CHAPTER 6

Applying the Tools to Multivariate Data

6.1 INTRODUCTION

In this chapter we come around full circle to the three substantive problems first introduced in Chapter 1. As recalled, each problem was based on a "toy" data bank and formulated in terms of three commonly used techniques in multivariate analysis: (a) multiple regression, (b) principal components analysis, and (c) multiple discriminant analysis.

We discuss the multiple regression problem first. The problem is structured so as to require the solution of a set of linear equations, called "normal" equations from least-squares theory. These equations are first set up in terms of the original data, and the parameters are found by matrix inversion. We then show how the same problem can be formulated in terms of either a covariance or a correlation matrix.

R^2, a measure of overall goodness of fit, and other regression statistics such as partial correlation coefficients, are also described. The results are interpreted in terms of the substantive problem of interest, and comments are made on the geometric aspects of multiple regression.

We then discuss variations on the general linear model of multiple regression: analysis of variance and covariance, two-group discriminant analysis, and binary-valued regression (in which all variables, criterion and predictors, are expressed as zero–one dummies). This discussion is presented as another way of showing the essential unity among single-criterion, multiple-predictor models.

Discussion then turns to the second substantive problem, formulated as a principal components model. Here the solution is seen to entail finding the eigenstructure of a covariance matrix. Component loadings and component scores are also defined and computed in terms of the sample problem.

After solving this sample problem, some general comments are made about other aspects of factor analysis, such as the factoring of other kinds of cross-product matrices, rotation of component solutions, and dimension reduction methods other than the principal components procedure.

The three-group multiple discriminant problem of Chapter 1 is taken up next. This problem is formulated in terms of finding the eigenstructure of a nonsymmetric matrix which, in turn, represents the product of two symmetric matrices. The discriminant

functions are computed, and significance tests are conducted. The results are interpreted in the context of the third sample problem.

We then turn to other aspects of multiple discriminant analysis (MDA), including classification matrices, alternative ways to scale the discriminant functions, and the relationship of MDA to principal components analysis. Finally, some summary-type comments are made about other techniques for dealing with multiple-criterion, multiple-predictor association.

The last major section of the chapter is, in some respects, a prologue to textbooks that deal with multivariate analysis per se. In particular, the concepts of transformational geometry, as introduced in earlier chapters, are now brought together as another type of descriptor by which multivariate techniques can be classified. Under this view multivariate methods are treated as procedures for matching one set of numbers with some other set or sets of numbers. Techniques can be distinguished by the nature of the transformation(s) used to effect the matching and the characteristics of the transformed numbers.

This organizing principle is described in some detail and suggests a framework that can be useful for later study of multivariate procedures as well as suggestive of new models in this field.

6.2 THE MULTIPLE REGRESSION PROBLEM

We are now ready to work through the details of the sample problem in Chapter 1 dealing with the relationship of employee absenteeism Y, to attitude toward the firm X_1 and number of years employed by the firm X_2. To simplify our discussion, the basic data, first shown in Table 1.2, are reproduced in Table 6.1.

As recalled from the discussion in Chapter 1, here we are interested in

1. finding a regression equation for estimating values of the criterion variable Y from a linear function of the predictor variables X_1 and X_2;
2. determining the strength of the overall relationship;
3. testing the significance of the overall relationship;
4. determining the relative importance of the two predictors X_1 and X_2 in accounting for variation in Y.

6.2.1 The Estimating Equation

As again recalled from Chapter 1, the multiple regression equation

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2}$$

is a linear equation for predicting values of Y that minimize the sum of the squared errors

$$\sum_{i=1}^{12} e_i^2 = \sum_{i=1}^{12} (Y_i - \hat{Y}_i)^2$$

TABLE 6.1

Basic Data of Sample Problem (from Table 1.2)

	Number of days absent			Attitude rating			Years with company		
Employee	Y	Y_d	Y_s	X_1	X_{d1}	X_{s1}	X_2	X_{d2}	X_{s2}
a	1	−5.25	−0.97	1	−5.25	−1.39	1	−3.92	−1.31
b	0	−6.25	−1.15	2	−4.25	−1.13	1	−3.92	−1.31
c	1	−5.25	−0.97	2	−4.25	−1.13	2	−2.92	−0.98
d	4	−2.25	−0.41	3	−3.25	−0.86	2	−2.92	−0.98
e	3	−3.25	−0.60	5	−1.25	−0.33	4	−0.92	−0.31
f	2	−4.25	−0.78	5	−1.25	−0.33	6	1.08	0.36
g	5	−1.25	−0.23	6	−0.25	−0.07	5	0.08	0.03
h	6	−0.25	−0.05	7	0.75	0.20	4	−0.92	−0.31
i	9	2.75	0.51	10	3.75	0.99	8	3.08	1.03
j	13	6.75	1.24	11	4.75	1.26	7	2.08	0.70
k	15	8.75	1.61	11	4.75	1.26	9	4.08	1.37
l	16	9.75	1.80	12	5.75	1.53	10	5.08	1.71
Mean	6.25			6.25			4.92		
Standard deviation	5.43			3.77			2.98		

Appendix A shows how the set of normal equations, used to find b_0, b_1, and b_2, are derived. In terms of the specific problem here, we have in matrix notation:[1]

$$\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 16 \end{bmatrix}; \quad \mathbf{X} = \begin{matrix} C & X_1 & X_2 \end{matrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 12 & 10 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}; \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_{12} \end{bmatrix}$$

The model being fitted by least squares is

$$\boxed{\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}}$$

Notice that the model, in matrix form, starts off with the observed vector y and the observed matrix X. As will be shown later, the device of including a column of ones as the first column of X (called C) is employed for estimating the intercept b_0.

We wish to solve for b, the vector of parameters, so that $\sum_{i=1}^{12} e_i^2 = \mathbf{e}'\mathbf{e}$ is minimized. As can be checked in Appendix A, the problem is a standard one in the calculus and leads to the so-called normal equations which, expressed in matrix form, are

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

[1] The sample entries in y and X are taken from Table 6.1.

That is, we first need to find the minor product moment of $\mathbf{X}$, which is $\mathbf{X}'\mathbf{X}$. Next, we find the inverse of $\mathbf{X}'\mathbf{X}$ and postmultiply this inverse by $\mathbf{X}'\mathbf{y}$.

In terms of the specific problem of Table 6.1, we have

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \left\{ \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 12 \\ 1 & 1 & 2 & \cdots & 10 \end{bmatrix}}_{\mathbf{X}'} \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 12 & 10 \end{bmatrix}}_{\mathbf{X}} \right\}^{-1} \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 12 \\ 1 & 1 & 2 & \cdots & 10 \end{bmatrix}}_{\mathbf{X}'} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 16 \end{bmatrix}}_{\mathbf{y}}$$

$$\mathbf{b} = \begin{bmatrix} -2.263 \\ 1.550 \\ -0.239 \end{bmatrix}$$

Hence, in terms of the original data of Table 6.1, we have the estimating equation

$$\hat{Y}_i = -2.263 + 1.550X_{i1} - 0.239X_{i2}$$

The 12 values of $\hat{Y}_i$ appear in the lower portion of Table 6.2, along with the residuals e_i.

If one adds up the squared residuals, one obtains (within rounding error) the residual term shown in the analysis of variance table of Table 6.2:

$$\text{residual} = 34.099$$

The total sum of squares is obtained from

$$\sum_{i=1}^{12} (Y - \bar{Y})^2 = 354.25$$

and the difference

$$\text{due to regression} = 320.15$$

6.2.2 Strength of Overall Relationship and Statistical Significance

The squared multiple correlation coefficient is R^2, and this measures the portion of variance in Y (as measured about its mean) that is accounted for by variation in X_1 and X_2. As mentioned in Chapter 1, the formula is

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{12} e_i^2}{\sum_{i=1}^{12} (Y_i - \bar{Y})^2}}$$

$$R^2 = 1 - \frac{34.099}{354.25} = 0.904$$

TABLE 6.2

Selected Output from Multiple Regression

$R^2 = 0.904$; $R = 0.951$; variance of estimate 3.789

Analysis of Variance for Multiple Regression

Source	df	Sums of squares	Mean squares	F ratio
Due to regression	2	320.151	160.075	42.25
Residual	9	34.099	3.789	
Total	11	354.250		

Variable	Regression coefficients	Standard errors	t values	Partial correlations	Proportion of cumulative variance
X_1	1.550	0.481	3.225	0.732	0.902
X_2	−0.239	0.606	−0.393	−0.130	0.002

Y intercept −2.263

Table of Residuals

Employee	Y	$\hat{Y}$	e	Employee	Y	$\hat{Y}$	e
a	1	−0.95	1.95	g	5	5.85	−0.84
b	0	0.60	−0.60	h	6	7.63	−1.63
c	1	0.36	0.64	i	9	11.33	−2.33
d	4	1.91	2.09	j	13	13.11	−0.11
e	3	4.53	−1.53	k	15	12.64	2.36
f	2	4.05	−2.05	l	16	13.95	2.05

The statistical significance of R, the positive square root of R^2, is tested via the analysis of variance subtable of Table 6.2 by means of the F ratio:

$$F = 42.25$$

which, with 2 and 9 degrees of freedom, is highly significant at the $\alpha = 0.01$ level. Thus, as described in Chapter 1, the equivalent null hypotheses

$$\boxed{\begin{array}{c} R_p = 0 \\ \beta_1 = \beta_2 = 0 \end{array}}$$

are rejected at the 0.01 level, and we conclude that the multiple correlation is significant.

Up to this point, then, we have established the estimating equation and measured, via R^2, the strength of the overall relationship between Y versus X_1 and X_2.

If we look at the equation again

$$\hat{Y}_i = -2.263 + 1.550X_{i1} - 0.239X_{i2}$$

we see that the intercept is negative. In terms of the current problem, a negative 2.263 days of absenteeism is impossible, illustrating, of course, the possible meaninglessness of extrapolation beyond the range of the predictor variables used in developing the parameter values.

The partial regression coefficient for X_1 seems reasonable; it says that predicted absenteeism increases 1.55 days per unit increase in attitude rating. This is in accord with the scatter plot (Fig. 1.2) that shows the association of Y with X_1 alone.

The partial regression coefficient for X_2, while small in absolute value, is negative, even though the scatter plot of Y on X_2 alone (Fig. 1.2) shows a positive relationship. The key to this seeming contradiction lies in the strong positive relationship between the predictors X_1 and X_2 (also noted in the scatter plot of Fig. 1.2). Indeed, the correlation between X_1 and X_2 is 0.95. The upshot of all of this is that once X_1 is in the equation, X_2 is so redundant with X_1 that its inclusion leads to a negative partial regression coefficient that effectively is zero (given its large standard error).

6.2.3 Other Statistics

The redundancy of X_2, once X_1 is in the equation, is brought out in Table 6.2 under the column

	Proportion of cumulative variance
X_1	0.902
X_2	0.002

That is, of the total $R^2 = 0.904$, the contribution of X_1 alone represents 0.902. The increment due to X_2 (0.002) is virtually zero, again reflecting its high redundancy with X_1.

This same type of finding is reinforced by examining the t values and the partial correlations in Table 6.2. These are

	t Values	Partial correlations
X_1	3.225	0.732
X_2	−0.393	−0.130

The Student t value is the ratio of a predictor variable's partial regression coefficient to its standard error. The standard error, in turn, is a measure of how well the predictor variable itself can be predicted from a linear combination of the other predictors. The higher the standard error, the more redundant (better predicted) that predictor variable is with the others. Hence, the less contribution it makes to Y on its own and the lower its t value.

