Chapter 14
Factor analysis

14.1 INTRODUCTION

Factor analysis is a method for investigating whether a number of variables of interest Y_1, Y_2, ..., Y_l, are linearly related to a smaller number of unobservable factors F_1, F_2, ..., F_k.

The fact that the factors are not observable disqualifies regression and other methods previously examined. We shall see, however, that under certain conditions the hypothesized factor model has certain implications, and these implications in turn can be tested against the observations. Exactly what these conditions and implications are, and how the model can be tested, must be explained with some care.

14.2 AN EXAMPLE

Factor analysis is best explained in the context of a simple example. Students entering a certain MBA program must take three required courses in finance, marketing and business policy. Let Y_1, Y_2, and Y_3, respectively, represent a student's grades in these courses. The available data consist of the grades of five students (in a 10-point numerical scale above the passing mark), as shown in Table 14.1.

Table 14.1
Student grades

Student no.	Grade in: Finance, Y_1	Marketing, Y_2	Policy, Y_3
1	3	6	5
2	7	3	3
3	10	9	8
4	3	9	7
5	10	6	5

It has been suggested that these grades are functions of two underlying factors, F_1 and F_2, tentatively and rather loosely described as quantitative ability and verbal ability, respectively. It is assumed that each Y variable is linearly related to the two factors, as follows:

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11} F_1 + \beta_{12} F_2 + e_1 \\ Y_2 &= \beta_{20} + \beta_{21} F_1 + \beta_{22} F_2 + e_2 \\ Y_3 &= \beta_{30} + \beta_{31} F_1 + \beta_{32} F_2 + e_3 \end{aligned} \tag{14.1}$$

The error terms e_1, e_2, and e_3, serve to indicate that the hypothesized relationships are not exact.

In the special vocabulary of factor analysis, the parameters β_{ij} are referred to as *loadings*. For example, β_{12} is called the loading of variable Y_1 on factor F_2.

In this MBA program, finance is highly quantitative, while marketing and policy have a strong qualitative orientation. Quantitative skills should help a student in finance, but not in marketing or policy. Verbal skills should be helpful in marketing or policy but not in finance. In other words, it is expected that the loadings have roughly the following structure:

	Loading on:	
Variable, Y_i	F_1, β_{i1}	F_2, β_{i2}
Y_1	+	0
Y_2	0	+
Y_3	0	+

The grade in the finance course is expected to be positively related to quantitative ability but unrelated to verbal ability; the grades in marketing and policy, on the other hand, are expected to be positively related to verbal ability but unrelated to quantitative ability. Of course, the zeros in the preceding table are not expected to be exactly equal to zero. By '0' we mean approximately equal to zero and by '+' a positive number substantially different from zero.

It may appear that the loadings can be estimated and the expectations tested by regressing each Y against the two factors. Such an approach, however, is not feasible because the factors cannot be observed. An entirely new strategy is required.

Let us turn to the process that generates the observations on Y_1, Y_2 and Y_3 according to (14.1). The simplest model of factor analysis is based on two assumptions concerning the relationships (14.1). We shall first describe these assumptions and then examine their implications.

A1: The error terms e_i are independent of one another, and such that $E(e_i) = 0$ and $Var(e_i) = \sigma_i^2$.

One can think of each e_i as the outcome of a random draw with replacement from a population of e_i-values having mean 0 and a certain variance σ_i^2. A similar assumption was made in regression analysis (Section 3.2).

> *A2*: The unobservable factors F_j are independent of one another and of the error terms, and are such that $E(F_j) = 0$ and $Var(F_j) = 1$.

In the context of the present example, this means in part that there is no relationship between quantitative and verbal ability. In more advanced models of factor analysis, the condition that the factors are independent of one another can be relaxed. As for the factor means and variances, the assumption is that the factors are standardized. It is an assumption made for mathematical convenience; since the factors are not observable, we might as well think of them as measured in standardized form.

