The Shape of Correlation Matrices

Peter J. ROUSSEEUW and Geert MOLENBERGHS

A correlation matrix between three variables has to satisfy certain conditions. Such a matrix essentially contains three numbers and thus can be represented by a point in three dimensions. The set of all possible correlation matrices yields a convex solid body with an uncommon shape. All its cross sections perpendicular to the axes are ellipses. At the same time, its surface contains the vertices and edges of a regular tetrahedron. Another unusual shape is obtained for banded correlation matrices between four variables.

KEY WORDS: Convexity; Correlation coefficient; Elliptical tetrahedron; Graphical display; Range restrictions.

1. INTRODUCTION

The correlation coefficient is one of the most frequently used statistical tools. The correlation r_{XY} between two variables X and Y can be interpreted intuitively by looking at scatterplots of observed data points or at bivariate population densities (see, e.g., Rodgers and Nicewander 1988). However, things become harder to visualize when there are three variables X, Y, and Z, which yield the pairwise correlations r_{XY}, r_{XZ}, and r_{YZ}. The latter correlations are somehow intertwined: for instance, knowing two of them gives some information about the third. In this article we will construct a three-dimensional graph of the set of all combinations (r_{XY}, r_{XZ}, r_{YZ}), which has very peculiar properties. We will also study correlation matrices between four variables. The results given throughout this article are valid both for empirical correlations computed from data as well as for their population versions.

In the simplest case we only have two variables, X and Y, and we can compute the correlation r_{XY} between them. The correlation matrix of X and Y is

$$\mathbf{C} = \begin{pmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{pmatrix} \tag{1}$$

where the diagonal entries are always 1. Moreover, any correlation matrix is symmetric, so it suffices to list the upper triangular part. From the definition of correlation, it follows that $-1 \leq r_{XY} \leq 1$. Conversely, if we take any number between -1 and 1 and plug it into (1), the resulting $\mathbf{C}$ is always a correlation matrix (e.g., there will exist a bivariate Gaussian distribution with correlation matrix $\mathbf{C}$). Therefore, it is trivial to see whether a given 2-by-2 matrix is a correlation matrix.

Peter J. Rousseeuw is Professor, Department of Mathematics and Computer Science, Universitaire Instelling Antwerpen (U.I.A.), Universiteitsplein 1, B-2610 Wilrijk, Belgium. Geert Molenberghs is Assistant Professor, Department of Mathematics, Physics, and Computer Science, Limburgs Universitair Centrum (L.U.C.), Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium. The authors wish to thank an associate editor and two referees for very helpful comments.

2. CORRELATIONS BETWEEN THREE VARIABLES

When considering three variables X, Y, and Z, the correlation matrix has the form

$$\mathbf{C} = \begin{pmatrix} 1 & r_{XY} & r_{XZ} \\ r_{XY} & 1 & r_{YZ} \\ r_{XZ} & r_{YZ} & 1 \end{pmatrix} \tag{2}$$

which is again symmetric with diagonal entries of 1. By definition, each of the three correlation coefficients must lie in the interval $[-1, 1]$. However, there is more to a correlation matrix than that. We cannot take any combination of three numbers in $[-1, 1]$, plug them into (2), and be certain that the resulting $\mathbf{C}$ is a correlation matrix. From elementary matrix algebra it follows that a matrix $\mathbf{C}$ of type (2) is a correlation matrix (e.g., of a multivariate Gaussian distribution) if and only if it is *positive semidefinite* (PSD), meaning that $\mathbf{v}'\mathbf{C}\mathbf{v} \geq 0$ for any column vector $\mathbf{v}$.