We see that the ratio of b_1 to its own standard error is

$$t(b_1) = \frac{1.550}{0.481} = 3.225$$

which is significant at the 0.01 level. The t value for X_2 of −0.393 is not significant, however. Without delving into formulas, the t test is a test of the separate significance of each predictor variable X_j, when included in a regression model, versus the same

regression model with all predictors included except it. We note here that only X_1 is needed in the equation.

The partial correlations also suggest the importance of X_1 rather than X_2 in accounting for variation in Y. The partial correlation of Y with some predictor X_j is their simple correlation when both variables are expressed on a residual basis, that is, net of the linear association of each with all of the other predictors. In the present problem, the partial correlation of Y with X_1 is considerably higher than Y with X_2, supporting the earlier conclusions.

But what if X_2 is entered first in the regression? What happens in this case to the various statistics reported in Table 6.2? As it turns out, the only statistic that changes if X_2 is credited with as much variance as it can account for before X_1 is allowed to contribute to criterion variance is the last column, proportion of cumulative variance. If X_2 is entered first, it is credited with 0.79, while X_1 is credited with only 0.11 of the 0.90 total. The rest of the output does not change, and X_2 is still eliminated from the regression on the basis of the t test results.

What this example points out is that in the usual case of correlated predictors, the question of "relative importance" of predictors is ambiguous. Many researchers interpret relative importance in terms of the change in R^2 occurring when the predictor in question is the last to enter. Other importance measures are also available, as pointed out by Darlington (1968). However, in the case of correlated predictors, no measure is entirely satisfactory.

6.2.4 Other Formulations of the Problem

In the sample problem of Table 6.1, the normal equations were formulated in terms of the original data. Alternatively, suppose we decided to work with the mean-corrected scores Y_d, X_{d1}, X_{d2}. In this case we would compute the covariance matrix

$$\mathbf{C} = \mathbf{X}_d'\mathbf{X}_d/m$$

and the vector of partial regression parameters would be found from

$$\boxed{\mathbf{b} = \mathbf{C}^{-1}\mathbf{a}(y)}$$

where $\mathbf{a}(y)$ is the vector of covariances between the criterion and each predictor in turn with elements

$$\boxed{\begin{aligned} a_1 &= \mathbf{y}_d'\mathbf{x}_{d1}/m \\ a_2 &= \mathbf{y}_d'\mathbf{x}_{d2}/m \end{aligned}}$$

in the sample problem.

The preceding formula for computing $\mathbf{b}$ would find only b_1 and b_2 since all data would be previously mean centered. To work back to original data, we can find the intercept of the equation by the simple formula:

$$\boxed{b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2}$$

If we decided to work with the standardized data Y_s, X_{s1}, and X_{s2}, the appropriate minor product moment is the correlation matrix

$$R = \mathbf{X}_s'\mathbf{X}_s/m$$

and the vector of parameters $\mathbf{b}^*$ (often called beta weights) would be found from

$$\boxed{\mathbf{b}^* = \mathbf{R}^{-1}\mathbf{r}(y)}$$

where $\mathbf{r}(y)$ is the vector of product-moment correlations between the criterion and each predictor in turn, with elements

$$\boxed{\begin{aligned} r_1 &= \mathbf{y}_s'\mathbf{x}_{s1}/m \\ r_2 &= \mathbf{y}_s'\mathbf{x}_{s2}/m \end{aligned}}$$

in the sample problem.

The vector $\mathbf{b}^*$ measures the change in Y per unit change in each of the predictors, when all variables are expressed in standardized units. To find the elements of $\mathbf{b}$, we use the conversion equations

$$\boxed{\begin{aligned} b_1 &= b_1^* \frac{s_y}{s_{x_1}} \\ b_2 &= b_2^* \frac{s_y}{s_{x_2}} \end{aligned}}$$

These simple transformations, involving ratios of standard deviations, enable us to express changes in Y per unit change in X_1 and X_2 in terms of the original Y units. Having done this, we can then solve for the intercept term in exactly the same way:

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

as shown in the covariance matrix case. Many computer routines for performing multiple regression operate on the correlation matrix. As seen here, any of the cross-product matrices—raw cross products, covariances, or correlations—can be used and, in the latter two cases, modified for expressing regression results in terms of original data.

6.2.5 Geometric Aspects—the Response Surface Model

Figure 1.2 showed two-dimensional scatter plots of Y versus X_1, Y versus X_2, and X_2 versus X_1 It is also a relatively simple matter to plot a three-dimensional diagram of Y versus X_1 and X_2. This is shown in Fig. 6.1.

We also show the fitted regression plane, as computed by least squares. This type of model, in which observations are represented by points and variables by dimensions, is often called the response surface or point model.

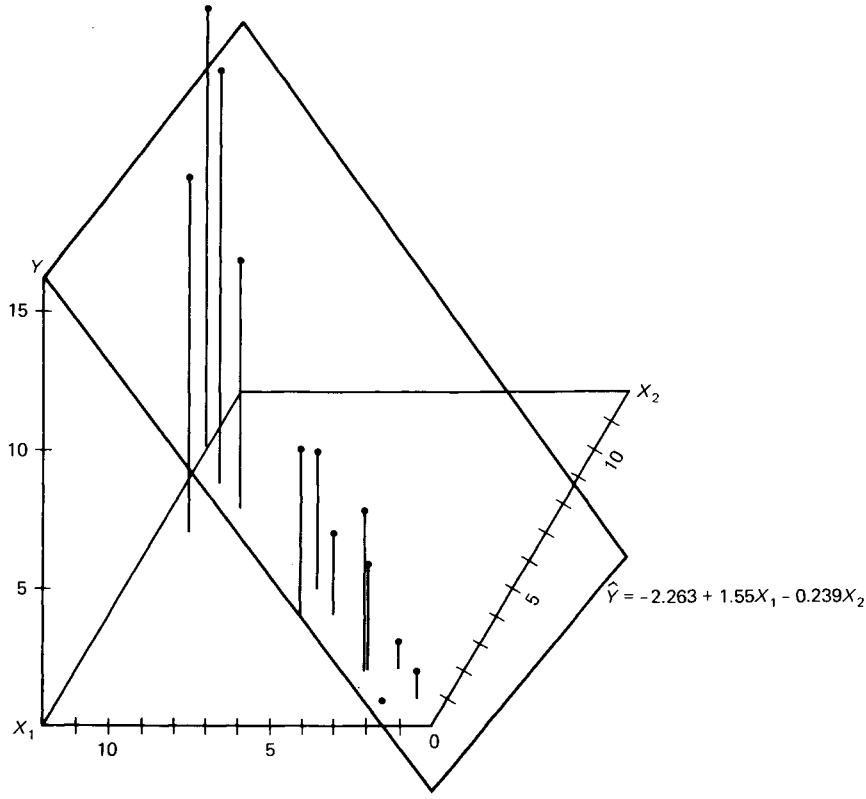

Fig. 6.1 Three-dimensional plot and fitted regression plane.

The intersection of the regression plane in Fig. 6.1 with the Y axis provides the estimate b_0, the intercept term. If we next imagine constructing a plane perpendicular to the X_1 axis, we, in effect, hold X_1 constant; hence b_2 represents the estimated contribution of a unit change in X_2 to a change in Y. Similar remarks pertain to the interpretation of b_1.

The regression plane itself is oriented so as to minimize the sum of squared deviations between each Y_i and its counterpart value on the fitted plane, where these deviations are taken along directions parallel to the Y axis. Similarly, we can find the sum of squared deviations about the mean of the Y_i's by imagining a plane perpendicular to the Y axis passing through the value $\bar{Y}$. Total variation in Y is thus partitioned into two parts. As indicated earlier, these separate parts are found by

1. subtracting *unaccounted-for variation,* involving squared deviations $(Y_i - \hat{Y}_i)^2$ about the fitted regression plane, from
2. *total* variation involving squared deviations $(Y_i - \bar{Y})^2$ from the plane imagined to be passing through $\bar{Y}$.

The quantity $\sum_{i=1}^{12}(Y_i - \hat{Y}_i)^2$ represents the unaccounted-for sum of squares, and the quantity $[\sum_{i=1}^{12}(Y_i - \bar{Y})^2 - \sum_{i=1}^{12}(Y_i - \hat{Y}_i)^2]$ represents the accounted-for sum of

squares. If no variation is accounted for, then we note that using $\bar{Y}$ is just as good at predicting Y as introducing the variation in X_1 and X_2.

6.2.6 An Alternative Representation

The foregoing representation of the 12 responses in variable space considers the 12 observations as points in three dimensions, where each variable, Y, X_1, or X_2, denotes a dimension. Alternatively, we can imagine that each of the 12 employees represents a dimension, and each of the variables constitutes a vector in this 12-dimensional person space. As we know from the discussion of matrix rank in Chapter 5, the three vectors will not span the whole 12-dimensional space but, rather, will lie in (at most) a three-dimensional subspace that is embedded in the 12-dimensional person space.

We also remember that if the vectors are translated to a mean-centered origin and are assumed to be of unit length, the (product-moment) correlation between each pair of vectors is given by the cosine of their angle. In this case we have three two-variable correlations: r_{yx_1}, r_{yx_2}, and $r_{x_1x_2}$.

This concept is pictured, in general terms, in Fig. 6.2. In the left panel of the figure are two unit length vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ emanating from the origin. Each is a 12-component vector of unit length, embedded in the "person" space. The cosine of the angle Ψ separating $\mathbf{x}_1$ and $\mathbf{x}_2$ is the simple correlation $r_{x_1x_2}$.

Since the criterion vector y is not perfectly correlated with $\mathbf{x}_1$ and $\mathbf{x}_2$, it must extend into a third dimension. The cosines of its angular separation between $\mathbf{x}_1$ and $\mathbf{x}_2$ are each measured, respectively, by its simple correlations r_{yx_1} and r_{yx_2}. However, one can project y onto the plane formed by $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection of y onto this plane is denoted by $\hat{\mathbf{y}}$.

In terms of this viewpoint, the idea behind multiple regression is to find the particular vector in the $\mathbf{x}_1$, $\mathbf{x}_2$ plane that minimizes the angle θ with y. This vector will be the projection $\hat{\mathbf{y}}$ onto the plane formed by $\mathbf{x}_1$, $\mathbf{x}_2$. Since any vector in the $\mathbf{x}_1$, $\mathbf{x}_2$ plane is a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, it follows that we want the vector $\hat{\mathbf{y}} = b_1^*\mathbf{x}_1 + b_2^*\mathbf{x}_2$, where the b_j^*'s are beta weights, that minimizes the angle, or maximizes the cosine of the angle with y. The cosine of this angle θ (see Fig. 6.2) is R, the multiple correlation. The problem then is to find a set of b_j^*'s that define a linear combination of the vectors $\mathbf{x}_1$

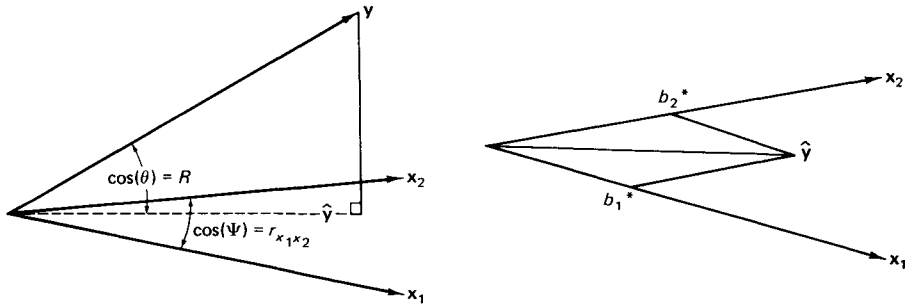

Fig. 6.2 Geometric relationship of $\hat{\mathbf{y}}$ to y in vector space. The graph on the right shows geometric interpretation of partial regression weights in vector space.

and x_2 maximizing the cosine R of θ, the angle separating y and $\hat{y}$. However, this is equivalent to minimizing the square of the distance from the terminus of y to its projection $\hat{y}$. This criterion

$$\text{minimize}\left[\sum_{i=1}^{12}(y_{si}-\sum_{j=1}^{n}b_j{}^*x_{sij})^2\right]$$

again leads to the least-squares equations. (Since all variables are assumed to be measured in standardized form, the intercept b_0 is zero.)

