymbol{R}$ of order $p = 3$.

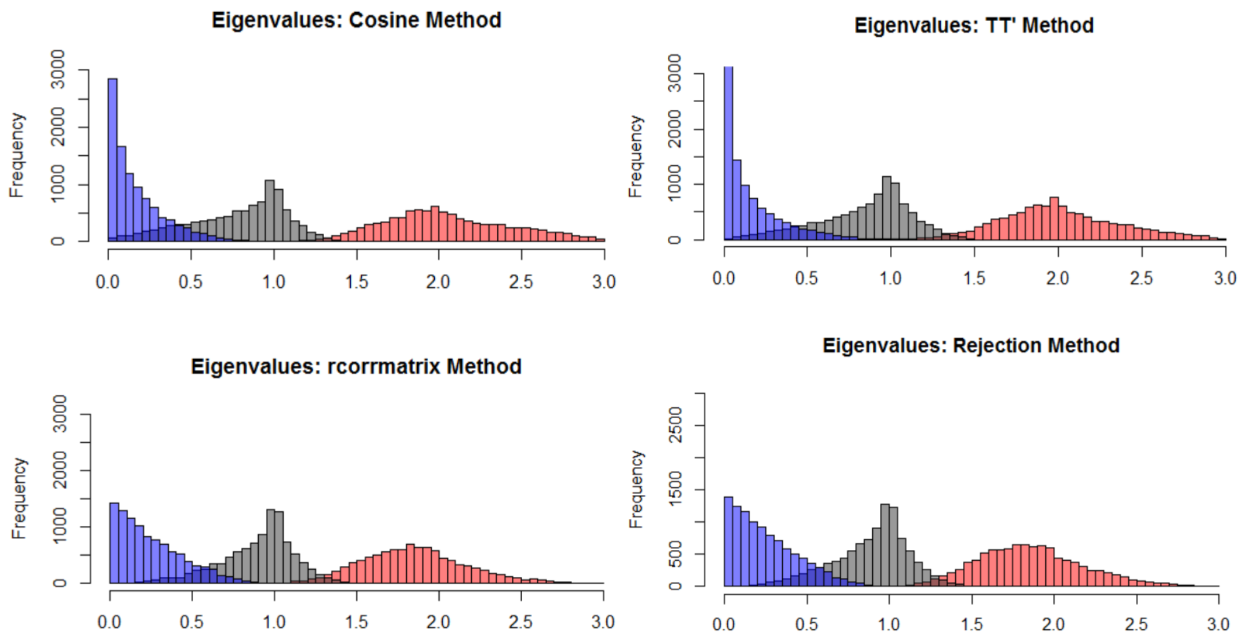


Figure 13. Distributions of the first (red), second (grey), and third (blue) eigenvalues for the four generation methods for $\boldsymbol{R}$ of order $p = 3$.

The distributions of eigenvalues across the four generation methods supports the findings from the distributions of correlations. All methods have modal eigenvalues near two, one, and zero for the first, second, and third eigenvalues respectively, with similar patterns of skewness across the four methods. However, the cosine and TT' method produce minimum eigenvalues closer to zero with more frequency than the *rcorrmatrix* and rejection methods. Further, the distributions of the first eigenvalue for the TT' and cosine methods are somewhat more leptokurtic than the distributions for the *rcorrmatrix* and rejection methods, with the TT' method being somewhat more "peaked" near two than the cosine method.

For $p = 6$, the rejection and *rcorrmatrix* methods produce similar distributions of correlations (Figure 14), as observed for the $p = 3$ condition. Again, this suggests that the *rcorrmatrix* method samples uniformly from the 6x6 correlation space. However, the TT' method does not produce a similar distribution of correlations to the cosine method, as in the $p = 3$

condition; instead, the TT’ method produces a distribution of correlations more similar to the rejection and *rcorrmatrix* methods, with somewhat larger tails of the distribution.

