Chapter 2

THE MATHEMATICS OF CORRELATION MATRICES

Numbers are basic mathematical elements. Most people have a good intuition for what a number is. Seven is a number, as is −182.1. π (the ratio of a circle's circumference to its diameter, approximately 3.14159) is also a number. There are different types of numbers, such as integer, rational, irrational, and real numbers. Algebra and more advanced number theory textbooks can be consulted to develop understanding of numbers.

Numbers have characteristics. For example, a number can be whole or a decimal. A number can be positive or negative, real or imaginary. The number 2 is even; the number 7 is prime. Integer numbers are either even or odd. Even integer numbers bigger than 2 cannot be prime. Knowing the characteristics of a number helps us better understand the number.

Matrices are mathematical elements, like numbers. Matrices are defined as a rectangular array of numbers (called "scalars" in matrix algebra) arranged into rows and columns. Several examples of a certain type of matrix—correlation matrices—have already been presented in Chapter 1. The scalars that compose the matrix rows and columns can be seen in those examples; for example, in Table 1.4 there are 25 scalars, each representing a correlation between variables that combine to define a 5 × 5 correlation matrix.

Like numbers, matrices also have characteristics. One characteristic of a matrix is its dimensions—how many rows and columns the matrix has. By convention, the number of rows is listed before the number of columns; a 4 × 3 matrix is one that has four rows and three columns, for a total of 4 × 3 = 12 scalar elements. Some matrices with special patterns of dimensions have specific names. For example, a vector is a matrix that has either a single column or a single row; conceptually, the reader may imagine that the scalars are strung into a line horizontally (a "row vector") or vertically (a "column vector"). A square matrix is a matrix that has the same number of rows and columns; all correlation matrices are square matrices.

Each square matrix has a set of numbers associated with it called its eigenvalues that further characterize the matrix. Eigenvalues have different interpretations depending on the matrix (e.g., if it is a correlation matrix or some other type of matrix) and the field of study; they can be interpreted geometrically, whereby a matrix is related to an object in space, and they can be interpreted algebraically, whereby they relate to how the matrix changes when it is multiplied by itself many, many times. We will discuss eigenvalues only as they relate to correlation matrices, as they tend to be used in social science data analysis.

To provide an example of eigenvalues, consider the correlation matrix shown in Table 1.2 of girls' intelligence across development. The eigenvalues of this correlation matrix (there are 10, the same as the number of variables) are 7.63, 0.64, 0.41, 0.38, 0.25, 0.22, 0.18, 0.12, 0.10, and 0.07. The correlation matrix describing racial composition of cities and their corresponding NBA teams (Table 1.5) has four eigenvalues: 2.18, 1.32, 0.49, and 0.01. Eigenvalues cannot be calculated from a single correlation, or a subset of the correlations in a correlation matrix. The entire matrix is needed to determine what the eigenvalues of a correlation matrix are; they are determined using an eigenvalue formula.

In this chapter, we will describe why eigenvalues are an essential part of understanding correlation matrices. We will also refer to the practical significance of eigenvalues for methods to test null hypotheses about eigenvalues (Chapter 3) and to analyze correlation matrices (Chapter 4). We cover eigenvalues only as they are relevant to an applied researcher who uses correlation matrices. We skip the details of the calculation of eigenvalues, leaving those for linear algebra textbooks and software systems such as R, SAS, SPSS, and Stata. Math packages or online utilities will also readily compute the eigenvalues for a given matrix. Although formulas are basic arithmetic, the eigenvalue formula for a large correlation matrix (even a 10 × 10 correlation matrix can be considered fairly large in this context) requires a great deal of computational effort. Thus, computer software is virtually always used to compute eigenvalues. In this chapter, we will also present a summary of several other important mathematical features of a correlation matrix (some of these as they relate to eigenvalues), as well as notation that will be used throughout the rest of the book.