In general, the x_1, x_2 axes will be oblique, as noted in Fig. 6.2. The right panel shows that a linear combination of x_1 and x_2, which results in the predicted vector $\hat{y}$, involves combining oblique axes via $b_1{}^*$and $b_2{}^*$. In this case, $b_1{}^*$and $b_2{}^*$are direction cosines.

Figure 6.3 shows some conditions of interest. In Panel I we see that y is uncorrelated with x_1 and x_2. This lack of correlation is indicated by the 90° angle between y and the x_1, x_2 plane. Panel II shows the opposite situation where y is perfectly correlated with x_1 and x_2 and, hence, can be predicted without error by a linear combination of x_1 and x_2.

Panel III shows the case where x_1 and x_2 are uncorrelated and y evinces some correlation with x_1 and none with x_2. Panel IV shows the case in which x_1 and x_2 are uncorrelated, but the projection of y lies entirely along x_2.

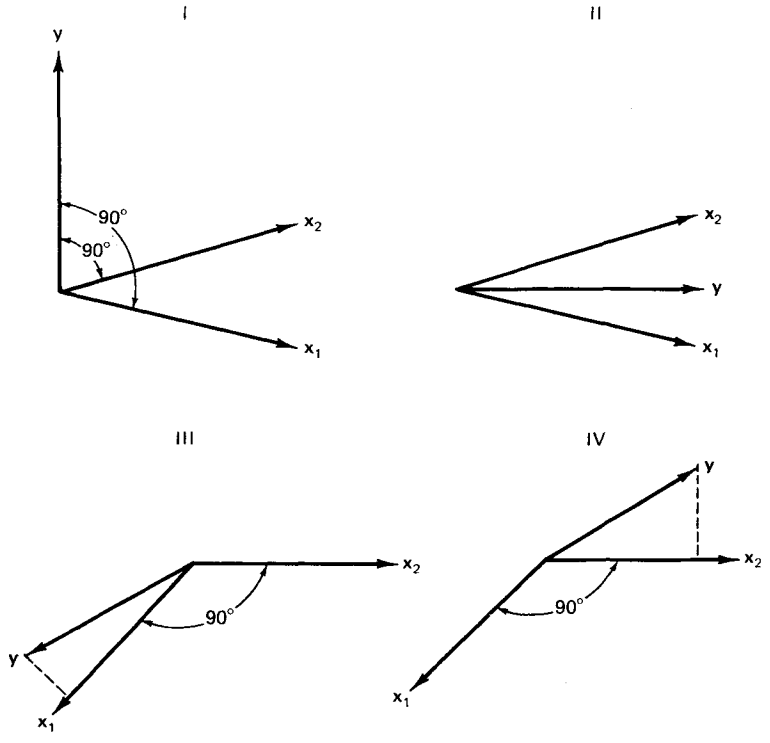

Fig. 6.3 Some illustrative cases involving multiple correlation. Key: I, no correlation; II, perfect correlation; III, y correlated with x_1 only; IV, y correlated with x_2 only.

In summary, the multiple correlation coefficient R is the cosine of the angle θ made by y and $\hat{y}$. The b_j*'s are normalized beta weights and represent coordinates of $\hat{y}$ in the oblique space of the predictor variables. If more than two predictors are involved, the same geometric reasoning applies, although in this case the predictors involve higher-dimensional hyperplanes.

Partial correlations between y and x_1 and x_2, respectively, can also be interpreted. For example, if we consider a plane perpendicular to x_2 and project y and x_1 onto this plane, $r_{yx_1 \cdot x_2}$, the partial correlation of y with x_1 (with x_2 partialed out) is represented by the cosine of the angle separating them on this plane. Similar remarks pertain to the partial correlation of y with x_2 and would involve a projection onto a space that is orthogonal to x_1. The same general idea holds for larger numbers of predictors.

6.3 OTHER FORMS OF THE GENERAL LINEAR MODEL

The typical multiple regression model considers each variable as intervally scaled. This representation is overly restrictive. Indeed, by employing the dummy-variable device, as introduced in Chapter 1, we can extend the linear regression model to a more general model that subsumes the techniques of

1. analysis of variance
2. analysis of covariance
3. two-group discriminant analysis
4. binary-valued regression

All of these cases are developed from two basic concepts: (a) the least-squares criterion for matching one set of data with some transformation of another set and (b) the dummy variable.

Figure 6.4 shows, in a somewhat abstract sense, various special cases in terms of the response surface or point model involving m observations in three dimensions.

Panel I of Fig. 6.4 shows each of the three columns of a data matrix as a dimension and each row of the matrix as a point. If we were then to append to the m x 2 matrix of predictors a unit vector, we have the familiar matrix expression for fitting a plane or, more generally, a response surface, in the three-dimensional space shown in Panel I. Predicted values $\hat{y}$ of the criterion variable y are given by

$$\boxed{\hat{y} = \mathbf{Xb}}$$

where **b** is a 3 x 1 column vector with entries b_0, b_1, b_2 denoting, respectively, the intercept, partial regression coefficient for x_1, and partial regression coefficient for x_2.

However—and this is the key point—nothing in the least-squares procedure precludes y (or x_1 or x_2 for that matter) from taking on values that are just zero or one. Panel II shows the case where y assumes only binary values, but x_1 and x_2 are allowed to be continuous. Panel III shows the opposite situation. Panel IV shows a "mixed" case where y and x_1 are continuous and x_2 is binary valued. Panel V shows a case where all three variables are binary valued.

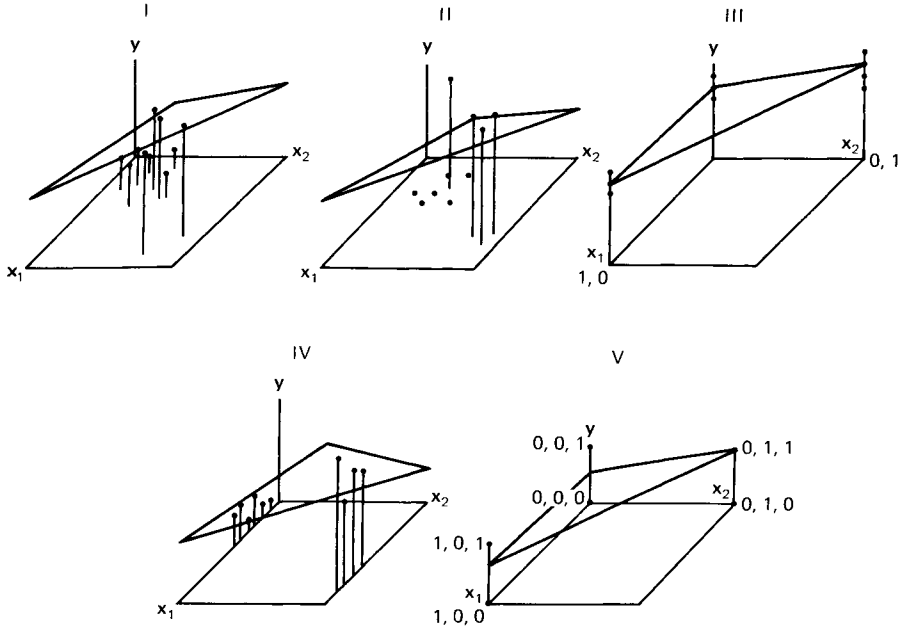

Fig. 6.4 Variations of the response surface model.

Not surprisingly, from Chapter 1 we recognize Panel I as a traditional multiple regression formulation. Panel II appears as a two-group discriminant function. Panel III appears as a one-way analysis of variance design with one treatment variable at three levels. Panel IV represents a simple analysis of covariance design with a single two-level treatment variable (x_2) and one continuous covariate (x_1). Panel V seems less familiar, but could be viewed as a type of binary-valued regression where the criterion and the predictors are each dichotomies (e.g., predicting high versus low attitude toward the firm as a function of sex and marital status).

As observed from Fig. 6.4, we can now conclude that all of these models are variations on a common theme—namely, one in which we are attempting to find some type of linear transformation that results in a set of scores that best match, in the sense of minimum sum of squared deviations, a set of criterion scores. In each case we are fitting a plane in the three-dimensional space of x_1, x_2, and y and then finding estimates $\hat{y}$ of y that result in a minimum sum of squared deviations.

All of the cases depicted in Fig. 6.4 are characterized by the fact that a single-criterion variable, either 0–1 coded or intervally scaled, is involved. Our interest is in finding some linear combination of predictors, where b denotes the set of combining weights, that leads to a set of predicted scores $\hat{y}$ that are most congruent with the original scores y.

Extension of the multiple regression model to handle binary-valued predictors is described in various texts (e.g., Neter and Wasserman, 1974) in terms of a general linear model.

If a further extension is made in order to allow for a binary-valued criterion, least squares can still be used to estimate parameter values, although the usual statistical tests

are not strictly appropriate since the normality and constant variance assumptions are missing. Still, as a descriptive device least squares can be used to find estimating equations for all of the cases depicted in Fig. 6.4.

Discussion of the multiple regression problem has thus resulted in a much wider scope of application than might first have been imagined. Through the dummy-variable coding device, one can subsume all cases of interest—analysis of variance and covariance, two-group discrimination, binary-valued regression—that involve a single-criterion variable and multiple predictors. Moreover, although detailed discussion of the geometric aspects of the models was more or less confined to multiple regression, all of these methods can be represented by either

1. the response surface or point model in variable space, or
2. the vector model in person or object space.

From the standpoint of matrix algebra, all of the preceding models entail solutions based on a set of linear (the normal) equations from least-squares theory. As such, the operation of matrix inversion becomes germane, as does the concept of matrix rank and related ideas such as determinants. In brief, the algebraic underpinnings of single-criterion, multiple-predictor association are concepts of matrix rank and inversion. Thus, it is no accident that much of the discussion in earlier chapters was devoted to these topics.

6.4 THE FACTOR ANALYSIS PROBLEM

If matrix inversion and rank are the hallmarks of single-criterion, multiple-predictor association, then eigenstructures are the key concepts in dimension-reducing methods like factor analysis. Eigenstructures are also essential in multiple-criterion, multiple-predictor association, as we shall see later in the chapter.

In Chapter 1 we introduced a small-scale problem in principal components analysis in the context of developing a reduced space for the two predictors: (a) X_1, attitude score and (b) X_2, number of years with company. Using the X_{d1} and X_{d2} data of Table 6.1 we wish to know if a change of basis vectors can be made that will produce an axis whose variance of point projections is maximal.

This is a standard problem in finding the eigenstructure of a symmetric matrix. Here we employ the covariance matrix, although in some cases one might want to use some other type of cross-products matrix. Table 6.3 details the steps involved in finding the eigenstructure of **C**, the simple 2×2 covariance matrix of the sample problem of predictor variables in Table 6.1. (Supporting calculations appear in Chapter 5.)

