# Important Matrices for Multivariate Analysis

There are several different matrices that will be important for traditional multivariate analysis. We will list them here. The formulas for computing the various matrices are much less important than the names and the meaning of the matrices and the matrix elements.

## The Data Matrix

The most important matrix for any statistical procedure is the data matrix. The observations form the rows of the data matrix and the variables form the columns. The most important requirement for the data matrix is that the rows of the matrix should be *statistically independent*. That is, if we pick any single row of the data matrix, then we should not be able to predict any other row in the matrix. Practically speaking, statistical independence is guaranteed when each row of the matrix is an independent observation.

To illustrate the data matrix and the other important matrices in this section, let us consider a simple example. Sister Sal of the Benevolent Beatific Bounty of Saints Boniface and Bridget was the sixth grade teacher at The Most Sacred Kidney of Saint Zepherinus School. During her tenure there Sister Sal not only kept track of the students' grades but also wrote down her own rating of the chances that a student will eventually grow up to become an ax murderer. Below is a table of five of Sister Sal's students, their age in sixth grade, Sister Sal's ax murder rating, and their scores as adults on the Psychopathic-deviate scale on the Minnesota Multiphasic Personality Inventory (MMPI Pd).

Table 1.1. Follow up of Sister Sal's sixth grade class.

| Student | Age | Rating | MMPI Pd |
|---|---|---|---|
| Abernathy | 10 | 3 | 38 |
| Beulah | 12 | 4 | 34 |
| Cutworth | 20 | 10 | 74 |
| Dinwitty | 10 | 1 | 40 |
| Euthanasia | 8 | 7 | 64 |

The data matrix would look like this:

$$\mathbf{X} = \begin{bmatrix} 10 & 3 & 38 \\ 12 & 4 & 34 \\ 20 & 10 & 74 \\ 10 & 1 & 40 \\ 8 & 7 & 64 \end{bmatrix}.$$

# Uncorrected Sums of Squares and Cross Products Matrix (USSCP)

The definition of an uncorrected sums of squares and cross products matrix is simply the data matrix premultiplied by its transpose. If **X** is the data matrix, then

$$\text{USSCP} = \mathbf{X}^t\mathbf{X}.$$

In the Sister Sal example, the USSCP matrix would equal

$$\text{USSCP} = \begin{pmatrix} 10 & 12 & 20 & 10 & 8 \\ 3 & 4 & 10 & 1 & 7 \\ 38 & 34 & 74 & 40 & 64 \end{pmatrix} \begin{pmatrix} \mathbf{10} & \mathbf{3} & \mathbf{38} \\ \mathbf{12} & 4 & \mathbf{34} \\ \mathbf{20} & \mathbf{10} & \mathbf{74} \\ \mathbf{10} & \mathbf{1} & \mathbf{40} \\ \mathbf{8} & \mathbf{7} & \mathbf{64} \end{pmatrix} = \begin{pmatrix} 808 & 344 & 3180 \\ 344 & 175 & 1478 \\ 3180 & 1478 & 13772 \end{pmatrix}$$

Note how the diagonals of the USSCP matrix are "sums of squares." For example, the first diagonal element equals $10^2 + 12^2 + 20^2 + 10^2 + 8^2 = 808$. Note also how the off diagonals are the "sums of cross products," the cross products in this case being a person's score on one variable multiplied by their score on another variable. For example, element (1,2) in the USSCP matrix is 10*3 + 12*4 + 20*10 + 10*1 + 8*7 = 344.

# Corrected Sums of Squares and Cross Products (CSSCP)

The corrected sums of squares and cross products matrix (CSSCP) is derived from a data matrix in which each score is expressed as a deviation from its mean. In the Sister Sal example, the means for the three variables are 12, 5, and 50. If we subtract these means from each data point we have a new matrix which we will call **D**.

$$\mathbf{D} = \begin{pmatrix} \mathbf{10}-12 & \mathbf{3}-5 & \mathbf{38}-50 \\ \mathbf{12}-12 & 4-5 & \mathbf{34}-50 \\ \mathbf{20}-12 & \mathbf{10}-5 & \mathbf{74}-50 \\ \mathbf{10}-12 & \mathbf{1}-5 & \mathbf{40}-50 \\ \mathbf{8}-12 & \mathbf{7}-5 & \mathbf{64}-50 \end{pmatrix} = \begin{pmatrix} -2 & -2 & -12 \\ 0 & -1 & -16 \\ 8 & 5 & 24 \\ -2 & -4 & -10 \\ -4 & 2 & 14 \end{pmatrix}.$$