Let us now examine some implications of these assumptions. Each observable variable is a linear function of independent factors and error terms, and can be written as

$$Y_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + (1)e_i.$$

The variance of Y_i can be calculated by applying the result in Appendix A.11:

$$\begin{aligned} Var(Y_i) &= \beta_{i1}^2 Var(F_1) + \beta_{i2}^2 Var(F_2) + (1)^2 Var(e_i) \\ &= \beta_{i1}^2 + \beta_{i2}^2 + \sigma_i^2. \end{aligned}$$

We see that the variance of Y_i consists of two parts:

$$Var(Y_i) = \underbrace{\beta_{i1}^2 + \beta_{i2}^2}_{\text{communality}} + \underbrace{\sigma_i^2}_{\text{specific variance}}.$$

The first, the *communality* of the variable, is the part that is explained by the common factors F_1 and F_2. The second, the *specific variance*, is the part of the variance of Y_i that is not accounted by the common factors. If the two factors were perfect predictors of grades, then $e_1 = e_2 = e_3 = 0$ always, and $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0$.

To calculate the covariance of any two observable variables, Y_i and Y_j, we can write

$$\begin{aligned} Y_i &= \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + (1)e_i + (0)e_j, \text{ and} \\ Y_j &= \beta_{j0} + \beta_{j1}F_1 + \beta_{j2}F_2 + (0)e_i + (1)e_j, \end{aligned}$$

and apply the result in Appendix A.12 to find

$$\begin{aligned} Cov(Y_i, Y_j) &= \beta_{i1}\beta_{j1}Var(F_1) + \beta_{i2}\beta_{j2}Var(F_2) + (1)(0)Var(e_i) \\ &\quad + (0)(1)Var(e_j) \\ &= \beta_{i1}\beta_{j1} + \beta_{i2}\beta_{j2}. \end{aligned}$$

We can arrange all the variances and covariances in the form of the following table:

	Variable:		
Variable:	Y_1	Y_2	Y_3
Y_1	$\beta_{11}^2 + \beta_{12}^2 + \sigma_1^2$	$\beta_{21}\beta_{11} + \beta_{22}\beta_{12}$	$\beta_{31}\beta_{11} + \beta_{32}\beta_{12}$
Y_2	$\beta_{11}\beta_{21} + \beta_{12}\beta_{22}$	$\beta_{21}^2 + \beta_{22}^2 + \sigma_2^2$	$\beta_{21}\beta_{31} + \beta_{22}\beta_{32}$
Y_3	$\beta_{11}\beta_{31} + \beta_{12}\beta_{32}$	$\beta_{21}\beta_{31} + \beta_{22}\beta_{32}$	$\beta_{31}^2 + \beta_{32}^2 + \sigma_3^2$

We have placed the variances of the Y variables in the diagonal cells of the table and the covariances off the diagonal. These are the variances and covariances implied by the model's assumptions. We shall call this table the *theoretical variance covariance matrix* (see Appendix A.11). The matrix is symmetric, in the sense that the entry in row 1 and column 2 is the same as that in row 2 and column 1, and so on.

Let us now turn to the available data. Given observations on the variables Y_1, Y_2, and Y_3, we can calculate the observed variances and covariances of the variables, and arrange them in the *observed variance covariance matrix*:

	Variable:		
Variable:	Y_1	Y_2	Y_3
Y_1	S_1^2	S_{12}	S_{13}
Y_2	S_{21}	S_2^2	S_{23}
Y_3	S_{31}	S_{32}	S_3^2

Thus, S_1^2 is the observed variance of Y_1, S_{12} the observed covariance of Y_1 and Y_2, and so on. It is understood, of course, the $S_{12} = S_{21}$, $S_{13} = S_{31}$, and so on; the matrix, in other words, is symmetric. It can be easily confirmed that the observed variance covariance matrix for the data of Table 14.1 is as follows:

$$\begin{pmatrix} 9.84 & -0.36 & 0.44 \\ -0.36 & 5.04 & 3.84 \\ 0.44 & 3.84 & 3.04 \end{pmatrix}$$

On the one hand, therefore, we have the observed variances and covariances of the variables; on the other, the variances and covariances implied by the factor model. If the model's assumptions are true, we should be able to estimate the loadings β_{ij} so that the resulting estimates of the theoretical variances and covariances are close to the observed ones. We shall soon see how these estimates can be obtained, but first let us examine an important feature of the factor model.