As observed from Table 6.3, the first eigenvalue $\lambda_1 = 22.56$ accounts for nearly all, actually 98 percent, of the variance of **C**, the covariance matrix. The linear composite $\mathbf{z}_1$, developed from $\mathbf{t}_1$, makes an angle of approximately 38° with the horizontal axis, as noted in Fig. 6.5. Thus, if we wished to combine the vectors of scores X_{di1} and X_{di2} into a single linear composite, we would have, in scalar notation,

$$\boxed{z_{i(1)} = 0.787X_{di1} + 0.617X_{di2}}$$

Note also that the second linear composite $\mathbf{z}_2$ is at a right angle to $\mathbf{z}_1$.

TABLE 6.3

Finding the Eigenstructure of the Covariance Matrix (Predictor Variables in Table 6.1)

Covariance matrix

$$\mathbf{C} = \begin{array}{c} \\ X_1 \\ X_1 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} 14.19 & 10.69 \\ 10.69 & 8.91 \end{bmatrix} \end{array}$$

Matrix equation

$$(\mathbf{C} - \lambda_i \mathbf{I})\mathbf{t}_i = \mathbf{0}$$

Characteristic equation

$$|\mathbf{C} - \lambda_i \mathbf{I}| = \begin{vmatrix} 14.19 - \lambda_i & 10.69 \\ 10.69 & 8.91 - \lambda_i \end{vmatrix} = 0$$

Expansion of determinant

$$\lambda_i^2 - 23.1\lambda_i + 126.433 - 114.276 = 0$$

Eigenvalues

$$\lambda_1 = 22.56; \qquad \lambda_2 = 0.54$$

Eigenvectors

$$\mathbf{t}_1 = \begin{bmatrix} 0.787 \\ 0.617 \end{bmatrix}; \qquad \mathbf{t}_2 = \begin{bmatrix} 0.617 \\ -0.787 \end{bmatrix}$$

6.4.1 Component Scores

Component scores are the projections of the twelve points on each new axis, z_1 and z_2, in turn. For example, the component score of the first point on z_1 is

$$z_{1(1)} = 0.787(-5.25) + 0.617(-3.92) = -6.55$$

as shown in Fig. 6.5. The full set of component scores appears in Table 5.1.

The variance of each column of **Z** will equal its respective eigenvalue. If one wishes to find a matrix of component scores with unit variance, this is done quite simply by a transformation involving the matrix of eigenvectors **T** and the reciprocals of the square roots of the eigenvalues:[2]

$$\boxed{\mathbf{Z}_s = \mathbf{X}_d \mathbf{T} \mathbf{\Lambda}^{-1/2}}$$

In the sample problem, the product of **T** and $\mathbf{\Lambda}^{-1/2}$ is given by

$$\mathbf{S} = \begin{array}{c} \\ \mathbf{T} \end{array} \begin{array}{c} \begin{array}{cc} & \mathbf{\Lambda}^{-1/2} \end{array} \\ \begin{bmatrix} 0.787 & 0.617 \\ 0.617 & -0.787 \end{bmatrix} \begin{bmatrix} 0.211 & 0 \\ 0 & 1.361 \end{bmatrix} \end{array}; \qquad \mathbf{S} = \begin{bmatrix} 0.166 & 0.840 \\ 0.130 & -1.071 \end{bmatrix}$$

In the sample problem, $\mathbf{Z}_s$ denotes the 12 x 2 matrix of unit-variance component scores; $\mathbf{X}_d$ is the 12 x 2 matrix of mean-centered predictor variables; **T** is the matrix of

[2] In this illustration we use Λ to denote the diagonal matrix of eigenvalues of C, the covariance matrix. Accordingly, $\Lambda^{-1/2}$ is a diagonal matrix whose main diagonal elements are the reciprocals of the square roots of the main diagonal elements of Λ.

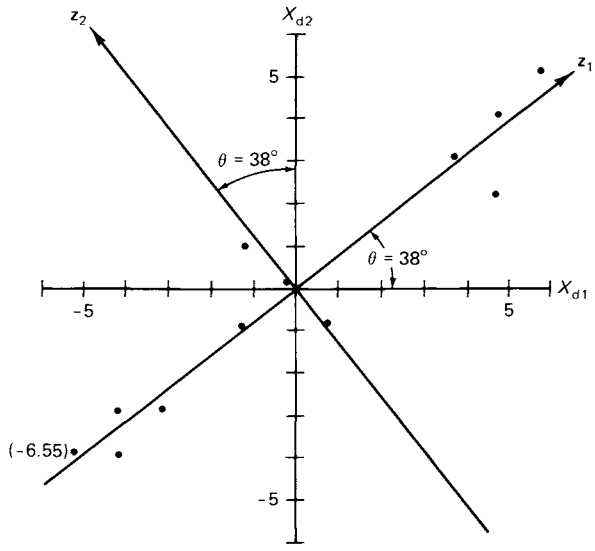

Fig. 6.5 Principal components rotation of mean-corrected predictor variables.

eigenvectors from Table 6.3; and $\Lambda^{-1/2}$ is a diagonal matrix of the reciprocals of the square roots of the eigenvalues. By application of the transformation matrix **S** (instead of **T**), we would obtain unit-variance component scores. That is, in this case,

$$\boxed{\mathbf{Z}_s'\mathbf{Z}_s/m = \mathbf{I}}$$

Geometrically, then, postmultiplication of $\mathbf{X}_d$ by **S** has the effect of transforming the ellipsoidal-like swarm of points in Fig. 6.5 into a circle, along the axes of the ellipse.

6.4.2 Component Loadings

Component loadings are simply product-moment correlations of each original variable X_{d1} and X_{d2} with each set of component scores.

To illustrate, the (unit-variance) component scores $z_{si(1)}$ on the first principal component are

a	−1.38	b	−1.21	c	−1.08	d	−0.92	e	−0.33	f	−0.07
g	−0.03	h	0.01	i	1.02	j	1.06	k	1.32	l	1.62

These represent the first column of $\mathbf{Z}_s$. For example,

$$z_{si(1)} = 0.166X_{di1} + 0.130X_{di2} = 0.166(-5.25) + 0.130(-3.92) = -1.38$$

The product-moment correlation of $z_{s(1)}$ with x_{d1} is 0.99, and the product-moment correlation of $z_{s(1)}$ with x_{d2} is 0.98. Not surprisingly, given the high variance accounted for by the first component, both loadings are high.

A more general definition of a component loading considers a loading as a weight, obtained for each variable, whose square measures the contribution that the variable makes to variation in the component. However, usually in applied work it is the

correlation matrix that is factored rather than the covariance, or some other type of cross products, matrix. Hence, the simpler definition of loading, namely, as the correlation of a variable with a component, is most prevalent.

In the present problem, the principal components analysis of a 2 x 2 correlation matrix would necessarily effect a 45° rotation, rather than the 38° rotation shown in Fig. 6.5. Hence the loadings of X_1 and X_2 on each component would necessarily be equal. However, this will not, in general, be the case with correlations based on three or more variables being analyzed by principal components.

The matrix of component "loadings" for the covariance matrix in the present problem is found quite simply from the relationship

$$\boxed{\mathbf{F} = \mathbf{T}\Lambda^{1/2}}$$

$$\mathbf{F} = \underset{\mathbf{T}}{\begin{bmatrix} 0.787 & 0.617 \\ 0.617 & -0.787 \end{bmatrix}} \underset{\Lambda^{1/2}}{\begin{bmatrix} 4.75 & 0 \\ 0 & 0.73 \end{bmatrix}}$$

$$\mathbf{F} = \begin{array}{c} \\ X_1 \\ X_2 \\ \lambda_i \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} 3.74 & 0.45 \\ 2.93 & 0.57 \end{bmatrix} \\ \begin{array}{cc} 22.56 & 0.54 \end{array} \end{array} \begin{array}{c} \\ 14.19 \\ 8.91 \\ 23.10 \end{array}$$

An interesting property of F is that the sum of the squared "loadings" of each component (column) equals its respective eigenvalue. For example,

$$\lambda_1 = (3.74)^2 + (2.93)^2 = 22.56$$

the variance of the first component, within rounding error.

Similarly, the sum of the squared entries of each variable (row) equals its respective variance. For example,

$$(3.74)^2 + (0.45)^2 = 14.19$$

the variance of X_1.

Finally, we see that both components together exhaust the total variance in the covariance matrix C. Furthermore, the first component itself accounts for 22.56/23.10 = 0.98 of the total variance.[3] Clearly, little is gained by inclusion of the second component insofar as the sample problem is concerned.

6.4.3 The Basic Structure of $\mathbf{X_d}$

Another way of looking at the principal components problem is in terms of the basic structure of a matrix, as described in Chapter 5. In line with our earlier discussion, suppose we wished to find the basic structure of

$$\boxed{\mathbf{X}_d/\sqrt{m} = \mathbf{U}\Delta\mathbf{T}'}$$

[3] Of additional interest is the fact that X_1 accounts for 14.19/23.10 or 0.61 of the total variance.

where, as we know, the minor product of the left-hand side of the equation represents the covariance matrix

$$\mathbf{C} = \mathbf{X}_d'\mathbf{X}_d/m$$

As shown in Chapter 5, C is symmetric and, hence, orthogonally decomposable into the triple product

$$\mathbf{C} = \mathbf{T}\Lambda\mathbf{T}'$$

$$\begin{matrix} & \mathbf{T} & & \Lambda & & \mathbf{T}' & \\ = \begin{bmatrix} 0.787 & 0.617 \\ 0.617 & -0.787 \end{bmatrix} & \begin{bmatrix} 22.56 & 0 \\ 0 & 0.54 \end{bmatrix} & \begin{bmatrix} 0.787 & 0.617 \\ 0.617 & -0.787 \end{bmatrix} & = \begin{bmatrix} 14.19 & 10.69 \\ 10.69 & 8.91 \end{bmatrix} \end{matrix}$$

where T is orthogonal, and Λ is diagonal. Note that Λ is the matrix of eigenvalues of $\mathbf{X}_d'\mathbf{X}_d/m$, and T is the matrix of eigenvectors, as shown in Table 6.3. As shown in Chapter 5, we can next solve for the orthonormal-by-columns matrix U by the equation

$$\mathbf{U} = \mathbf{X}_d/\sqrt{m}\,\mathbf{T}\Delta^{-1}$$

where $\Delta^{-1} = \Lambda^{-1/2}$. This, in turn, leads to the basic structure of $\mathbf{X}_d/\sqrt{m}$:

$$\mathbf{X}_d/\sqrt{m} = \mathbf{U}\Delta\mathbf{T}'$$

As recalled, U is orthonormal by columns; Δ is diagonal (a stretch transformation); and T' is orthogonal (a rotation). Moreover, as also pointed out in Chapter 5, if the eigenstructure of the major product moment $\mathbf{X}_d\mathbf{X}_d'/m$ is found instead, the matrix of its eigenvalues Λ will still be the same, and the representation is now

$$\boxed{\mathbf{X}_d\mathbf{X}_d'/m = \mathbf{U}\Lambda\mathbf{U}'}$$

where U is the same matrix found above. One then goes on to solve for T' in a manner analogous to that shown above.[4]

Finally, by similar procedures we could find the basic structure of any of the following matrices of interest in Table 6.1:

$$\mathbf{X}; \quad \mathbf{X}_d; \quad \mathbf{X}_s; \quad \mathbf{X}/\sqrt{m}; \quad \text{or} \quad \mathbf{X}_s/\sqrt{m}$$

by procedures identical to those shown above. As we know, division of X, $\mathbf{X}_d$, or $\mathbf{X}_s$ by the scalar $\sqrt{m}$ has no effect on the eigenvectors of either the minor or major product moments of X, $\mathbf{X}_d$, or $\mathbf{X}_s$. Corresponding eigenvalues of the product-moment matrix are changed by multiplication by $1/m$, which, in this case, represents the sample size.