Our goal is to picture the set of combinations $\mathbf{r} = (r_{XY}, r_{XZ}, r_{YZ})$ that arises in this way. Each of these combinations can be seen as a point in the cube $[-1, 1]^3$ in three-dimensional space. Let us denote the set of these points by $\mathbf{R}$, so that each point $\mathbf{r}$ in $\mathbf{R}$ corresponds to a PSD matrix $\mathbf{C}$ and vice versa. Therefore $\mathbf{R}$ is bounded (being a subset of $[-1, 1]^3$) and it contains the point $(0, 0, 0)$ that corresponds to the identity matrix. Moreover, the region $\mathbf{R}$ is convex because the set of PSD matrices is convex. For all points in $\mathbf{R}$, the determinant of $\mathbf{C}$ must be nonnegative:

$$\det(\mathbf{C}) = 1 + 2r_{XY}r_{XZ}r_{YZ} - r_{XY}^2 - r_{XZ}^2 - r_{YZ}^2 \geq 0. \tag{3}$$

For matrices $\mathbf{C}$ that are *positive definite* (i.e., $\mathbf{v}'\mathbf{C}\mathbf{v} > 0$ for any column vector $\mathbf{v} \neq 0$), the determinant (3) is even strictly positive. On the boundary of $\mathbf{R}$ the determinant becomes zero, yielding the equation

$$r_{XY}^2 + r_{XZ}^2 + r_{YZ}^2 - 2r_{XY}r_{XZ}r_{YZ} = 1. \tag{4}$$

We can verify that in the cube $[-1, 1]^3$, Equation (4) determines a closed surface, and that $\det(\mathbf{C})$ is strictly positive inside and strictly negative outside. Therefore, the surface (4) determines the solid region $\mathbf{R}$.

Figure 1a gives a picture of $\mathbf{R}$. It is clearly convex with four sharp vertices. It is symmetric with respect to certain mirror reflections and to rotations corresponding to permutations of the components of (r_{XY}, r_{XZ}, r_{YZ}). Although (4) has no flat surfaces anywhere, it does contain six straight line segments connecting the four vertices. These line segments form a regular tetrahedron. Note that the surface is rather smooth and that it is only sharp at the four vertices. The surface (4) also contains three complete circles (orthogonal to each other) with unit radius, corresponding to the cross sections with $r_{XY} = 0$, $r_{XZ} = 0$, and $r_{YZ} = 0$.

Surprisingly, any horizontal cross section of this surface is an ellipse. Indeed, if we fix $r_{YZ} = c$ with $|c| < 1$, we find

$$r_{XY}^2 + r_{XZ}^2 - 2cr_{XY}r_{XZ} = 1 - c^2. \tag{5}$$

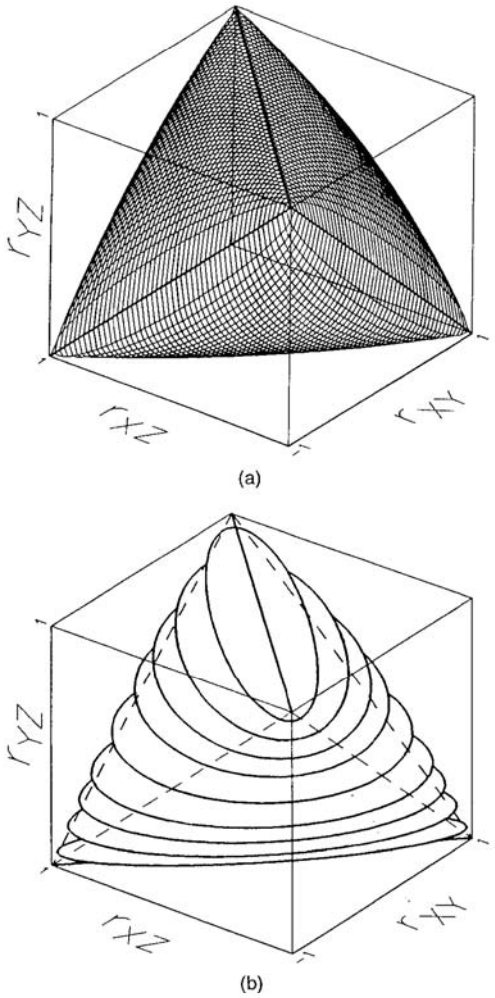

Figure 1. (a) Set of all Possible Correlations Between X, Y, and Z. (b) Slicing this set at r_{YZ} yields ellipses.