The CSSCP matrix is then defined as

$$\text{CSSCP} = \mathbf{D}^t\mathbf{D}.$$

In the present example, this matrix is

$$\text{CSSCP} = \begin{pmatrix} -2 & 0 & 8 & -2 & -4 \\ -2 & -1 & 5 & -4 & 2 \\ -12 & -16 & 24 & -10 & 14 \end{pmatrix} \begin{pmatrix} -2 & -2 & -12 \\ 0 & -1 & -16 \\ 8 & 5 & 24 \\ -2 & -4 & -10 \\ -4 & 2 & 14 \end{pmatrix} = \begin{pmatrix} 88 & 44 & 180 \\ 44 & 50 & 228 \\ 180 & 228 & 1272 \end{pmatrix}.$$

You should verify that the diagonals of this matrix are indeed sums of squares and that the off diagonals are cross products.

There is a second formula that expresses the CSSCP matrix in terms of the raw data matrix. This formula is used mostly for computational reasons. Let $\bar{\mathbf{x}}$ denote a column vector of means and let $N$ denote sample size. Then

$$\text{CSSCP} = \mathbf{X}^t\mathbf{X} - N\bar{\mathbf{x}}\bar{\mathbf{x}}^t.$$

For the present example,

$$\text{CSSCP} = \begin{pmatrix} 3 & 4 & 10 & 1 & 7 \\ 38 & 34 & 74 & 40 & 64 \end{pmatrix} \begin{pmatrix} 10 & 3 & 38 \\ 12 & 4 & 34 \\ 20 & 10 & 74 \\ 10 & 1 & 40 \\ 8 & 7 & 64 \end{pmatrix} - 4 \begin{pmatrix} 12 \\ 5 \\ 50 \end{pmatrix} \begin{pmatrix} 12 & 5 & 50 \end{pmatrix}.$$

## Covariance Matrix

A covariance matrix is a symmetric matrix where each diagonal element equals the variance of a variable and each diagonal element is the covariance between the row variable and the column variable.

The definition of the variance for variable $X$ is

$$V_X = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}.$$

The definition of a covariance between two variables, $X$ and $Y$, is

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}.$$

You should verify that the covariance of a variable with itself equals the variance of the variable.

A covariance is a statistic that measures the extent to which two variables "vary together" or 'covary." Covariances have two properties--magnitude and sign. Covariances that are close to 0, relative to the scale of measurement of the two variables, imply that the two variables are not related--i.e. one cannot predict scores on one variable by knowing scores on the other variable. Covariances that are large (either positive large or negative large) relative to the measurement scale of the variables indicate that the variables are related. In this case, one can predict scores on one variable from knowledge of scores on another variable.

The sign of a covariance denotes the direction of the relationship. A positive covariance signifies a direct relationship. Here high scores on one variable are associated with high scores on the other variable, and conversely low scores on one variable are associated with low scores on the other variable. A negative covariance denotes an inverse relationship. Here, high scores on one variable predict low scores on the other variable, and conversely low scores on the first variable are associated with high scores on the second variable. The covariance between amount of time spent studying and grades is positive while the covariance between amount of time spent partying and grades would be negative.

In matrix terms, a covariance matrix equals the corrected sums of squares and cross products matrix in which each element is divided by ($N$ - 1). Let **C** denote the covariance matrix. Then

$$\mathbf{C} = \text{CSSCP}\left(\frac{1}{N-1}\right) = \mathbf{D}^t\mathbf{D}\left(\frac{1}{N-1}\right).$$

For the present example,

$$\mathbf{C} = \begin{pmatrix} 88 & 44 & 180 \\ 44 & 50 & 228 \\ 180 & 228 & 1272 \end{pmatrix} \div 4 = \begin{pmatrix} 22 & 11 & 45 \\ 11 & 12.5 & 57 \\ 45 & 57 & 318 \end{pmatrix}.$$

## Correlation Matrix

A correlation matrix is a special type of covariance matrix. A correlation matrix is a covariance matrix that has been calculated on variables that have previously been standardized to have a mean of 0 and a standard deviation of 1.0. Many texts refer to variables standardized in this way as $Z$ scores.