6.4.4 Other Aspects of Principal Components Analysis

The example of Table 6.3 represents only one type of principal components analysis, namely, a components analysis of the covariance matrix C. As indicated above, one can component-analyze the averaged raw cross-products matrix $\mathbf{X}'\mathbf{X}/m$ or the correlation

[4] Alternatively, we could find U and T' simply by finding the eigenvectors of $\mathbf{X}_d\mathbf{X}_d'/m$ and $\mathbf{X}_d'\mathbf{X}_d/m$ separately.

matrix $\mathbf{X}_s'\mathbf{X}_s/m$. In general, the eigenstructures of these three matrices will differ. That is, unlike some factoring methods, such as canonical factor analysis (Van de Geer, 1971), principal components analysis is *not* invariant over changes in scale or origin.

Principal components analysis does exhibit the orthogonality of axes property in which the axes display sequentially maximal variance. That is, the first axis displays the largest variance, the second (orthogonal) axis, the next largest variance, and so on. In problems of practical interest a principal components analysis might involve a set of 30 or more variables, rather than the two variables X_1 and X_2, used here for illustrative purposes. Accordingly, the opportunity to replace a large number of highly correlated variables with a relatively small number of uncorrelated variables, with little loss of information, represents an attractive prospect. It is little wonder that principal components analysis has received much attention by researchers working in the behavioral and administrative sciences.

It is also not surprising that a large variety of other kinds of factoring methods have been developed to aid the researcher in reducing the dimensionality of his data space. Still, principal components represents one of the most common procedures for factoring matrices and, if anything, its popularity is on the rise.

However, from a substantive viewpoint, the orientation obtained from principal components may not be the most interpretable. Accordingly, applications researchers often rotate the component axes that they desire to retain to a more meaningful orientation from a content point of view. A number of procedures (Harman, 1967) are available to accomplish this task. Generally, the applied researcher likes to rotate component axes with a view to having each variable project highly on only one rotated dimension and nearly zero on others.

Another problem in any type of factoring procedure concerns the number of components (or factors, generally) to retain. Most data matrices will be full rank; hence, assuming that the number of objects exceeds the number of variables, one will obtain as many components as there are variables. Often the "lesser" components (those with lesser variance) are discarded; one often keeps only the first r ($<n$) components that account for some appreciable proportion (e.g., 80 to 90 percent) of the total variance in the data. Other rules for deciding how many factors to retain are also in use, including various statistical and graphical criteria. Still, the decision is largely a judgmental one, and factor analysis remains something of an ad hoc set of procedures.

Factor analysis—either principal components or other type of factoring procedure—represents only one class of methods for effecting dimensional reduction of one's data. More recently, new classes of techniques, such as multidimensional scaling (Green and Wind, 1973), have been used to develop reduced spaces. Many of these newer methods require only rank order input data. For example, the elements of a covariancelike matrix need only be ranked in order for these "nonmetric" procedures to be used.

However, insofar as the metric procedure of principal components analysis is concerned, we see that the major mathematical tool involves the eigenstructure of symmetric matrices. Related concepts such as the singular value decomposition of a matrix into its basic structure, quadratic forms, and matrix rank are also of interest.

From a geometric viewpoint we seek a rotation of the original basis of the space that coincides with the axes of the hyperellipsoid of points, assumed to represent the objects, in the original n-dimensional space. The eigenvalues correspond to the variances of these

new axes, and the normalized eigenvectors of the particular cross-product matrix employed are the direction cosines that define the rotation.

6.5 THE MULTIPLE DISCRIMINANT ANALYSIS PROBLEM

The third problem described in Chapter 1 concerns the development of linear composites of X_{d1} and X_{d2} with the property of maximally separating the three groups (shown in Fig. 1.5). That is, in this multivariate application, the 12 employees, based on the data of Table 6.1, were split into three groups with regard to degree of absenteeism:

Group 1–low	(employees a, b, c, d)
Group 2–intermediate	(employees e, f, g, h)
Group 3–high	(employees i, j, k, l)

While it happens to be the case here that the three groups are ordered with respect to extent of absenteeism, this is not a requirement of multiple discriminant analysis (MDA). Any polytomy consisting of a set of mutually exclusive and collectively exhaustive groups is sufficient for application of MDA.

In the sample problem application of MDA, we wish to find a linear composite of X_{d1} and X_{d2} with the property of maximizing among-group variation relative to (pooled) within-group variation. Like principal components analysis, this involves finding the eigenstructure of a matrix. However, in this case the matrix is nonsymmetric, although it, in turn, represents the product of two symmetric matrices.

6.5.1 Finding the Eigenstructure

The quantity to be maximized in MDA consists of the ratio

$$\lambda_1 = \frac{\mathbf{v}_1'\mathbf{A}\mathbf{v}_1}{\mathbf{v}_1'\mathbf{W}\mathbf{v}_1} = \frac{SS_A(\mathbf{w}_1)}{SS_W(\mathbf{w}_1)}$$

where, as it turns out, λ_1 is an eigenvalue, and **A** and **W** denote among-group and pooled within-group SSCP matrices, respectively. The vector $\mathbf{v}_1$ denotes the set of weights used to develop the linear composite (denoted as $\mathbf{w}_1$) of the original mean-corrected score matrix $\mathbf{X}_d$, while SS_A and SS_W denote the among-group and within-group sums of squares of the linear composite. In scalar notation,

$$w_{i(1)} = v_1 X_{di1} + v_2 X_{di2}$$

Let us develop these concepts, step by step.

Figure 6.6 shows the first linear composite $\mathbf{w}_1$ that we seek. We see that $\mathbf{w}_1$ makes an angle of 25° with the horizontal axis. We can project the 12 points onto $\mathbf{w}_1$ and find their discriminant scores $w_{i(1)}$. (The grand mean of these scores will be zero.) Also, we can find the three group means on $\mathbf{w}_1$ and the associated among-group and pooled

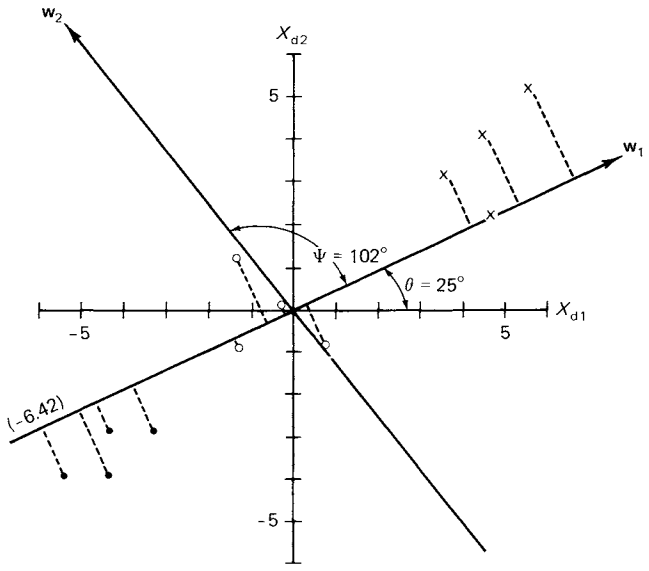

Fig. 6.6 Discriminant transformation of the mean-corrected predictor variables from Table 6.1. Key: • Group 1; ○ Group 2; x Group 3.

within-group sums of squares. According to the preceding criterion, we have found a set of scores $w_{i(1)}$ with the property that

$$\lambda_1 = \frac{SS_A(\mathbf{w}_1)}{SS_W(\mathbf{w}_1)}$$

is maximal. That is, if we find (a) the sum of squares of the three group means from the grand mean, which is zero in the case of mean-corrected data and (b) the pooled within-group sum of squares of each of the scores about their respective group means on $\mathbf{w}_1$, then (c) the ratio λ_1 of these two sums of squares is greater than that found by any other suitably normalized[5] axis in the space of Fig. 6.6.

Table 6.4 shows the preliminary calculations of interest. First, we compute each group mean on X_{d1} and X_{d2}, respectively; with equal-size groups these means, of course, sum to zero, within rounding error. Then we find the matrix of within-group deviations and the matrix of among-group deviations from group and grand mean, respectively.

From here, we compute the minor product moment of

$\mathbf{X}_k - \bar{\mathbf{X}}_k$, the matrix of within-group deviations and, similarly, the minor product moment of

$\bar{\mathbf{X}}_k - \bar{\bar{\mathbf{X}}}_k$, the matrix of among-group deviations (for $k = 1,2,\ldots,K = 3$ groups)

[5] That is, we seek a linear composite in which the coefficients v_1 and v_2 are direction cosines. Also, it should be remembered in the computation of $SS_A(\mathbf{w}_1)$ that each group mean is based on four observations.

TABLE 6.4

Preliminary Calculations for Multiple Discriminant Analysis

	Employee	X_{d1}	X_{d2}	$X_k - \bar{X}_k$ Within-group deviations		$\bar{X}_k - \bar{\bar{X}}$ Between-group deviations	
1	a	−5.25	−3.92	−1	−0.5	−4.25	−3.42
	b	−5.25	−3.92	0	−0.5	−4.25	−3.42
	c	−4.25	−2.92	0	0.5	−4.25	−3.42
	d	−3.25	−2.92	1	0.5	−4.25	−3.42
	Mean	−4.25	−3.42				
2	e	−1.25	−0.92	−0.75	−0.75	−0.5	−0.17
	f	−1.25	1.08	−0.75	1.25	−0.5	−0.17
	g	−0.25	0.08	0.25	0.25	−0.5	−0.17
	h	0.75	−0.92	1.25	−0.75	−0.5	−0.17
	Mean	−0.50	−0.17				
3	i	3.75	3.08	−1	−0.50	4.75	3.58
	j	4.75	2.08	0	−1.50	4.75	3.58
	k	4.75	4.08	0	0.50	4.75	3.58
	l	5.75	5.08	1	2.50	4.75	3.58
	Mean	4.75	3.58				

$$\mathbf{W} = (\mathbf{X}_k - \bar{\mathbf{X}}_k)'(\mathbf{X}_k - \bar{\mathbf{X}}_k); \qquad \mathbf{A} = (\bar{\mathbf{X}}_k - \bar{\bar{\mathbf{X}}})'(\bar{\mathbf{X}}_k - \bar{\bar{\mathbf{X}}})$$

$$\mathbf{W} = \begin{bmatrix} 6.75 & 1.75 \\ 1.75 & 8.75 \end{bmatrix}; \qquad \mathbf{A} = \begin{bmatrix} 163.50 & 126.50 \\ 126.50 & 98.17 \end{bmatrix}$$

$$\mathbf{T} = \mathbf{W} + \mathbf{A} = \begin{bmatrix} 170.25 & 128.25 \\ 128.25 & 106.92 \end{bmatrix}$$

to find **W**, the pooled within-group SSCP matrix, and **A**, the among-group SSCP matrix, as shown in Table 6.4. Their sum equals the total sample SSCP matrix **T**, which is also shown in Table 6.4.

The problem, as shown in Appendix A, is to maximize λ_1 with respect to $\mathbf{v}_1$. The resulting matrix equation is

$$\boxed{(\mathbf{A} - \lambda_1 \mathbf{W})\mathbf{v}_1 = \mathbf{0}}$$

Assuming that **W** is nonsingular and, hence, that $\mathbf{W}^{-1}$ exists, we can premultiply both sides of the above equation to get

$$(\mathbf{W}^{-1}\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{v}_1 = \mathbf{0}$$

with characteristic equation

$$|\mathbf{W}^{-1}\mathbf{A} - \lambda_1 \mathbf{I}| = 0$$

Note, then, that we have another eigenstructure problem, one now involving the nonsymmetric matrix $\mathbf{W}^{-1}\mathbf{A}$.