Some typical slices are shown in Figure 1b. For $c = 0$, we obtain the circle mentioned previously. For $0 < c < 1$, we find an ellipse with a major axis in the direction of the line $r_{XY} = r_{XZ}$ and a minor axis in the direction of $r_{XY} = -r_{XZ}$. For $-1 < c < 0$, the major and minor axes are interchanged. In the limiting cases $c = 1$ and $c = -1$, the ellipse degenerates to a line segment on a face of the cube (e.g., $r_{YZ} = 1$ implies that $r_{XY} = r_{XZ}$ and that the latter may take any value between -1 and $+1$.)

In conclusion, the surface is made up of many ellipses, and the same is true if we slice the surface according to fixed values of r_{XY} or r_{XZ}. Nevertheless, the surface is not an ellipsoid (in fact, (4) is a cubic in $\mathbf{r}$ because the term $-2r_{XY}r_{XZ}r_{YZ}$ is of degree 3), and it contains all the edges of a regular tetrahedron. We call this surface an *elliptical tetrahedron.* It appears to have interesting physical properties (e.g., it resembles an elastic-membrane tetrahedron that is being inflated), but it may not be round enough for a ball game. It would be nice to make a solid model, for instance, carved out of wood or using rapid prototyping, or to rotate a model on a computer screen using video software.

Projections of $\mathbf{R}$ can have different shapes. If we project $\mathbf{R}$ on a horizontal plane (e.g., $r_{YZ} = 0$), we obtain a whole square. The same happens if we project $\mathbf{R}$ on the vertical plane orthogonal to $(1, 0, 0)$ or the plane orthogonal to $(0, 1, 0)$. If we project $\mathbf{R}$ on a plane orthogonal to one of the main diagonals, such as $(1, 1, 1)$, then we obtain a triangle with rounded edges, as used in a Wankel engine or a movie projector.

The volume of $\mathbf{R}$ can be computed by elementary calculus, yielding $V = \pi^2/2 \approx 4.93$. This means that if we generate three numbers r_{XY}, r_{XZ}, and r_{YZ} independently of each other and uniform in $[-1, 1]$, the probability that the resulting $\mathbf{C}$ is a true correlation matrix equals only $V/8 = \pi^2/16 \approx 61.7\%$.

Remark 1. For any matrix $\mathbf{C}$ of the type (2) with off-diagonal elements belonging to $[-1, 1]$, it holds that $\mathbf{C}$ is PSD if and only if condition (3) is satisfied. This is a special case of a more general result in matrix algebra (see, for instance, Rao 1965, sec. 1.c), saying that a symmetric $p \times p$ matrix $\mathbf{C}$ is PSD if and only if all its symmetric submatrices, including $\mathbf{C}$ itself, have a nonnegative determinant. (Here, a symmetric submatrix is defined by removing at most $p - 1$ rows and the corresponding columns.)

Remark 2. Condition (3) is algebraically equivalent to formula (3) of Leung and Lam (1975), which gives upper and lower bounds on r_{XZ} assuming that r_{XY} and r_{YZ} are known. Olkin (1981) considered range restrictions in the multivariate case. Thomas and O'Quigley (1993) obtained a related formulation in the trivariate case, making use of spherical trigonometry in the space of the variables. Also some particular cases have received attention: for instance, Brown and Eagleson (1984) described a situation where the sample correlations r_{XY}, r_{XZ}, and r_{YZ} are pairwise independent but not independent as a triple. More recently, Hamilton (1987) obtained an example where $r_{XZ}^2 + r_{YZ}^2$ is smaller than the coefficient of determination. These articles also provide references to earlier literature on the subject. In this context, Figure 1 yields a visual aid to intuition.

Remark 3. We arrived at the set $\mathbf{R}$ in a different way, in connection with the following question: given a matrix $\mathbf{C}$ of type (2) that is not PSD, how can we transform it to a true correlation matrix? Several approaches to this were reviewed in Rousseeuw and Molenberghs (1993). In retrospect, some of these transformations can be interpreted geometrically. The *linear shrinking method* starts from a point in $[-1, 1]^3$, outside of $\mathbf{R}$, and moves the point to the boundary of $\mathbf{R}$ along the ray through $(0, 0, 0)$. The *scaling method* searches for the point in $\mathbf{R}$ closest (in Euclidean distance) to the starting point, thereby carrying out an orthogonal projection on $\mathbf{R}$.