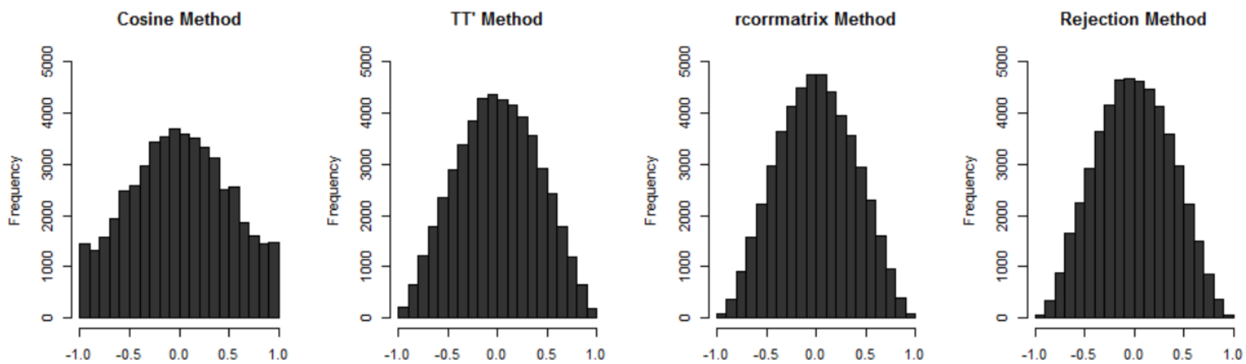


Figure 14. Distribution of correlations across the four generation methods for $\boldsymbol{R}$ of order $p = 6$.

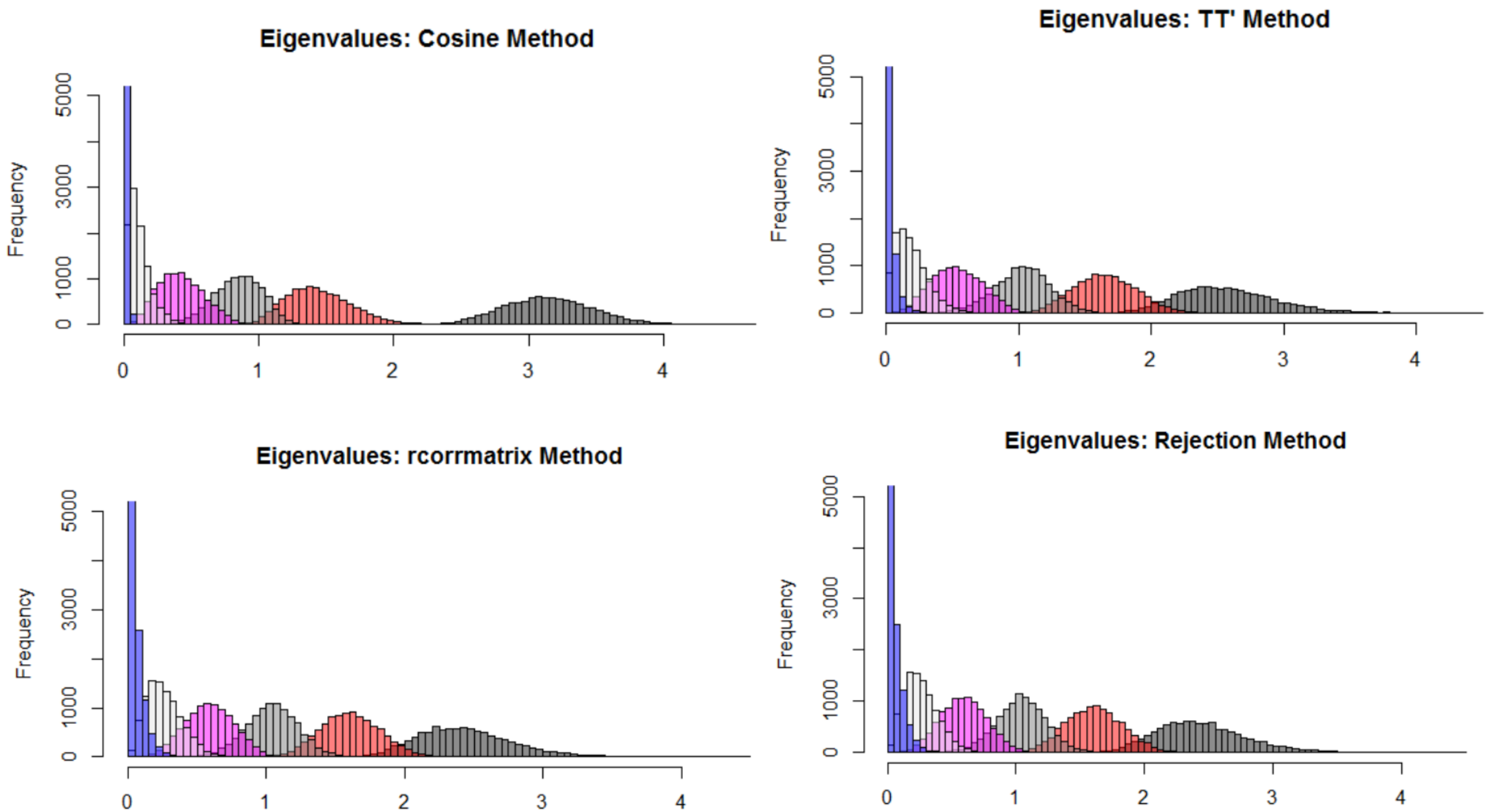


Figure 15. Distributions of the eigenvalues for the four generation methods for $\boldsymbol{R}$ of order $p = 6$.