Requirements of Correlation Matrices

To understand a correlation matrix, it helps to start with the correlation coefficients themselves, which collectively define a correlation matrix. Correlations among a set of variables (e.g., $X_1, X_2, \ldots, X_p$) are typically summarized in a correlation matrix, which we will generically call $\boldsymbol{R}$. $\boldsymbol{R}$ will always be a square matrix of order p ("square" meaning that the matrix has the same number of rows as it has columns, and "order" here refers to the number of rows and columns, i.e., the number of variables); the rows and columns of $\boldsymbol{R}$ indicate the same p variables, those being correlated, and the entries in $\boldsymbol{R}$, r_{ij}, are correlation coefficients between pairs of variables X_i and X_j. For example, if the element of $\boldsymbol{R}$ in the fourth row and seventh column is $r_{47} = .24$, that would indicate that the correlation between variables X_4 and X_7 is .24. Note that, because r_{47} necessarily equals r_{74} (i.e., the correlation coefficient is a symmetric measure), the correlation matrix is symmetric.

The upper and lower triangles of a correlation matrix are defined in relation to the diagonal elements—that is, the elements in the r_{ii} positions in the matrix from the upper left to the lower right of the matrix. (Note that all diagonal elements, r_{ii}, in correlation matrices equal 1.0, because a variable correlates perfectly with itself.) The upper triangle consists of all elements in a correlation matrix that are above the diagonal; the lower triangle consists of all elements below the diagonal.

Elements r_{ij} in a correlation matrix—the correlations themselves—are constrained in the following four ways. Any researcher who has seen a correlation matrix, or studied the basic correlation coefficient, is likely familiar with the first three requirements:

1. $r_{ij} = r_{ji}$ (i.e., $\boldsymbol{R}$ is symmetric; the correlation between X_i and X_j is the same as the correlation between X_j and X_i)
2. $r_{ij} = 1$ if $i = j$ (i.e., the diagonal elements of $\boldsymbol{R}$ are 1; the correlation of a variable with itself is 1)
3. $-1 \leq r_{ij} \leq 1$ if $i \neq j$ (i.e., the off-diagonals of $\boldsymbol{R}$ are correlation coefficients bounded inside the interval between −1 and +1)

It is important to emphasize that these features of the correlation coefficient are simply mathematical properties of the formula by which the Pearson (and other) correlation coefficients are computed. Algebraic treatment that is relatively simple (but which we do not present here) exists to show that each of these properties is necessarily a feature of the correlation coefficient itself and, therefore, of all elements of a correlation matrix.

There is one final requirement for a correlation matrix that involves eigenvalues. The eigenvalues are related to the variances of the variables on which the correlation matrix is based; that is, the p eigenvalues are related to the variances of the p variables. True variances must be nonnegative, because they are computed from sums of squares, which themselves are each nonnegative. Thus, the final requirement for a correlation matrix is a check on its eigenvalues. Specifically, the fourth (and final) requirement is that all the eigenvalues of the correlation matrix are nonnegative:

4. $\lambda_1, \lambda_2, \dots \lambda_p \geq 0$ where λ_i, $i = 1, 2, \dots, p$ are the eigenvalues of $\boldsymbol{R}$.

We will discuss eigenvalues of a matrix in conceptual detail in the next section. Here, it is worth mentioning that this fourth requirement is often framed in terms of another characteristic of matrices, the matrix determinant, rather than eigenvalues. (Specifically, for readers with some matrix algebra background, this fourth property is equivalently ensured if the determinant of $\boldsymbol{R}$ and all principal minor submatrices of $\boldsymbol{R}$ are nonnegative.) We prefer to frame the fourth property in terms of eigenvalues, because eigenvalues may be calculated directly from $\boldsymbol{R}$; the determinant equivalency requires the calculation of determinants from an increasingly large group of submatrices of $\boldsymbol{R}$ as p gets large.