TABLE 6.5

Finding the Eigenstructure of the $\mathbf{W}^{-1}\mathbf{A}$ *Matrix*

$$\mathbf{W}^{-1} = \begin{bmatrix} 0.156 & -0.031 \\ -0.031 & 0.121 \end{bmatrix}; \qquad \mathbf{W}^{-1}\mathbf{A} = \begin{bmatrix} 21.594 & 16.698 \\ 10.138 & 7.880 \end{bmatrix}$$

Eigenvalues of $\mathbf{W}^{-1}\mathbf{A}$

$$\Lambda = \begin{bmatrix} 29.444 & 0 \\ 0 & 0.0295 \end{bmatrix};$$

Eigenvectors of $\mathbf{W}^{-1}\mathbf{A}$

$$\mathbf{V} = \begin{bmatrix} 0.905 & -0.612 \\ 0.425 & 0.791 \end{bmatrix}$$

As is the case with principal components analysis, generally the characteristic equation will have more than a single root. In fact, in this problem we shall be able to find two eigenvalues, λ_1 and λ_2, and their associated eigenvectors.

Table 6.5 shows the eigenvalues and eigenvectors obtained for this sample application. Note the parallel between this problem and the principal components problem. In each case we are finding the eigenstructure of a matrix, but here the matrix is nonsymmetric.

From Table 6.5 we see that the first discriminant function displays a relatively large eigenvalue of

$$\lambda_1 = 29.444$$

with associated, and normalized, eigenvector

$$\mathbf{v}_1 = \begin{bmatrix} 0.905 \\ 0.425 \end{bmatrix}$$

representing an angle of 25° from the horizontal axis.

Also, similar to principal components analysis, we can obtain a second discriminant function $\mathbf{w}_2$ with scores that are uncorrelated with those of the first function. The eigenvalue associated with $\mathbf{w}_2$ is

$$\lambda_2 = 0.0295$$

with normalized eigenvector

$$\mathbf{v}_2 = \begin{bmatrix} -0.612 \\ 0.791 \end{bmatrix}$$

Note that λ_2 is much smaller than λ_1; for all practical purposes it appears that a single discriminant function might account for these data.

In general, with K groups and n predictors one obtains

$$\boxed{\min(K-1, n)}$$

different discriminant functions; here, of course $K - 1 = n = 2$, and we note that two functions are obtained. Usually, however, the number of predictors will greatly exceed the number of groups, and a great deal of parsimony can often be achieved by the use of discriminant scores.

As before, discriminant scores are found by projecting the points onto the discriminant axes. For example, the discriminant score of the first observation on $\mathbf{w}_1$ is

$$w_{1(1)} = 0.905(-5.25) + 0.425(-3.92) = -6.42$$

as shown in Fig. 6.6.

We also observe in Fig. 6.6 that $\mathbf{v}_1$ and $\mathbf{v}_2$ are not orthogonal, even though the scores on $\mathbf{w}_1$ versus those on $\mathbf{w}_2$ are uncorrelated. From the matrix of eigenvectors in Table 6.5 we can compute the cosine between $\mathbf{v}_1$ and $\mathbf{v}_2$ as follows:

$$\cos \Psi = (0.905 \quad 0.424)\begin{bmatrix} -0.612 \\ 0.791 \end{bmatrix} = -0.21, \qquad \text{so that} \quad \Psi = 102^\circ$$

Thus, the angle Ψ separating $\mathbf{v}_1$ and $\mathbf{v}_2$ is $90^\circ + 12^\circ = 102^\circ$, as shown in Fig. 6.6.

6.5.2 Statistical Significance and Classification

It is one thing, of course, to find linear composites with the properties described above; it is quite another to test their statistical significance and to use them for classifying observations. Accordingly, each of these problems is taken up, in turn. At this point we have found two eigenvalues:

$$\lambda_1 = 29.444; \qquad \lambda_2 = 0.0295$$

Bartlett (1947) has proposed a statistic that can be used to test the significance of the discriminant functions (actually, their eigenvalues).

Bartlett's statistic starts out by testing the null hypothesis that group centroids are all equal in the *full* discriminant space, in this case involving both the $\mathbf{w}_1$ and $\mathbf{w}_2$ axes.

Bartlett's statistic is expressed as follows:

$$V = 2.3026[m - 1 - (n + K)/2] \sum_{i=1}^{r} \log(1 + \lambda_i)$$

where m denotes sample size, n denotes number of predictor variables, K denotes number of groups, and λ_i denotes the ith eigenvalue ($i = 1, 2, \ldots, r$). In terms of the sample problem,

$$\begin{aligned} V &= 2.3026[12 - 1 - (2 + 3)/2](\log 30.444 + \log 1.0295) \\ &= 2.3026(8.5)(1.48350 + 0.01263) \\ &= 29.035 + 0.247 \\ &= 29.282 \end{aligned}$$

Bartlett's V statistic is approximately distributed as chi square with $n(K - 1) = 4$ degrees of freedom. In the sample problem, V is significant beyond the 0.01 alpha level.

However, one wonders whether the second discriminant function, whose eigenvalue is almost zero, adds anything beyond the first. Fortunately, V can be decomposed into the separate parts

$$V_1 = 29.035; \qquad V_2 = 0.247$$

The first portion V_1 has already been tested in the context of V. However, if V_1 is "partialed out," is V_2 statistically significant? As it turns out, V_2 can be tested in the same way that V was tested: $V_2 = 0.247$ is also approximately distributed as chi square with

$$n(K-1)-(n+K-2) = (n-1)(K-2) = 1$$

degree of freedom. This approximate chi square ($V_2 = 0.247$) is clearly nonsignificant. Not surprisingly, we conclude that only the first discriminant function need be retained.

Bartlett's procedure can be used for more than two discriminant functions in a similar manner. Had a third discriminant function been involved, its associated degrees of freedom would be $(n-2)(K-3)$; those associated with a fourth discriminant function would be $(n-3)(K-4)$, and so on. However, the reader interested in applying this test in substantive research should be aware of its assumptions (Harris, 1975; pp. 109-113).

In the sample problem it is not hard to see why only the first discriminant function is significant. The following ratio:

$$\frac{\lambda_1}{\lambda_1+\lambda_2} = \frac{29.444}{29.444+0.0295} = 0.999$$

shows that $\mathbf{w}_1$ exhausts virtually all of the variation in the discriminant space.

Classifying the twelve observations by means of $\mathbf{w}_1$, the retained and significant discriminant function, is quite straightforward. All that is entailed is to compute a discriminant score for each observation, according to

$$w_{(1)} = v_1 X_{d1} + v_2 X_{d2}$$

One also computes the discriminant scores for the three group means

$$\bar{w}_1(\text{Group 1}) = 0.905(-4.25) + 0.425(-3.42) = -5.30$$

$$\bar{w}_1(\text{Group 2}) = 0.905(-0.50) + 0.425(-0.17) = -0.52$$

$$\bar{w}_1(\text{Group 3}) = 0.905(4.75) + 0.425(3.58) = 5.82$$

One then assigns each observation to that group whose mean score on $\mathbf{w}_1$ is closest to the individual score, $w_{i(1)}$.

When this procedure is implemented for the sample problem, it turns out that all twelve cases are correctly assigned to their respective groups. Had $\mathbf{w}_2$ also been statistically significant and retained for classification purposes, the classification procedure would have been modified to involve the computation of Euclidean distances between each individual observation and each group centroid in discriminant function space.[6] Each observation would then be assigned to the group whose centroid, in discriminant function space, was nearest.

It should be mentioned, however, that the use of Bartlett's statistic and the classification rules enumerated above only scratch the surface of the topics of statistical

[6] It should be noted that in discriminant function space the pooled within-group SSCP matrix would first be spherized by means of the procedure described in Section 5.9.2; it is *this* space in which ordinary Euclidean distance is appropriate (given equal prior probabilities and equal costs of misclassification) for assigning objects to groups.

significance and assignment. For example, Bartlett's statistic can be modified by Schatzoff's tables (Schatzoff, 1966) to produce an exact test. Also, other tests (Rao, 1952) are available for testing the null hypothesis of group centroid equality in the full discriminant space.

The classification rules described above also need modification in cases where the prior probabilities of inclusion differ across the groups or where the costs of misclassification differ. Modern approaches to the problem (e.g., Eisenbeis and Avery, 1972) formulate the classification task in terms of statistical decision theory. As such, both prior probabilities of an observation belonging to each of the groups and costs of misclassification can be explicitly introduced into the assignment procedure.

6.5.3 Other Aspects of Multiple Discriminant Analysis

One of the questions posed in Chapter 1 concerned the relative importance of the two predictors X_{d1} and X_{d2} in effecting group discrimination. In the case of correlated predictors, this represents an ambiguous question and shares, along with multiple regression and other multivariate techniques, the difficulties of parceling out variance among nonorthogonal predictors. While we do not go into this question in detail, a few procedures that have been suggested for ascribing relative importance to X_{d1} versus X_{d2} can be mentioned.

First, the entries in the normalized eigenvector $\mathbf{v}_1$ are 0.905 and 0.425 for X_{d1} and X_{d2}, respectively. These are analogous to partial regression coefficients in multiple regression. To convert them into standardized (beta-type) numbers, each is multiplied by that predictor variable's pooled within-group standard deviation:[7]

$$\text{Standardized weight } (X_{d1}); \qquad 0.905 \times \sqrt{0.75} = 0.783$$

$$\text{Standardized weight } (X_{d2}); \qquad 0.425 \times \sqrt{0.972} = 0.419$$

As can be noted, on either a standardized or nonstandardized basis, X_{d1} receives the larger weight.

Cooley and Lohnes (1971) recommend what they call structure correlations to ascertain predictor importance. These are merely the product-moment correlations between scores on each original variable and the discriminant scores. In this example they turn out to be

$$\text{Structure correlation } (X_{d1}) = 0.998$$

$$\text{Structure correlation } (X_{d2}) = 0.976$$

In this case both predictors correlate highly with the retained discriminant function $\mathbf{w}_1$, although the correlation for X_{d1} is slightly greater.

Still other procedures, such as Bock and Haggard's (1968) step-down F ratios, can be employed to measure the relative importance of various predictors. However, we do not delve into these more esoteric methods, other than to say that the question of ascribing "relative importance" remains ambiguous in the case of correlated predictor variables no matter what procedure is used.

[7] Other standardization procedures, based on multiplication of each discriminant coefficient by the total-sample (as opposed to pooled within-group) standard deviation of the variables of interest, are also in use.

Another topic of interest concerns the relationship of MDA to other multivariate techniques. For example, an intimate connection exists between MDA and principal components analysis. Without delving deeply into the technical details, it turns out that a preliminary "spherizing" of the data matrix via

$$\boxed{\mathbf{X}_d\mathbf{W}^{-1/2}}$$

results in a new set of coordinates with spherical (pooled) within-group variation.

One can then find the eigenstructure of the matrix

$$\boxed{\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}}$$

so as to satisfy the equation

$$\boxed{[\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}]\mathbf{Q} = \mathbf{Q}\Lambda}$$

where $\mathbf{Q}$ is orthogonal and Λ is diagonal.[8] The final transformation is then

$$\boxed{\mathbf{X}_d\mathbf{W}^{-1/2}\mathbf{Q}}$$

which, of course, is a spherizing transformation followed by a rotation of the spherized within-group variation to principal axes, on the basis of among-group variation. This idea was described in the discussion in Chapter 5 of the simultaneous diagonalization of two different quadratic forms.[9]

Probably the most important point to mention, however, is that MDA is one member of the same general family that includes

1. canonical correlation,
2. multivariate analysis of variance and covariance,
3. categorical canonical correlation.