Summary of Simulation Findings

The goal of the Chapter 3 was to demonstrate how the 3x3 cosine space could be applied to methods of generating random correlation matrices. Four generation methods, including a new method that samples randomly from 3x3 cosine space as an initial step, were used to generate random $\boldsymbol{R}$, and the distributions of correlations and eigenvalues from these correlation matrix samples were compared. Overall, the findings from the study supported the hypotheses. Specifically, the cosine method produced correlations closer to the extremes with higher frequency (and, hence, produced more variable correlation distributions) than either the rejection method or the *rcorrmatrix* method. In both the $p = 3$ and $p = 6$ conditions, the *rcorrmatrix* and rejection methods performed similarly.

One seemingly surprising result of the simulation is the behavior of the TT' method across conditions. For $p = 3$, the TT' method performed similarly to the cosine method, although the cosine method is meant to sample more correlations near -1 and 1 (and more $\boldsymbol{R}$ near the boundary of the correlation space) than the TT' method. However, for $p = 6$, the TT' method performed comparably to the *rcorrmatrix* and rejection methods, indicating fewer extreme $\boldsymbol{R}$ were generated than the cosine method. These results may be explained by the increase in the number of columns of **T** from three to six between the two conditions; the TT' method has been shown to produce varying distributions of correlations, with less extreme correlations and more homogenous $\boldsymbol{R}$ associated with more columns of **T** (Botha, Shapiro, & Steiger, 1988). For few columns of **T**, the TT' method may well oversample near the boundary of the correlation space. Conversely, further increases in the number of columns of **T** than were used in this study may produce *less* extreme correlations and $\boldsymbol{R}$ than either the *rcorrmatrix* or rejection methods.

Another surprising result is the scarcity of eigenvalues near two for the cosine method for $p = 6$. While the other three methods have eigenvalues that spread generally across the range of potential eigenvalues (for $p = 6$, the range of observed eigenvalues across generation methods was approximately [0,4.5]), the cosine method was more likely to produce a larger first eigenvalue and smaller second, third, etc. eigenvalues. This phenomenon may be a result of the oversampling from extreme correlations; as correlations nearer to -1 and 1 are generated, the correlation matrix is more likely to have a near-zero eigenvalue. On average, this smaller last eigenvalue may be compensated for by a larger first eigenvalue, which would have the observed effect of forcing the distribution of the first eigenvalue further to the right. Another possible explanation involves the method by which sampling from the 3x3 cosine space was implemented. For $p = 6$, two independent draws from 3x3 cosine space were performed to generate more extreme $\boldsymbol{R}$. If, however, only one draw from 3x3 cosine space was conducted before implementing the rest of the TT' algorithm, the distribution of correlations and eigenvalues may look different.

Finally, it is worth noting the change in the distribution of correlation coefficients between the $p = 3$ and $p = 6$ conditions for all four generation methods – particularly for the rejection method. Because this method most closely samples randomly and uniformly from the correlation space of a given size, that the distribution of correlations for the rejection method changes depending on p indicates that the distribution of correlations across the correlation space also depends on p.

CHAPTER 4

DISCUSSION AND CONCLUSION

In the present work, I sought to explore the properties of correlation matrices and the space they occupy when correlations are interpreted as cosines of angles. Chapter 1 generated this cosine space for 3x3 correlation matrices, identified the shape of the cosine space as a regular tetrahedron, and discussed properties of the 3x3 cosine space. Chapter 2 expanded on the findings from Chapter 1 by first investigating 4x4 cosine spaces before expanding the scope of interest to higher dimension correlation matrices. Finally, Chapter 3 demonstrated how the 3x3 cosine space can be used to randomly generate correlation matrices with more extreme correlation coefficients – and hence more linear dependency – than other generation methods.

There are several limitations to the current research. First, and perhaps most obvious, is the lack of a more complete understanding of how correlation and cosine spaces generalize beyond three dimensions. While some aspects of these spaces are known – for instance, the number of dimensions in which they exist, the number of vertices of the space, and certain constraints that project higher dimension spaces into subsets of known spaces (i.e., the 3x3 cosine space) – there are many aspects of the higher dimension cosine spaces still left to discover. For example, although the cosine transformation linearizes the curvature of the 3x3 correlation space, the subsets of 4x4 cosine space shown in Figures 7-10 all have varying degrees of curvature. Hence, it is unclear if cosine spaces in higher dimensions are flat-sided polytopes, or some other convex shapes with curved faces and edges. Future research may focus on such correlation and cosine space generalizations.