14.3 FACTOR LOADINGS ARE NOT UNIQUE

We would like to demonstrate that the loadings are not unique, that is, that there exist an infinite number of sets of values of the β_{ij} yielding the same theoretical variances and covariances. Consider first an arbitrary set of loadings defining what we shall call Model A:

$$\begin{aligned} Y_1 &= 0.5\ F_1 + 0.5\ F_2 + e_1 \\ Y_2 &= 0.3\ F_1 + 0.3\ F_2 + e_2 \\ Y_3 &= 0.5\ F_1 - 0.5\ F_2 + e_3 \end{aligned}$$

It is not difficult to verify that the theoretical variance covariance matrix is

$$\begin{pmatrix} 0.5 + \sigma_1^2 & 0.3 & 0 \\ 0.3 & 0.18 + \sigma_2^2 & 0 \\ 0 & 0 & 0.5 + \sigma_3^2 \end{pmatrix}$$

For example, $Var(Y_1) = (0.5)^2 + (0.5)^2 + \sigma_1^2 = 0.5 + \sigma_1^2$; $Cov(Y_1, Y_2) = (0.5)(0.3) + (0.5)(0.3) = 0.3$; and so on.

Next, consider Model B, having a different set of β_{ij}:

$$\begin{aligned} Y_1 &= (\sqrt{2}/2)\ F_1 + 0\ F_2 + e_1 \\ Y_2 &= (0.3\sqrt{2})\ F_1 + 0\ F_2 + e_2 \\ Y_3 &= 0\ F_1 - (\sqrt{2}/2)\ F_2 + e_3 \end{aligned}$$

It can again be easily confirmed that the theoretical variances and covariances are identical to those of Model A. For example, $Var(Y_1) = (\sqrt{2}/2)^2 + (0)^2 + \sigma_1^2 = 0.5 + \sigma_1^2$; $Cov(Y_1, Y_2) = (\sqrt{2}/2)(0.3\sqrt{2}) + (0)(0) = 0.3$; and so on.

Examine now panel (a) of Figure 14.1. Along the horizontal axis we plot the coefficient of F_1, and along the vertical axis the coefficient of F_2 for each equation of Model A. The coefficients of F_1 and F_2 in the first equation are plotted as the point with coordinates (0.5, 0.5); those of the second equation as the point (0.3, 0.3), and those of the third as the point (0.5, −0.5).