3. CORRELATIONS BETWEEN FOUR VARIABLES

When correlating four variables X, Y, Z, and U, the matrix $\mathbf{C}$ becomes a symmetric 4-by-4 matrix with six upper diagonal entries. Therefore, each matrix $\mathbf{C}$ corresponds to a point $(r_{XY}, r_{XZ}, r_{XU}, r_{YZ}, r_{YU}, r_{ZU})$ in six-dimensional

space. Using Remark 1, we can verify that a symmetric 4-by-4 matrix $\mathbf{C}$ with diagonal elements of 1 and off-diagonal elements belonging to $[-1, 1]$ is PSD if and only if (a) all its symmetric 3-by-3 submatrices are PSD as in Section 2 and (b) it holds that $\det(\mathbf{C}) \geq 0$. The expression of $\det(\mathbf{C})$ is a lengthy polynomial of degree 4 in r_{XY}, r_{XZ}, r_{XU}, r_{YZ}, r_{YU}, and r_{ZU}.

It becomes much harder than in the previous section to visualize the set $\mathbf{R}$ of the points $(r_{XY}, r_{XZ}, r_{XU}, r_{YZ}, r_{YU}, r_{ZU})$ for which $\mathbf{C}$ is PSD. $\mathbf{R}$ remains convex but is now a subset of the six-dimensional cube $[-1, 1]^6$. However, we can still compute the (hyper-)volume of $\mathbf{R}$. If we divide it by the volume of the cube, we find the probability that a matrix $\mathbf{C}$ obtained by uniformly generating six numbers in $[-1, 1]$ is a true correlation matrix. This probability is 18.3%, which is lower than the 61.7% of the trivariate case because the requirements on $\mathbf{C}$ include the condition (3) on the first three variables X, Y, and Z.

4. BANDED CORRELATION MATRICES

The reasoning in this article can also be applied to correlation matrices with a special structure. A very useful type is the *banded* (or *tridiagonal*) correlation matrix, in which there can only be a correlation between adjacent variables. This arises naturally in practice, for instance, when the variables $X, Y, Z, \ldots$ are measured one after the other and the only dependence is between succesive measurements. (This is called the *one-dependent* model.)

An important application of banded correlation matrices is in the popular method of Liang and Zeger (1986) for analyzing longitudinal responses. Its computation is much simpler than specifying the joint distribution of the outcomes and estimating its parameters by maximum likelihood, whereas consistency and related properties still hold. The strength of this technique is that the true correlation matrix between the outcomes need not be known exactly and can be replaced by a simple "working" correlation matrix, for instance, of banded type.

In the 3-by-3 case, a banded correlation matrix is of the form

$$\mathbf{C} = \begin{pmatrix} 1 & r_{XY} & 0 \\ r_{XY} & 1 & r_{YZ} \\ 0 & r_{YZ} & 1 \end{pmatrix}.$$

This matrix is PSD if (3) holds, which reduces to

$$r_{XY}^2 + r_{YZ}^2 \leq 1. \tag{6}$$

Therefore, the set $\mathbf{R}$ of points (r_{XY}, r_{YZ}) is simply the unit disk in the plane.

Things become more interesting in the 4-by-4 case where

$$\mathbf{C} = \begin{pmatrix} 1 & r_{XY} & 0 & 0 \\ r_{XY} & 1 & r_{YZ} & 0 \\ 0 & r_{YZ} & 1 & r_{ZU} \\ 0 & 0 & r_{ZU} & 1 \end{pmatrix}.$$

Using Remark 1 we find that $\mathbf{C}$ is PSD if and only if

$$\det(\mathbf{C}) = (1 - r_{XY}^2)(1 - r_{ZU}^2) - r_{YZ}^2 \geq 0 \tag{7}$$

because (7) also implies the conditions on the symmetric 3-by-3 submatrices. We see that r_{XY} and r_{ZU} may be interchanged in (7) but that r_{YZ} plays a different role, so we will plot the latter on the vertical axis. The points in $[-1, 1]^3$ where (7) is zero form the surface shown in Figure 2a. It contains the points $(0, 0, 1)$ and $(0, 0, -1)$ on the vertical axis. In the horizontal plane it contains four vertices and four line segments connecting them, which together form a square. The surface has these features in common with a regular octahedron, but is more rounded.