The generic formula for a correlation coefficient between variables $X$ and $Y$ is

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{s_X s_Y}$$

where $s_X$ and $s_Y$ are, respectively, the standard deviations of variables $X$ and $Y$. Because a correlation is a specific form of a covariance, it has the same two properties--magnitude and sign--as a covariance. The sign indicates the direction of the relationship. Positive correlations imply a direct relationship, and negative correlations imply an inverse relationship. Similarly, correlations close to 0 denote no statistical association or predictability between the two variables. Correlations that deviate from 0 in either direction (positive or negative) indicate stronger statistical associations and predictability.

The correlation coefficient has one important property that distinguishes it from other types of covariances. The correlation coefficient has a mathematical lower boundary of -1.0 and an upper bound of 1.0. This property permits correlation coefficients to be compared, while ordinary covariances usually cannot be compared. For example, if $X$ and $Y$ correlate .86 while $X$ and $Z$ correlate .32, then we can conclude that $X$ is more strongly related to $Y$ than to $Z$. However, if variables $A$ and $B$ have a covariance of 103.6 while $A$ and $C$ have a covariance of 12.8, then we cannot conclude anything about the magnitude of the relationship. The reason is that the magnitude of a covariance depends upon the measurement scale of the variables. If the measurement scale for variable $C$ has a much lower variance than that for variable $B$, then $A$ might actually be more strongly related to $C$ than to $B$. The correlation coefficient avoids this interpretive problem by placing all variables on the same measurement scale--the $Z$ score with a mean of 0 and a standard deviation of 1.0.

The formula for a correlation matrix may also be written in matrix algebra. Let **S** denote a diagonal matrix of standard deviations. That is, the standard deviation for the first variable in on the first diagonal element, that for the second variable is the second diagonal element, and so on. All off diagonals elements are 0. Matrix **S** may be easily computed from the covariance matrix, **C**, by letting

taking the square root of the diagonal elements and setting all off diagonal elements to 0. For the present example,

$$\mathbf{S} = \begin{pmatrix} \sqrt{22} & 0 & 0 \\ 0 & \sqrt{12.5} & 0 \\ 0 & 0 & \sqrt{318} \end{pmatrix} = \begin{pmatrix} 4.69 & 0 & 0 \\ 0 & 3.54 & 0 \\ 0 & 0 & 17.83 \end{pmatrix}.$$

Let **R** denote the correlation matrix. Then the general formula for **R** is

$$\mathbf{R} = \mathbf{S}^{-1}\mathbf{C}\mathbf{S}^{-1}.$$

Although we do not need to know how to compute an inverse, the inverse of a diagonal matrix is quite easy to calculate--simply take the inverse of each diagonal element. For example,

$$\mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{4.69} & 0 & 0 \\ 0 & \frac{1}{3.54} & 0 \\ 0 & 0 & \frac{1}{17.83} \end{pmatrix}.$$

Consequently, the correlation matrix for the Sister Sal data is

$$\mathbf{R} = \begin{pmatrix} \frac{1}{4.69} & 0 & 0 \\ 0 & \frac{1}{3.54} & 0 \\ 0 & 0 & \frac{1}{17.83} \end{pmatrix} \begin{pmatrix} 22 & 11 & 45 \\ 11 & 12.5 & 57 \\ 45 & 57 & 318 \end{pmatrix} \begin{pmatrix} \frac{1}{4.69} & 0 & 0 \\ 0 & \frac{1}{3.54} & 0 \\ 0 & 0 & \frac{1}{17.83} \end{pmatrix}$$

$$= \begin{pmatrix} 1.000 & .663 & .538 \\ .663 & 1.000 & .904 \\ .538 & .904 & 1.000 \end{pmatrix}.$$

## The Good News: Read This!

If we analyze only two variables, we can visualize the data quite well by constructing a scatterplot. If we analyze three variables, we could also plot the data, but in this case the plot would be in three dimensions. We can visualize this by imaginging that each data point is a star suspended in space in a room. In short, if we analyze *p* variables then we are geometrically dealing with points in a *p* dimensional space. Of course, we cannot visualize dimensions higher than three, but we can still deal with such *hyperspace* using mathematics and geometry.

The key point in understanding classic multivariate analyses is to realize that the classic techniques try to *summarize p dimensional space in terms of three types of matrixes*: (1) matrices of means; (2) diagonal matrices of standard deviations; and (3) correlation matrices. (The only additional information used in some multivariate analyses is sample and group size.) For example, the analysis of just one variable measured on, say, 734 individuals would involve placing 734 dots on top of a straight number line. We would try to summarize the information about all 734 dots into two statistics, the mean and the standard deviation. The mean would tell us about the general location of the dots, and the

standard deviation would inform us about how spread out the dots are around the mean.