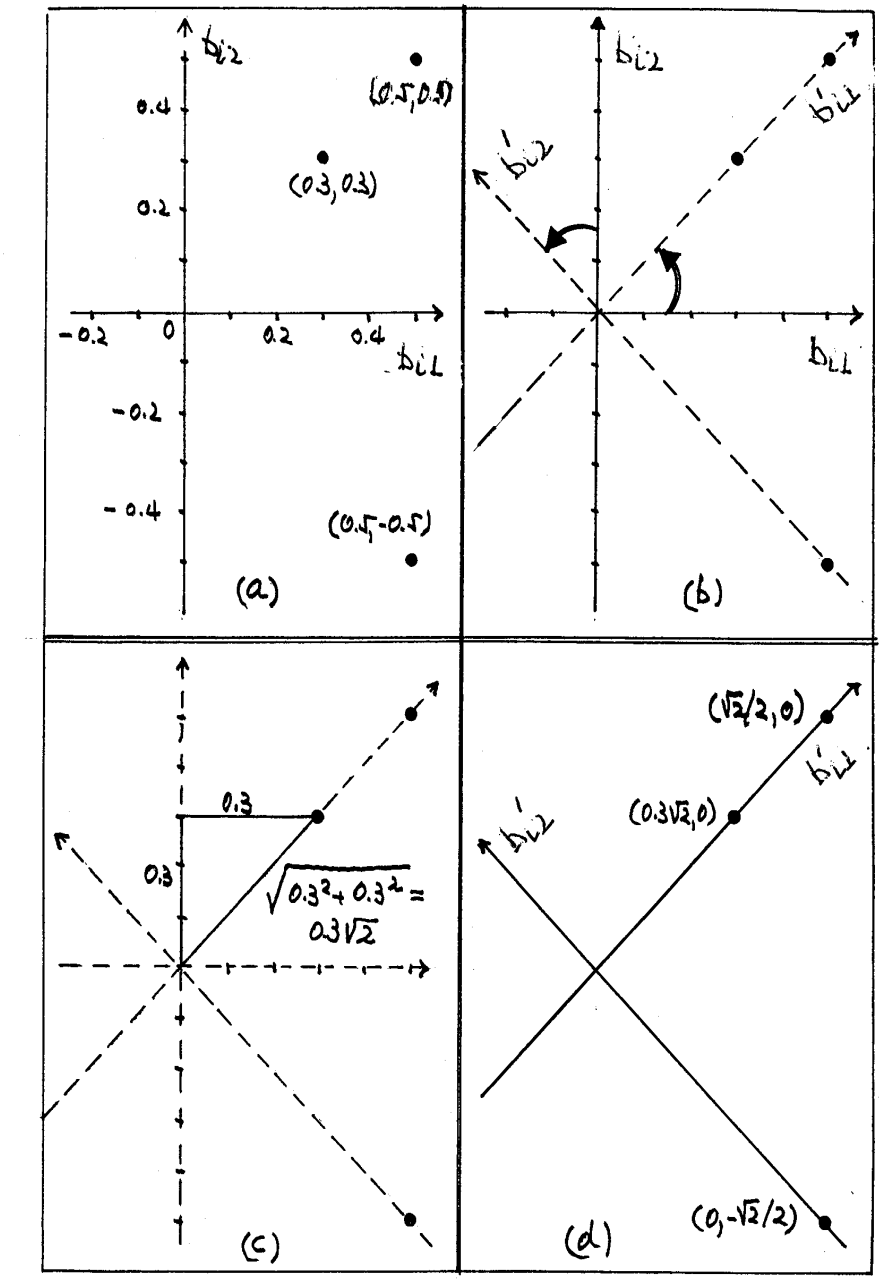

Figure 14.1
Rotation of loadings illustrated

Imagine rotating the coordinate axes anticlockwise by 45° as shown in Figure 14.1(b) to arrive at the new coordinate axes indicated by the dotted lines.

The coordinates of the three points with respect to the new axes can be calculated. For example, as illustrated in Figure 14.1(c), the point with original coordinates (0.5, 0.5) now has coordinates $(0.3\sqrt{2}, 0)$. The new coordinates for all three points are shown in Figure 14.1(d).

We see that the loadings of Model B are the result of applying to the loadings of Model A the rotation described above. It can be shown that any other rotation of the original loadings will produce a new set of loadings with the same theoretical variances and covariances as those of the original model. The number of such rotations is, of course, infinite large.

Rather than being a disadvantage, this property of the factor model is put to good use in practice. Recall that there are usually some prior expectations concerning the loadings. In particular, it is expected that some loadings will be close to zero, while others will be positive or negative and substantially different from zero. For this reason, factor analysis usually proceeds in two stages. In the first, one set of loadings β_{ij} is calculated which yields theoretical variances and covariances that fit the observed ones as closely as possible according to a certain criterion. These loadings, however, may not agree with the prior expectations, or may not lend themselves to a reasonable interpretation. Thus, in the second stage, the first loadings are "rotated" in an effort to arrive at another set that fit equally well the observed variances and covariances, but are more consistent with prior expectations or more easily interpreted.

Suppose, for example, that the loadings of Model A happened to be the estimates producing the best fit. All the loadings are of the same order of magnitude, and suggest that all variables depend on the same two factors. The loadings of Model B, however, fit the observed variances and covariances equally well and indicate clearly that Y_1 and Y_2 depend on one factor only, while Y_3 depends on the other factor only.

14.4 FIRST FACTOR SOLUTIONS

Perhaps the most widely used method for determining a first set of loadings is the *principal component method*.* This method seeks values of the loadings that bring the estimate of the total communality as close as possible to the total of the observed variances. The covariances are ignored. Table 14.2 shows (for the example of this chapter) the elements of the factor

* This is not the only method for factor analysis. Among others are the principal factor (also called principal axis) and maximum likelihood methods. See, for example, Johnson and Wichern (1992, Ch. 9), Rencher (1995, Ch. 13).

Table 14.2
Elements of principal component method

Variable, Y_i	Observed variance, S_i^2	Communality, $\beta_{i1}^2 + \beta_{i2}^2$
Finance grade, Y_1	S_1^2	$\beta_{11}^2 + \beta_{12}^2$
Marketing grade, Y_2	S_2^2	$\beta_{21}^2 + \beta_{22}^2$
Policy grade, Y_3	S_3^2	$\beta_{31}^2 + \beta_{32}^2$
Total	T_o	T_t

model on which the principal component method concentrates.