The linkage among these multiple-criterion techniques is provided by a generalized canonical correlation model that allows for dummy variables on one or both sides of the equation. For example, one could have developed a multiple discriminant function for the sample problem by means of a canonical correlation in which the criterion variables were represented by the dummies

1	0	(Group 1)
0	1	(Group 2)
0	0	(Group 3)

By a similar judicious choice of dummy and continuous variables, one can find linear composites of both the criterion and predictor batteries that relate to any of the specific multivariate techniques described above, and in Chapter 1 as well.

[8] In this illustration Λ denotes the diagonal matrix of eigenvalues of $\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}$, while $\mathbf{Q}$ denotes the associated matrix of eigenvectors.

[9] Still other procedures are available for finding the eigenstructure of $\mathbf{W}^{-1}\mathbf{A}$ (see Overall and Klett, 1972).

Full discussion of the interrelationships among techniques would take us far beyond the scope of the book. As we have illustrated in Chapter 5, however, the eigenstructure of nonsymmetric matrices and the simultaneous diagonalization of two different quadratic forms figure prominently in the computation of discriminant functions for three or more groups. These concepts are also central in canonical correlation, multivariate analysis of variance and covariance, and categorical canonical correlation; in the last case, all variables are expressed as dummies.

6.6 A PARTING LOOK AT MULTIVARIATE TECHNIQUE CLASSIFICATION

In Chapter 1 a number of characteristics were enumerated that provided guidance for classifying the large, and still growing, variety of multivariate techniques. In particular, the following descriptors represented the main bases of classification:

1. whether the data matrix is kept intact versus partitioned into criterion and predictor subsets;
2. the number of variables in each subset (if partitioning is undertaken);
3. the types of scales by which the variables are measured.

At this point, however, the various types of linear transformations described in Chapters 4 and 5 are behind us. And, even in the introductory material of Chapter 1, it was suggested that multivariate analysis is largely concerned with transformations for matching one set of numbers, such as a data vector, a data matrix, or a linear composite, with some other set of numbers.

The degree of matching is usually assessed by a residual sum of squared deviations or some other measure that can be related to this. This idea was illustrated at the beginning of the present chapter in the context of multiple regression. Here we desired to minimize the quantity

$$\sum_{i=1}^{m} e_i^2 = (Y_i - \hat{Y}_i)^2$$

where Y_i denotes a datum, and $\hat{Y}_i$ denotes a predicted value of Y_i. *As a further aid to technique classification, we now take the view that multivariate techniques may differ according to the nature of the allowable transformations and the properties that the transformed matrices exhibit in the matching process.*

Partly by way of review of Chapters 4 and 5 and partly by way of prologue, let us list the major classes of transformations that vectors or matrices can undergo in the course of achieving various types of matching. For illustration, let us assume a general data matrix, denoted by $\mathbf{X}$, of m rows and n columns ($m \geqslant n$).

Our objective here will be to recapitulate various types of transformations described in earlier chapters as a way to make explicit the present descriptor, the nature of the linear transformation, for characterizing multivariate techniques.

6.6.1 Types of Transformations

By way of an overview, Fig. 6.7 shows a directed graph of the transformations that are considered. This list of transformations is not meant to be exhaustive. However, those shown in Fig. 6.7 appear to be the most frequently encountered ones in multivariate analysis. We consider the more general classes first, followed by the more restricted transformations.

We shall let $\mathbf{T}$ denote an arbitrary matrix. The matrices $\mathbf{U}$ and $\mathbf{V}'$ denote either orthogonal matrices or orthonormal sections, while Δ denotes a diagonal matrix. From Section 5.7 we know that $\mathbf{T}$ can always be decomposed into the triple product

$$\boxed{\mathbf{T} = \mathbf{U}\Delta\mathbf{V}'}$$

We shall take advantage of this type of singular value (Eckart–Young, 1936) decomposition in describing various special cases of a general linear (or affine) transformation.

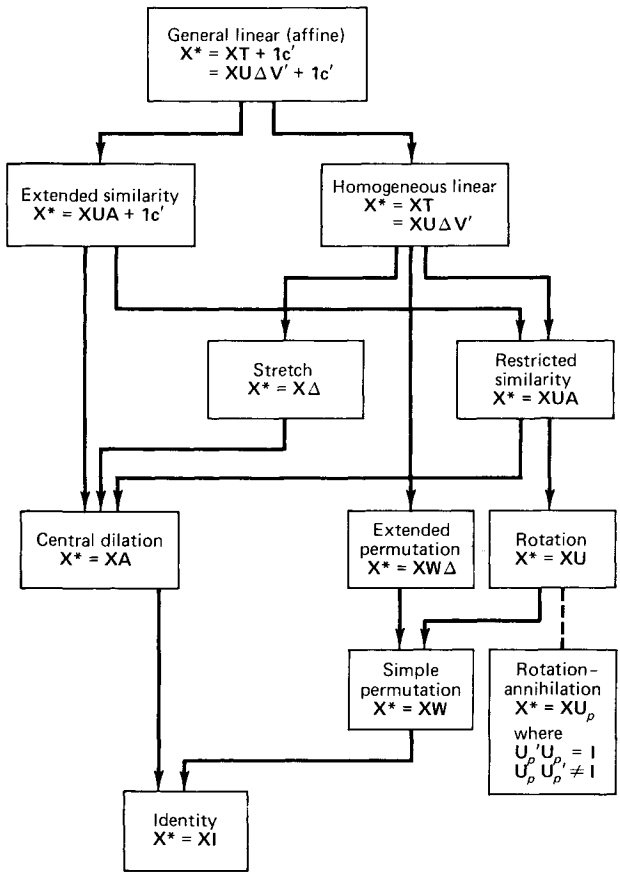

Fig. 6.7 A directed graph of various types of linear transformations.

6.6.1.1 General Linear (Affine) Transformation The most general transformation of **X** to be considered is an affine transformation, defined as

$$\boxed{\mathbf{X}^* = \mathbf{XT} + \mathbf{1c}'}$$

where $\mathbf{T}(=\mathbf{U}\Delta\mathbf{V}')$ denotes an arbitrary linear transformation. The matrix product of the $m \times 1$ unit column vector **1** and $\mathbf{c}'$ a $1 \times n$ row vector of constants defines the permissible shift of origin (as illustrated in Chapter 4 in the case of a centroid-centered orientation).[10]

6.6.1.2 Homogeneous Linear Transformation A homogenous linear transformation can be defined as

$$\boxed{\mathbf{X}^* = \mathbf{XT}}$$

with no shift in origin, but **T**, defined as before, is otherwise not restricted.

6.6.1.3 Similarity Transformations An extended similarity transformation involves a rotation, achieved by the orthogonal matrix **U**, a central dilation, effected by the scalar matrix **A**, and a shift in origin:

$$\boxed{\mathbf{X}^* = \mathbf{XUA} + \mathbf{1c}'}$$

where **1** and $\mathbf{c}'$ are defined as before.

A restricted similarity transformation is a special case of this in which no shift in origin is permitted:

$$\boxed{\mathbf{X}^* = \mathbf{XUA}}$$

where **U** and **A** are defined as before.

6.6.1.4 Rotation As illustrated in earlier chapters, a rotation is a transformation that is carried out by an orthogonal matrix. This type of matrix is denoted by **U**, where $\mathbf{U}'\mathbf{U} = \mathbf{UU}' = \mathbf{I}$. The transformation is written as

$$\boxed{\mathbf{X}^* = \mathbf{XU}}$$

If the determinant $|\mathbf{U}| = 1$, then a proper rotation of **X** is entailed. If $|\mathbf{U}| = -1$, then a rotation of **X** followed by a reflection is entailed (i.e., an improper rotation).

6.6.1.5 Rotation–Annihilation One type of transformation stipulates that only the condition $\mathbf{U}_p'\mathbf{U}_p = \mathbf{I}$ be met; that is, $\mathbf{U}_p$ can be an orthonormal section (rectangular rather than square) whose columns are mutually orthogonal and of unit length. This amounts to a rotation followed by annihilation of some dimensions.

In Fig. 6.7 we show this transformation with a dotted rather than solid line. This is because $\mathbf{U}_p$ is not a special case of **U** for the reason that $\mathbf{U}_p\,\mathbf{U}_p' \neq \mathbf{I}$. As such, $\mathbf{U}_p$ is rather tangentially related to the overall schema of Fig. 6.7.

[10] Note that an affine transformation is nonhomogeneous in the sense that there are no fixed points (e.g., the 0 or origin vector) under this type of mapping. However, Section 4.4.1 shows how it can be carried out via matrix multiplication.

6.6.1.6 Permutation An extended permutation permits both a reordering of dimensions and a stretch or rescaling of the configuration. This is written as

$$\boxed{X^* = XW\Delta}$$

where W is a permutation matrix and Δ is diagonal. As recalled, a permutation matrix is an orthogonal matrix, all of whose entries are either 0 or 1, that changes the order of dimensions.

A simple permutation is written as

$$\boxed{X^* = XW}$$

with W defined as before.

6.6.1.7 Stretch A stretch transformation involves a simple rescaling of the configuration by a diagonal matrix Δ. That is,

$$\boxed{X^* = X\Delta}$$

6.6.1.8 Central Dilation A special case of a stretch transformation involves a central dilation, given by the scalar matrix A. That is,

$$\boxed{X^* = XA(=AX)}$$

6.6.1.9 Identity Transformation A special case of the central dilation transformation is the identity transformation

$$\boxed{X^* = IX = XI = X}$$

where I, of course, is the identity matrice.

While still other combinations are possible, the preceding ones cover most cases of practical interest.

6.6.2 Constructing the Classification

With the various geometric illustrations presented in the preceding section, it is now appropriate to discuss the nature of multivariate techniques from the standpoint of configuration matching. We consider the following classes:

1. vector-matrix matching,
2. matching of two matrices,
3. matching of three or more matrices,
4. matching of a data-based and an internally derived matrix.

Within each of these classes, two additional aspects are discussed:

1. types of scores—continuous or binary (dummy variable),
2. type of permissible transformation applicable to each matrix or vector.

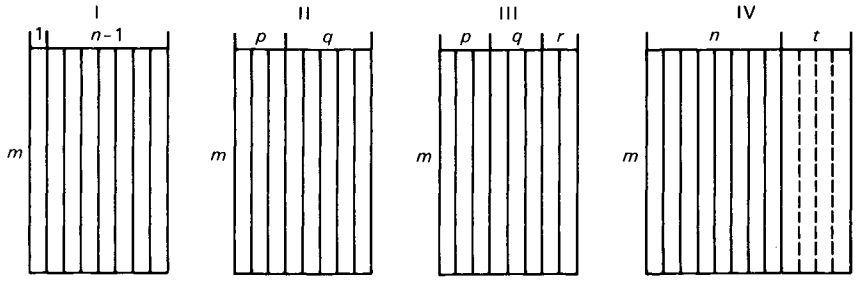

Fig. 6.8 Illustrative partitionings of data matrix.

Each of the above major classes is examined in turn. To assist us in this regard, we reproduce in Fig. 6.8 the schema that appeared as in Fig. 1.1. However, now we emphasize the nature of the linear transformation.

6.6.2.1 Vector-Matrix Matching Panel I of Fig. 6.8 is the prototype of the family of multivariate techniques illustrated by the matrix equation

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}}$$

where $\mathbf{b}$ is a vector of combining weights, and $\hat{\mathbf{y}}$ is a set of predicted values for $\mathbf{y}$, the criterion variable. As illustrated earlier, by allowing $\mathbf{y}$ or $\mathbf{X}$ to be mixtures of continuous or binary-coded variables, this family is broad enough to include

1. multiple (and simple) regression;
2. two-group discriminant analysis, where $\mathbf{y}$ is binary valued;
3. analysis of variance and covariance, where some columns of $\mathbf{X}$ are binary valued;
4. binary-valued regression, where both $\mathbf{y}$ and $\mathbf{X}$ are binary valued.