Eigenvalues of a Correlation Matrix

Every square matrix has a set of eigenvalues and an associated set of eigenvectors. These are defined by mathematical definition, using specific formulas that can be found in any linear algebra text, or online (but which we do not present in any detail in this book). The eigenvalues and eigenvectors of a matrix are linked—each eigenvalue has a corresponding eigenvector, and vice versa. If a square matrix is of order p (i.e., p rows and columns), then the matrix has p eigenvalues and p eigenvectors. There may be repeating values among this set of eigenvalues, but the number of eigenvalues, with duplications, will still be p. Furthermore, the sum of the eigenvalues is equal to the sum of the diagonal elements of the matrix. Therefore, in the case of correlation matrices, in which the diagonal elements all equal 1 (and therefore the sum of the diagonal elements is p), the sum of the eigenvalues for the correlation matrix will also equal p. As examples of this property, in the fourth paragraph of this chapter where two sets of eigenvalues are presented, it is easy to verify that the first set of eigenvalues, from a 10 × 10 matrix, add to 10; the second set of eigenvalues, from a 4 × 4 matrix, add to 4.

Eigenvalues and eigenvectors are frequently invoked in fields that use statistical analysis. Readers don't need to have deep understanding and appreciation of these mathematical terms to use correlation matrices, but some conceptual understanding is useful to explain why some matrices that appear to be correlation matrices are not correlation matrices. Furthermore, and of more substantive interest, eigenvalues and eigenvectors have geometric interpretations that allow researchers to reduce complex information into simpler summaries. For example, facial recognition researchers use eigenvalues and eigenvectors to summarize similarities between many faces into a much smaller set of "eigenfaces"; audio recognition researchers can construct similar "eigenvoices" to break down complex speech into simpler dimensions. Intelligence researchers often summarize the information in a whole battery of instruments measuring human abilities by using factor analysis to smooth out the redundancies (i.e., overlapping variance) across the many different measures. Methods such as PCA and factor analysis rely on eigenvalues and eigenvectors to develop component and factor models of a set of variables.

In the previous section, we indicated that p eigenvalues are related to variances that underlie the correlation matrix. More specifically, eigenvalues, relative to p, are measures that are related to proportions of variance. In a correlation matrix with one or a few large eigenvalues, relative to p, substantial redundancy is indicated among the variables; that is, many of the variables share a great deal of variance and thus map into a central construct or dimension (technically, these dimensions are often called "principal components"). For example, a correlation matrix of order 4 may have eigenvalues 2.8, 0.9, 0.2, and 0.1 (note that these four eigenvalues sum to four, as required). The presence of the relatively large first eigenvalue of 2.8 indicates that the variables share substantial common variance—roughly 2.8/4 = 0.7, or 70% of the variance among all the variables may be expressed with a single linear combination of the four variables. Next, 0.9/4 = 23% of the variance is accounted for by a second dimension—this second dimension is constructed to be unrelated (uncorrelated) to the first dimension. Thus, underlying the four variables with these eigenvalues is one dominant dimension and a second uncorrelated, less dominant, dimension, with very little variance accounted for by the third and fourth dimensions (around 7%), which means that these four variables can be (almost completely) summarized by two dimensions (or components, or factors).

Pseudo-Correlation Matrices and Positive Definite Matrices

The constraint on correlation matrices that all eigenvalues must be nonnegative occurs because eigenvalues are related to variances. As noted in the examples above, a given eigenvalue divided by the sum of all eigenvalues gives the proportion of variance associated with the particular direction or dimension defined by the associated eigenvector. The presence of a negative eigenvalue would therefore indicate a negative proportion of variance, which is conceptually and mathematically intractable for statistical settings. However, it is not unusual that a matrix may *look* like a correlation matrix because each element of the apparent correlation matrix meets the first three requirements of a correlation matrix, but the overall matrix fails to meet the fourth requirement of nonnegative eigenvalues. We call such matrices pseudo-correlation matrices.

For example, consider the (apparent) correlation matrix presented in Table 1.6. One would not be able to tell upon naive inspection that this correlation matrix—constructed using real data and modified slightly[1]—is in

[1] Only one correlation was altered from the correlation matrix computed from real data using polychoric correlations: The correlation between ideal and expected children in 1979 was increased from .756 to .876.

fact a pseudo-correlation matrix. The eigenvalues of this correlation matrix are 2.26, 1.59, 0.65, 0.50, and −0.0049. Because the correlation matrix has a negative eigenvalue, it is not a true correlation matrix.