Because classic multivariate analysis involves three types of matrices, it is important for us to take time to reflect on the meaning of these three matrices. Let us do that now.

A matrix or vector of means tell us where variables are located along the number lines in a multidimensional space. For example, a vector consisting of two means tell us where the "dots" in a scatterplot are centered, and a vector of three means tells us where the dots in a three dimensional space are centered. In techniques such as MANOVA, the multivariate analysis of variance, we will often compare a vector of means for one group to a vector from another group. Effectively this comparison is equal to asking whether the "dots" for one group are centered in the same place as the "dots" for the other group.

The matrix of standard deviations is a measure of the extent to which the dots in space are spread out around their center. If variable *X* has a standard deviation of 12, then we should expect a number of "dots" within + or - 24 units of the center of *X* and relatively few dots beyond +24 or -24 units away from the center of *X*. In many cases, it is convenient to think of standard deviations as "scaling factors" for the variables, analogous to currency conversions. For example, if variable *X* has a standard deviation of 2 and variable *Y* has a standard deviation of 5, then one unit of *X* is "worth" .4 units of *Y* and one unit of *Y* is "worth" 2.5 units of *X*.

Finally, the correlation matrix expresses the geometric *shape* of the dots in hyperspace when each variable is measured on the same scale (i.e., each variable has a standard deviation of 1.0). Specifically, the correlation matrix informs us about the extent to which the dots are spherical or elliptical in various dimensions. For example, the correlation matrix for two variables tells us whether the dots in a scattergram are circular (when the correlation is close to 0), elliptical (when the correlation is greater to 0 but not close to 1.0 or -1.0), or approach forming a straight line (when the correlation approaches 1.0 or -1.0). The correlation matrix also indicates the direction of the dots. For example, a positive correlation for two variables implies that the dots are oriented from the "southwest towards the northeast" while a negative correlation denotes that the dots are going from the "northwest towards the southeast."

To summarize, classic multivariate analysis uses summary statistics to inform us about three properties of the data points in hyperspace. The first property is *location* and is summarized by the means. The second property is *spread* or *scale* and is summarized by the standard deviations. The third property is *shape* and is summarized by the correlations.

A final comment is in order. We have seen how a covariance matrix is a function of the standard deviations and the correlation matrix. Consequently, we could logically conclude that two types of matrices are needed for classic multivariate analysis--matrices of means and covariance matrices. The means would inform us about location and the covariance matrix would inform us about the spread and shape of the dots in hyperspace. Indeed, this is true and many multivariate techniques are expressed in just this way. However, it is much easier for us humans to think in terms of correlations than it is in terms of covariances. So it is best for us to conceptualize multivariate analysis in terms standard deviations and correlations instead of covariances.

## The Bad News: Read This!

To really understand multivariate analysis, we need only consider the three types of matrices noted above. Unfortunately, computer algorithms for multivariate analysis have been stuck in the dark ages when computations were performed by hand and every effort was made to simplify calculations by avoiding unnecessary computations. Although we human beings think in correlational terms, traditional multivariate analysis is expressed in “corrected sums of squares and cross products terms.” Why? Because once the CSSCP matrix has been calculated, one has to spend considerable effort to transform the CSSCP matrix into a covariance matrix and then the covariance matrix into a correlation matrix. In the days before digital computers, this was wasted effort. Although modern computers can now transform these matrices in microseconds, we are still faced with the legacy of ancestral methods. Hence, programs like SAS and SPSS refer to “error sums of squares and cross products matrices” while today we should be speaking in terms of “error correlation matrices" and "error standard deviations.”

Those of you that have performed ANOVAs have already been subjected to the anachronistic nature of modern statistics. The traditional ANOVA table has columns for degrees of freedom, sums of squares, and mean squares because these were the computational columns in the old paper-and-pencil spreadsheets used to arrive at an *F* ratio. The key interpretative statistics in a simple oneway ANOVA are the group means and the standard deviations, NOT the sums of squares and mean squares. The group means tell us about location and the group standard deviations tell us how spread out the group scores are around their central location. Similarly, when we get to MANOVA, we will find that the computer printouts are full of sums of squares and cross products matrices. These are the old computational intermediaries for constructing tests of significance. The main interpretive statistics for a oneway MANOVA are the vectors of means, a diagonal matrix of standard deviations, and the correlation matrix.