In least-squares theory the scalars R^2 or η^2 (eta squared) are usually the quantities being maximized.[11] Both R^2, in the context of regression, and η^2, in the context of analysis of variance, are invariant over linear transformations of $\mathbf{y}$ or $\mathbf{X}$. Moreover, both R^2 and η^2 can be simply related to the criterion of minimizing the sum of the squares of $\mathbf{y} - \hat{\mathbf{y}}$, as described earlier.

6.6.2.2 Matching Two Matrices In Panel II of Fig. 6.8 we have the case of two matrices, $\mathbf{Y}_{m \times p}$ and $\mathbf{X}_{m \times q}$, and are interested in the association between these two batteries of variables. If we assume that both sets of variables represent continuous values, the canonical correlation problem can be represented by separate affine transformations of $\mathbf{Y}$ and $\mathbf{X}$ such that each pair of linear composites is most congruent with each other, subject to being uncorrelated with previously "extracted" composites. This uncorrelatedness condition is an illustration of the kinds of restrictions that may be placed on the transformed values.

[11] The scalar η^2 (eta squared) is computed in just the same way as R^2 except for the fact that all predictor values are dummy variables. This is equivalent to

$$\eta^2 = \frac{SS_A}{SS_T}$$

while SS_A denotes the among-group sum of squares and SS_T denotes the total-group sum of squares.

Other possibilities come to mind, however. For example, one could allow only a separate homogenous linear transformation of each matrix with no shift in origin permitted. Or, one could permit a shift in origin but require each transformation to be an extend similarity transformation which, as shown earlier, is less general than an affine transformation.

In other kinds of applications we may desire **Y** to remain fixed (i.e., transformed by an identity matrix) but permit **X** to be transformed by an affine transformation, extended similarity, or a similarity transformation. Some "procrustes" solutions, as used in matching factor score solutions from different studies, are of this general type (Rummel, 1970).

Still other restrictions are possible. Schönemann and Carroll (1970) describe a matching procedure in which one matrix undergoes an extended similarity transformation, while the other undergoes either (a) an extended similarity, (b) a similarity, (c) a rotation, or (d) an identity transformation. Cliff's procedure (Cliff, 1966) allows a similarity transformation on one side and a similarity, rotation, or identity transformation on the other.

If one matrix consists of two or more binary-valued variables, we have an instance of either multiple discriminant analysis or multivariate analysis of variance, depending upon how one frames the problem. From the standpoint of permissible transformations, however, the techniques are similar. That is, one can formulate either a multiple discriminant problem or a multivariate analysis of variance problem in terms of the canonical correlation model with one of the two matrices represented by binary-valued dummies. Generally, however, we are interested in special kinds of output that are related to the particular procedure employed. Therefore, while one *could* use a canonical correlation program to find discriminant weights, ordinarily we would not do so since we would be interested in various ancillary outputs as well.

If both data sets consist of dummy variables, we may have a case of categorical canonical correlation or categorical conjoint measurement (Carroll, 1973). Insofar as the solution to the problem is concerned, these techniques are special cases of canonical correlation in which *both* matrices consist of dummy variables.

Variations can be developed, however. For example, Horst (1956) describes a type of multiple discriminant analysis in which the dummy-variable criterion matrix, defining group membership, remains fixed. The predictor matrix is transformed linearly to best match it, subject to the predicted values maintaining the same column means and variances as the columns of the criterion-variable matrix.

6.6.2.3 Three or More Matrices Heretofore, we have described multivariate analysis of covariance in terms of a matrix of criterion variables and a matrix of predictor variables. The latter matrix consists of a mixture of dummy variables, the design variables, and covariates, whose effect on the criterion variables we desire to remove. Alternatively, we can partition the data matrix into three matrices: criterion, design dummies, and covariates, as illustrated in Panel III of Fig. 6.8.

Problems involving three, or more, matrices fall into two major types:

1. a multivariate analysis of covariance situation, or multiple, partial correlation (Cooley and Lohnes, 1971), in which one of the data matrices consists of a set of

covariates, moderators, or contingency variables whose effect is to be removed before considering the association between the remaining matrices;

2. a generalized canonical correlation situation where the status of all three, or more, matrices is considered to be the same (in this case, we extend two-group canonical correlation to cover three or more matrices).

Multivariate analysis of covariance problems occurs frequently in the behavioral and administrative sciences. For example, one may set up various experiments in which several response measures are sought from the subjects and, furthermore, certain covariates like task familiarity, education, and IQ level are also included in the analysis.

In multivariate analysis of covariance one matrix, the response matrix, is typically made up of continuous scores, while the design matrix is typically made up of dummy variables. The matrix of covariates is usually made up of continuous scores. However, this is not necessary. In principle, any (or all three) of the matrices could consist of continuous or binary-valued scores, or, indeed, as mixtures. In this class of problems one generally allows affine transformations to be applied to any of the three matrices, in the spirit of two-set canonical correlation.

Generalized canonical correlation, employing three or more data-based matrices of equal status, is concerned primarily with configuration matching. Horst (1961), Carroll (1968), and Kettenring (1972) have all proposed models for this type of problem. For example, in the Carroll and Chang approach, an $r + 1$st space is defined such that the r original spaces, each consisting of the same m observations on r sets of variables, are transformed to match it as well as possible. This procedure allows an affine transformation of each "contributing" configuration.

While generalized canonical correlation has usually been considered in the context of all scores being continuous, this, again, is not necessary provided that the researcher's interest is centered on data description and summarization, rather than on statistical inference. Binary-valued scores, or mixtures of continuous and binary valued, can be dealt with just as readily. Again, affine transformations would generally be permitted.

6.6.2.4 Matching Based on an Internal Criterion Multivariate techniques can also cover the possibility of deriving a matrix (e.g., a "latent" matrix) that best reproduces the scores of a data-based matrix, or some matrix derived from it, subject to meeting certain internal criteria. For example, in our earlier discussion of principal components analysis employing the covariance matrix as input, we found a rotation of the space whose successive dimensions accounted for the greatest amount of residual variance. This can also be viewed as defining successively higher-dimensional subspaces that maximize variance for that dimensionality.

As pointed out earlier, most factor analytic techniques (e.g., principal components analysis) are not independent of scale. That is, different results are obtained depending upon whether the averaged raw sums of squares and cross products, covariance, or correlation matrix is the one being factored. A major exception to this is canonical factor analysis (McDonald, 1968). This technique produces results that are comparable across various types of data scaling. That is, the solution obtained from one type of scaling can be readily transformed to a solution obtained from a different scaling of the original data matrix. Maximum likelihood factor analysis (Van de Geer, 1971) also yields results that are independent of scale.

As pointed out earlier, some types of factor rotation (e.g., Varimax) are based on achieving internal criteria of "simple" structure (Horst, 1966). Simple structure entails the idea of a hypothetical zero–one matrix in which each variable is, ideally, supposed to load with unity on one factor only (i.e., with zeros appearing elsewhere). In this sense a type of "matching" of one matrix to another is also involved.

In brief, a useful descriptor in characterizing multivariate methods is the type of linear transformation involved in the matching process and the restrictions placed on the nature of the transformed data. As we have illustrated, if only briefly, the various possibilities are extensive. Combined with the descriptors of Chapter 1, the type of linear transformation descriptor provides a rather comprehensive system for characterizing all current multivariate techniques. Moreover, it can be suggestive of still other combinations to be invented.

6.7 SUMMARY

In this chapter we have tried to show how the mathematical tools developed in the foregoing chapters and the appendixes underlie the formulation and solution of various multivariate techniques. In particular, multiple regression, principal components analysis, and multiple discriminant analysis were presented as prototypical techniques.

In multiple regression, the concepts of matrix inversion, determinants, and matrix rank figured prominently in the solution. We also showed how the multiple regression problem could be described geometrically, both from the standpoint of a response surface or point model and from the standpoint of a vector model. Finally, the notion of generalized regression, as a least-squares model that encompasses analysis of variance and covariance, two-group discrimination, and binary-valued regression, was illustrated graphically.

Principal components, the technique described next, entailed the rotation of a set of basis vectors to a new orthogonal basis with projections whose variance was sequentially maximal. The concepts of matrix eigenstructure of a symmetric matrix, matrix rank, and quadratic forms were most important here.

Multiple discriminant analysis then provided us with a procedure for extending our discussion to cover the eigenstructure of a nonsymmetric matrix. The simultaneous diagonalization of two different quadratic forms represented the central concept from matrix algebra. Geometrically, this entailed a rotation to align the configuration with the principal axes of the within-group SSCP matrix, a spherizing along these axes and then a further rotation to principal axes of the transformed among-group SSCP matrix.

The various matrix transformations of Chapter 4 were then recapitulated and organized into a framework within which various multivariate techniques could be described. In conjunction with the descriptors of Chapter 1, the specific nature of the linear transformation provided a useful way to characterize various multivariate procedures.

This chapter (and the entire book) has served as something of a prologue for textbooks dealing with multivariate methods per se. A large number of such texts are listed in the references, although no attempt has been made to be exhaustive. We do hope, however, that this book will make the going a bit easier as the reader delves more deeply into the subject matter of multivariate analysis.

REVIEW QUESTIONS

1. Using the data of Table 6.1,

a. compute the parameter values of Y regressed on X_1 and X_2 by means of the covariance matrix;

b. repeat the process, now employing the correlation matrix;

c. regress Y on X_1 and X_1^2 (in place of X_2) by means of a raw cross-products matrix. How does the R^2 of this compare with the simple squared correlation found from the regression of Y on X_1 alone?

2. Again using the data of Table 6.1,

a. find the principal components of the correlation matrix $\mathbf{R}$, obtained from X_1 and X_2. How do the eigenvectors compare with those obtained from $\mathbf{C}$, the covariance matrix?

b. find the principal components of the averaged raw cross-products matrix $\mathbf{X}'\mathbf{X}/m$, obtained from X_1 and X_2;

c. returning to Section 6.4.1, find the multiple regression of Y on the two columns of component scores computed from the covariance matrix $\mathbf{C}$. How does the value of this R^2 compare to that obtained by regressing Y on X_1 and X_2 originally? What happens to the squared correlation if only the scores on the first component are used?

d. find the eigenstructure of $\mathbf{C}$, the covariance matrix based on all three variables, Y, X_1, and X_2. Compare the eigenvectors obtained here with those appearing in Fig. 6.5.

3. Again using the data of Table 6.1,

a. perform a three-group discriminant analysis on the standardized columns $\mathbf{x}_{s1}$ and $\mathbf{x}_{s2}$ using the same group designation as before. How do these discriminant weights compare with those found earlier?

b. using the procedure of Chapter 5, perform a simultaneous diagonalization of the $\mathbf{W}$ and $\mathbf{A}$ matrices in Table 6.4 and compare your results with those of Table 6.5.

c. split the mean-corrected columns, X_{d1} and X_{d2} in Table 6.1, into two groups (viz., the first six versus the second six employees) and compute a two-group discriminant function. What simplifications in the computations are noted in this case?

4. Regress Y on X_1 and X_2, where the latter predictor is now dichotomized with $X_2 < 5$ receiving the code value 0 and $X_2 \geqslant 5$ receiving the code value 1.

a. How does the regression equation compare to the original shown in Table 6.2?

b. What is the effect on R^2 and the proportion of cumulative variance column, as illustrated in Table 6.2?