Matrices that do satisfy all four requirements are called true correlation matrices. All true correlation matrices have nonnegative eigenvalues; in the language of matrix algebra, these are referred to as positive semidefinite (PSD) matrices. Correlation matrices that have strictly positive (i.e., no negative or zero) eigenvalues are positive definite (PD) matrices. Pseudo-correlation matrices are referred to in this language as indefinite matrices, indicating the presence of at least one negative and one positive eigenvalue.

If a correlation matrix has one or more eigenvalues that are exactly 0, this/these eigenvalues correspond to directions or dimensions (related to the corresponding eigenvectors) that explain zero proportion of the variance in the original variables. This circumstance may happen in practice if there is linear dependence among the variables in the correlation matrix. For example, a researcher may unintentionally create one variable that is a linear combination of one or more of the other variables, such as by including as variables in the correlation matrix both the total score and the individual items (which are summed to create the total score) in the same research setting. Correlation matrices that have one or more zero eigenvalues, even though a true correlation matrix, are problematic for most statistical software, and the researcher who tries to analyze such a correlation matrix may receive an error message from the computer program. In such cases, the researcher can check the data to ensure that they were entered correctly, or the researcher may be able to identify one or more variables that caused a linear dependence and that can be removed from the analysis.

Pseudo-correlation matrices are not just of theoretical interest; researchers often and regularly may have to diagnose and deal with such matrices. How do pseudo-correlation matrices exist? A pseudo-correlation matrix may look like a true correlation matrix, but there does not exist a set of complete quantitative variables to which the Pearson correlation formula can be applied in a pairwise fashion to produce the matrix. A pseudo-correlation matrix may arise from real data under one of several conditions. First, if there exist missing data among the variables, a correlation matrix created using pairwise complete cases (i.e., computed from correlations between pairs of variables, but because of missing data patterns using different subsets of the observations) may have one or more negative eigenvalues. The correlation matrix based on complete cases of numeric data using the Pearson product–moment correlation formula will necessarily be a true correlation matrix, but many researchers calculate correlations using as many observations as possible for each pair of variables, resulting in correlations within a correlation matrix based on different sample sizes and based on different subsets of the total data set, which can lead to pseudo-correlation matrices.

Second, a pseudo-correlation matrix may occur if the variables used to construct the correlation matrix are not numeric/quantitative but are rather binary or ordinal, in which case a researcher may choose to use polychoric or tetrachoric correlation formulas to form the correlation matrix. Polychoric and tetrachoric correlations are calculated by assuming that the binary/ordinal variables that are being correlated are attempting to measure traits that are inherently normally distributed; although these types of correlations are recommended in many research settings, as the assumption of underlying normality is often reasonable, an entire correlation matrix populated by polychoric or tetrachoric correlations may not be PSD and may be a pseudo-correlation matrix.

Third, if the correlation matrix is the result of averaging more than one correlation matrix (such as may be done in two-stage meta-analysis), then there is no guarantee that the resulting correlation matrix is PSD. In all of these cases, if the correlation matrix is in actuality a pseudo-correlation matrix, warnings or errors are likely to be generated by the statistical software system used to analyze the correlation matrix. The correlation matrix presented in Table 1.6, for example, is a pseudo-correlation matrix, and trying to analyze it will likely cause an error message in software.

Although problematic in substantive research settings, pseudo-correlation matrices can inform quantitative methods. Recent work has focused on using pseudo-correlations to provide insight into a true correlation matrix of interest (Waller, 2016). Other work has dealt with statistical issues surrounding pseudo-correlation matrices in real-data settings (Bentler & Yuan, 2011; Higham, 2002), particularly with large correlation matrices where the relative risk of a correlation matrix being non-PSD is greater.