Relatedly, another limitation of the current project is that random sampling from cosine space is practical only in three dimensions. This is a limitation of both the sampling method (which requires a tetrahedron with known vertices) and of the knowledge of the general form of cosine spaces. Presently, to capitalize on the greater frequency of extreme correlation matrices in cosine space, sampling from 3x3 cosine space must be integrated with at least one other generation method. Depending on the choice of method, or the number of independent draws from 3x3 cosine space, the researcher may end up with markedly different structures for their generated correlation matrices. An ideal solution to this issue would be a known generalization of the cosine space as well as an efficient algorithm to sample from such a space. A more practical remedy would be to combine the resulting samples from several different generation methods until the researcher achieves the desired distribution of correlations or eigenvalues. Further research into efficient methods of generating correlation matrices across cosine space, or near the boundary of the correlation space, would benefit researchers in many fields, such as finance, where correlation matrices with high redundancy and near-zero eigenvalues are the norm (Higham, 2002).

As a final note, future work should be done investigating how strategic cross-sections of higher dimensions project into 2- or 3-dimensions. Such work may shed light on the general form of cosine spaces in higher dimensions, but the intermediate reward is deeper understanding into how correlations in a given correlation matrix are interdependent. Such graphical demonstrations as presented in Chapter 2 may help the novice statistical student understand concepts such as linear dependence, restriction of range, and eigenvalues using a geometric rather than an algebraic approach. These topics can be abstruse – but fascinating – to even those with statistical predilections.

REFERENCES

Böhm, W., & Hornik, K. (2014). Generating random correlation matrices by the simple rejection method: Why it does not work. *Statistics and Probability Letters, 87*, 27-30. doi:10.1016/j.spl.2013.12.012

Botha, J. D., Shapiro, A., & Steiger, J. H. (1988). Uniform indices-of-fit for factor analysis models. *Multivariate Behavioral Research, 23*(4), 443-450.

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist*, New York: John Wiley.

Browne, M., & Shapiro, A. (1986). The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra and its Applications, 82*, 169–176. doi:10.1016/0024-3795(86)90150-3

Chai, K. M. A. (2014). Three-by-three correlation matrices: Its exact shape and a family of distributions. *Linear Algebra and its Applications, 458*, 589-604. doi:10.1016/j.laa.2014.06.039

Chalmers, C. P. (1975). Generation of correlation matrices with a given eigen–structure. *Journal of Statistical Computation and Simulation, 4*(2), 133-139.

Coxeter, H. S. M. (1973). *Regular Polytopes*, Courier Corporation.

Glass, G. V., & Collins, J. R. (1970). Geometric proof of the restriction on the possible values of rxy when ryz are fixed. *Educational and Psychological Measurement, 30*, 37-39. doi: 10.1177/001316447003000103

Higham, N. J. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis, 22*(3), 329-343. doi: 10.1093/imanum/22.3.329

Hubert, L. J. (1972). A note on the restriction of range for Pearson product-moment correlation coefficients. *Educational and Psychological Measurement, 32*, 767-770.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis, 97*, 2177–2189.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*(3), 457-477. doi: 10.1207/s15327906mbr2603_5

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*(9), 1989-2001.

Marsaglia, G., & Olkin, I. (1984). Generating correlation matrices. *SIAM Journal on Scientific and Statistical Computing 5*(2), 470-475. doi: 10.1137/0905034

Numpacharoen, K., & Atsawarungruangkit, A. (2012). Generating correlation matrices based on the boundaries of their coefficients. *PLOS ONE, 7*(11), e48902.

Olkin, I. (1981). Range restrictions for product-moment correlation matrices. *Psychometrika, 46*(4), 469-472.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66. doi: 10.1080/00031305.1988.10475524

Rocchini, C., & Cignoni, P (2000). Generating random points in a tetrahedron. *Journal of Graphics Tools*, *5*(4), 9-12. doi: 10.1080/10867651.2000.10487528

Rousseeuw, P. J., & Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, *48*(4), 276-279. doi: 10.2307/2684832

Stein, P. (1966). A note on the volume of a simplex. *American Mathematical Monthly, 73*(3), 299-301. doi: 10.2307/2315353

Thomas, G., & O'Quigley, J. (1993). A geometric interpretation of partial correlation using spherical triangles. *The American Statistician*, 47, 30-32.

Trosset, M. W. (2005). Visualizing correlation. *Journal of Computational and Graphical Statistics*, *14*(1), 1-19. doi: 10.1198/106186005X27004

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421-459.

Waller, N. G. (2016). Fungible correlation matrices: A method for generating nonsingular, singular, and improper correlation matrices for Monte Carlo research. *Multivariate Behavioral Research*, 1-15. doi: 10.1080/00273171.2016.1178566