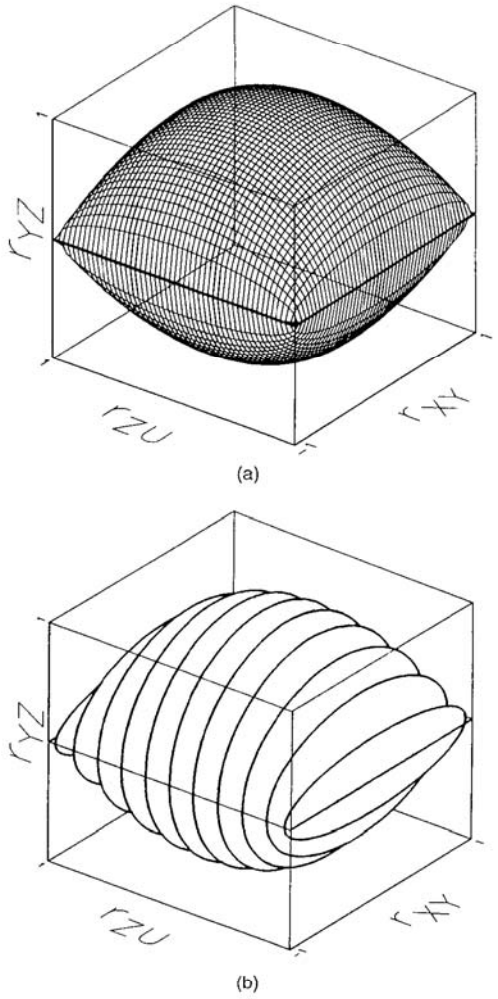

Figure 2. (a) Set of all Banded Correlation Matrices Between X, Y, Z, and U. (b) Slicing this set at r_{ZU} yields ellipses.

In Figure 2b we see that cutting the surface for fixed values of r_{ZU} yields ellipses of which the major axis is horizontal and the minor axis is vertical. (Note that the ellipse in the middle is a circle.) This implies that the surface, which we might call an *elliptical octahedron*, passes smoothly through the edges of the square. The surface is actually smooth everywhere, except at the four vertices of the square. Physically, it resembles inflating two elastic membranes attached to a rigid square frame.

If we project this solid on a horizontal plane, we obtain a square. On the other hand, projecting the solid on a vertical plane orthogonal to $(1, 0, 0)$ or $(0, 1, 0)$ yields a perfect circle. Therefore, this object casts very different shadows depending on the direction of the light.

5. CONCLUDING REMARKS

It is hoped that Figures 1 and 2 may help to develop intuition for the constraints that exist between correlations computed from three or more variables. Unlike the tabular format of correlation matrices, these graphical displays remind us that correlations combine in certain ways. The solids formed by these combinations are uncommon, but they can be investigated by elementary methods.

[*Received January 1992. Revised March 1994.*]

REFERENCES

Brown, T. C., and Eagleson, G. K. (1984), "A Useful Property of Some Symmetric Statistics," *The American Statistician*, 38, 63–65.

Hamilton, D. (1987), "Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$," *The American Statistician*, 41, 129–132.

Leung, C.-K., and Lam, K. (1975), "A Note on the Geometric Representation of the Correlation Coefficients," *The American Statistician*, 29, 128–130.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Olkin, I. (1981), "Range Restrictions for Product-Moment Correlation Matrices," *Psychometrika*, 46, 469–472.

Rao, C. R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Rodgers, J. L., and Nicewander, W. A. (1988), "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician*, 42, 59–66.

Rousseeuw, P. J., and Molenberghs, G. (1993), "Transformation of Non Positive Semidefinite Correlation Matrices," *Communications in Statistics—Theory and Methods*, 22, 965–984.

Thomas, G., and O'Quigley, J. (1993), "A Geometric Interpretation of Partial Correlation Using Spherical Triangles," *The American Statistician*, 47, 30–32.