Smoothing Techniques

In cases where a pseudo-correlation matrix did not arise from error, and the researcher does not wish to remove variables or alter how the correlation matrix was calculated to amend the non-PSD matrix, smoothing techniques are recommended before proceeding with statistical analyses. The goal of a smoothing technique is to produce a true correlation matrix that closely approximates the pseudo-correlation matrix. Generally, programs that implement smoothing techniques take as input the pseudo-correlation matrix and allow the user to indicate their tolerance for how much change is allowed to smooth the pseudo-correlation matrix into a true correlation matrix. The program will then output a smoothed, PD true correlation matrix that is “close” to the provided pseudo-correlation matrix.

There are a variety of smoothing techniques that can be broadly sorted into three categories: (1) shrinking techniques, (2) simultaneous-variable techniques, and (3) single-variable techniques. Shrinking techniques simply reduce the magnitude of all correlations toward zero; after a sufficient amount of shrinking, the correlation matrix typically will be PSD or strictly PD. Simultaneous-variable techniques seek the true correlation matrix that minimizes the “distance” to the pseudo-correlation matrix, using a variety of definitions for how “distance” is measured. Most smoothing techniques are simultaneous-variable techniques, and several of these have been proposed (e.g., Rousseeuw & Molenberghs, 1993); however, techniques performed with comparable tolerance levels will provide similar smoothed correlation matrices (Kracht & Waller, 2018). Table 2.1 demonstrates two different simultaneous-variable smoothing techniques for the correlation matrix in Table 1.6. Both techniques were implemented in the free software program R and are functions in popular R packages. Finally, single-variable techniques focus on only shrinking rows and columns of the correlation matrix associated with one or a few “problem” variables (e.g., Bentler & Yuan, 2011).

Table 2.1 Two Smoothed, PD Correlation Matrices Calculated From the Non-PSD Matrix in Table 1.6 Using the Software Package R

Variable	*1*	*2*	*3*	*4*	*5*
1. Ideal number of children (1979)	1.000	***.763***	−.053	**.473**	.121
2. Expected number of children (1979)	***.762***	1.000	**−.425**	**.374**	**.010**
3. Number of children (1980)	**−.053**	**−.425**	1.000	.118	**.429**
4. Ideal number of children (1982)	**.473**	**.374**	.118	1.000	.207
5. Number of children (2004)	.121	**.010**	**.429**	.207	1.000

Note: Correlations above the diagonal were smoothed with the *nearPD()* function in the Matrix package, and correlations below the diagonal were smoothed with the *cor.smooth()* function in the psych package. The eigenvalues of both true smoothed correlation matrices are the same to two decimals: 2.15, 1.57, .65, .49, and .14. The smoothed matrices are identical to the third decimal except for one correlation (italicized). The two smoothing algorithms were implemented such that the smallest eigenvalue between the two procedures would be comparable. Correlations that have changed in magnitude by more than .10 from the corresponding element in the non-PSD matrix are bolded. PD = positive definite; PSD = positive semidefinite.

Single-variable techniques leave larger portions of the non-PSD correlation matrix unchanged but often result in greater change to the affected rows and columns compared with the simultaneous-variable methods.

Restriction of Correlation Ranges in the Matrix

We reiterate that a correlation matrix is not just a matrix filled with correlations. Not every set of correlations, arbitrarily inscribed symmetrically in the $p \times p$ matrix, will produce a true correlation matrix. Once one correlation coefficient in the matrix is known, or fixed, then other correlations in the matrix are constrained, or bounded, if a true correlation matrix is to be produced. Stanley and Wang (1969) derived the formula for the simple 3 × 3 correlation matrix case showing how fixing two of the correlation coefficients constrains the third correlation coefficient. That is, if a researcher has three variables of interest—say X_1, X_2, and X_3—and the values of r_{12} and r_{13} are known, the range of possible values for r_{23} can be mathematically derived and will generally be much tighter than the range of $[-1, +1]$. Hubert (1972) extended this formula for any number of variables. In Chapter 6, we show how this restriction of correlation range can be directly observed using the geometric representation of the set of all pseudo-correlation matrices.

The Inverse of a Correlation Matrix

Another characteristic of a matrix is its inverse. The inverse of a matrix is conceptually similar to the reciprocal of a scalar, the types of numbers that we deal with on a regular basis. For example, the scalars 8 and 36.9 have as reciprocals $1/8 = .125$, and $1/36.9 = .\overline{02710}$, respectively. Reciprocals or scalar inverses are the numbers that, when multiplied by the original number, produce the identity, 1.0. The identity scalar, 1.0, is the number that, when multiplied by any other number, returns the original number.

Although many scalar operations have equivalent operations on matrices (e.g., matrices of "matching" or conformable sizes can be added, subtracted, or multiplied), there is no matrix operation for division. For scalars a and b, you can simply calculate $\frac{a}{b}$, unless b is zero, in which case the ratio is mathematically impossible to calculate; for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, it is impossible to calculate $\frac{\boldsymbol{A}}{\boldsymbol{B}}$, much like it is impossible to divide a scalar by 0. However, as for scalars, matrix inverses can be multiplied by other matrices as a substitute for division. For scalars, multiplying $a * \frac{1}{b}$ or $a * b^{-1}$, where $\frac{1}{b}$

or b^{-1} is the reciprocal of b, is the same as calculating $\frac{a}{b}$. For matrices, it may be possible to calculate $\boldsymbol{AB}^{-1}$, where $\boldsymbol{B}^{-1}$ is the inverse of matrix $\boldsymbol{B}$, even though $\frac{\boldsymbol{A}}{\boldsymbol{B}}$ can never be calculated.

There is a matrix identity that, when multiplied by another matrix, returns that original matrix, and we can define the matrix inverse as the matrix that, when multiplied by the original matrix, will equal the identity. The details of actually computing matrix inverses are not important to studying correlation matrices, but it is important to know that in the computations used to do statistical analysis, the inverse of the correlation matrix ($\boldsymbol{R}^{-1}$) is often used rather than $\boldsymbol{R}$ itself. $\boldsymbol{R}^{-1}$ has direct interpretations in advanced statistical methods such as multiple regression, factor analysis, and discriminant analysis (Raveh, 1985) and, sometimes, also serves as a weight matrix in analysis. However, just as the scalar number zero has no reciprocal, certain matrices also do not have inverses, including, for example, pseudo-correlation matrices and PSD correlation matrices. Among correlation matrices, only true, strictly PD correlation matrices have inverses—which explains why many statistical programs will return an error message if the researcher tries to analyze a correlation matrix with one or more zero or negative eigenvalues. A cryptic message that "the correlation/covariance matrix cannot be inverted," or equivalently, "the correlation/covariance matrix is not full rank," is referencing the absence of a valid inverse for the correlation matrix.

The Determinant of a Correlation Matrix

The final characteristic of a matrix we find relevant to (briefly) discuss in this book is the determinant of a matrix. All square matrices have a determinant (denoted as $|\boldsymbol{R}|$ for a given correlation matrix $\boldsymbol{R}$), which is a single number equal to the product of all of the eigenvalues of the matrix. Computer programs can readily calculate the determinant of a matrix, along with the eigenvalues and eigenvectors. Although inspecting the p eigenvalues of $\boldsymbol{R}$ is often useful, the determinant can provide some quick diagnosis for issues about the correlation matrix. For example, if $|\boldsymbol{R}| < 0$, then one or more of the eigenvalues of $\boldsymbol{R}$ is negative, and $\boldsymbol{R}$ is therefore a pseudo-correlation matrix. If $|\boldsymbol{R}| = 0$, then one or more of the eigenvalues of $\boldsymbol{R}$ is equal to 0, and there is linear dependence among the variables of the correlation matrix. Finally, $|\boldsymbol{R}| > 0$ for true correlation matrices that are PD and have eigenvalues that are strictly positive (although in some fairly unusual cases a pseudo-correlation matrix may have $|\boldsymbol{R}| > 0$, such as if an even number of eigenvalues are negative). We discuss determinants primarily because they appear in some test statistics for null hypotheses on correlation matrices, which we discuss in Chapter 3.

Examples

Racial Composition of NBA and Sponsor Cities

The correlation matrix in Table 1.5 has four variables. For this correlation matrix, X_1 = number of Black teammates in 1983, X_2 = number of Black teammates in 1989, X_3 = percent Black of city residents in 1980, and X_4 = percent Black of city residents in 1990. This correlation matrix is obviously of order 4 ($p = 4$). Each entry in the correlation matrix is between [−1, 1], and each element is a correlation coefficient. For example, $r_{12} = .41$ indicates that the correlation between the number of Black teammates on NBA teams between 1983 and 1989 is .41, or positively related at a moderate level. The value $r_{24} = .29$ indicates that there is also a positive (but weaker) relationship between the percentage of Black city residents in 1990 and the number of Black teammates on that city's NBA team.

The eigenvalues of this correlation matrix are 2.18, 1.33, 0.49, and 0.01 (which sum to 4 within rounding error). All the eigenvalues are positive, and so this matrix is strictly PD, and is therefore a true correlation matrix. The first eigenvalue, 2.18, is linked to an eigenvector that corresponds to a dimension accounting for 2.18/4 = .55, or about 55% of the total variance in the correlation matrix. The second eigenvalue corresponds to an eigenvector associated with an uncorrelated dimension that accounts for an additional 1.33/4 = .32 proportion of variance (and which is constrained to be uncorrelated with the first dimension). Therefore, two uncorrelated underlying dimensions corresponding to the first two eigenvectors have eigenvalues large enough to indicate that these two dimensions account for about 87% of the total variance among the four variables.

Girls' Intelligence Across Development

The correlation matrix in Table 1.2 has 10 variables—girls' intelligence measured each year from ages 8 to 17. The correlation matrix is obviously of order 10 ($p = 10$). The eigenvalues of this correlation matrix are 7.63, 0.64, 0.41, 0.38, 0.25, 0.22, 0.18, 0.12, 0.10, and 0.07 (which sum to 10 within rounding error). Because all the eigenvalues are positive, this correlation matrix is strictly PD and is, therefore, a true correlation matrix. The first eigenvalue is relatively large compared with the other eigenvalues (7.63/10 = 0.763), indicating that the first dimension is associated with a large portion of the variance (around 76%) among intelligence scores measured between ages 8 and 17.

Summary

Certain characteristics of correlation matrices are important for statistical applications. Correlation matrices contain correlations, but not all matrices that contain correlations are true correlation matrices. Those that appear to be correlation matrices by virtue of containing correlations, but are not true correlation matrices, are called pseudo-correlation matrices. How can we tell them apart? Pseudo-correlation matrices can be diagnosed by computing the eigenvalues that correspond to a particular correlation matrix (many software routines, or online computational applications, can be used to compute eigenvalues). True correlation matrices have eigenvalues that are only positive and/or zero. Pseudo-correlation matrices have at least one eigenvalue that is negative. Matrices with only positive eigenvalues are called PD matrices. Matrices whose eigenvalues are positive and/or zero are called PSD matrices. Matrices with at least one positive and one negative eigenvalue are called indefinite matrices; all pseudo-correlation matrices are indefinite. Smoothing techniques are algorithms that replace a pseudo-correlation matrix with the closest true correlation matrix, with "closest" defined differently across different techniques. In addition to eigenvalues, other important characteristics of the correlation matrix include its determinant and its inverse, both of which appear in statistical tests on correlation matrices described in the next chapter.

In the next chapter, we discuss a number of statistical procedures that have been developed to analyze correlation matrices. The material in the current chapter informs those analytic methods, because most of those approaches cannot be applied to pseudo-correlation matrices. The reader will see eigenvalues discussed in the next chapter (and also later in the book) and should by now be aware of their substantial value as diagnostic indices that reveal important features of both true and pseudo-correlation matrices.