The communality, it will be recalled, is the part of the variance of the variable that is explained by the factors. The larger this part, the more successful the postulated factor model can be said to be in explaining the variable. The principal component method determines the values of the β_{ij} which make the total communality (T_t in Table 14.2) approximate as closely as possible the sum of the observed variances of the variables (T_o in Table 14.2).

All major statistical programs implement the principal component method. The solution for the example of this chapter (data of Table 14.1) is shown in Table 14.3.*

Table 14.3
Principal component solution, data of Table 14.1, unstandardized variables

Variable, Y_i (1)	Observed variance, S_i^2 (2)	Loadings on F_1, b_{i1} (3)	F_2, b_{i2} (4)	Communality, $b_{i1}^2 + b_{i2}^2$ (5)	Percent explained (6)=100×(5)/(2)
Finance, Y_1	9.84	3.136773	0.023799	9.8399	99.999
Marketing, Y_2	5.04	−0.132190	2.237858	5.0255	99.712
Policy, Y_3	3.04	0.127697	1.731884	3.0157	99.201
Overall	17.92	9.873125[a]	8.007997[a]	17.8811	99.783

[a] Sum of squared loadings

Estimated loadings are indicated by b_{ij}, to be distiguished from the theoretical loadings β_{ij}. The reported accuracy is excessive for practical

* Based on the output of program SAS with the statements `proc factor n = 2 cov vardef=n eigenvectors;` and additional calculations by hand.

purposes (two or three decimal places would have been sufficient); it is intended to assist comparisons with the output of other computer programs.

We see that the solution supports the expectations. The loadings on F_1 are relatively large for Y_1 but close to zero for Y_2 and Y_3; the loadings on F_2 are close to zero for Y_1 but relatively high for Y_2 and Y_3. Thus, F_1 could be interpreted as quantitative, and F_2 as verbal, ability. We also observe that the factor model explains nearly 100%, 99.7%, and 99.2% respectively of the observed variance of finance, marketing and policy grades. Overall, the two factors explain 99.78% of the sum of all observed variances.

The sum of squared loadings on F_1, $\sum_i b_{i1}^2$, can be interpreted as the contribution of F_1, and that on F_2, $\sum_i b_{i2}^2$, as the contribution of F_2 in explaining the sum of the observed variances. In Table 14.3,

$$\sum_i b_{i1}^2 = (3.136773)^2 + \cdots + (0.127697)^2 = 9.873125,$$

and

$$\sum_i b_{i2}^2 = (0.023799)^2 + \cdots + (1.731884)^2 = 8.007997.$$

Thus, F_1 explains about 9.873/19.92 or 55.1%, and F_2 about 44.7% of the sum of the observed variances.

The estimate of the specific variance of Y_i, σ_i^2, is the difference between the observed variance and estimated communality of Y_i:

Variable, Y_i	Observed variance	Estimated communality	Estimated spec. variance
Finance, Y_1	9.84	9.8399	0.0001
Marketing, Y_2	5.04	5.0255	0.0145
Policy, Y_3	3.04	3.0157	0.0243

Having the total communality approximate as closely as possible the sum of the observed variances (in effect, attaching the same weight to each variable) makes sense when the Y variables are measured in the same units, as in the example of this chapter. When this is not so, however, it is clear that the principal component method will favor the variables with large variances at the expense of those with small ones.

For this reason, it is customary to standardize the variables prior to subjecting them to the principal component method so that all have mean zero and variance equal to one. This is accomplished by subtracting from each observation (Y_{ij}) the mean of the variable ($\bar{Y}_i$) and dividing the result by the standard deviation (S_i) of the variable to obtain the standardized observation Y'_{ij},

$$Y'_{ij} = \frac{Y_{ij} - \bar{Y}_i}{S_i}.$$

It can be shown that the covariances of the standardized variables are equal to the correlation coefficients of the original variables (the variances of the standardized variables are, of course, equal to 1).

This last result can be easily verified using the data of Table 14.1. First, we calculate

	Y_1	Y_2	Y_3
Mean, $\bar{Y}_i$:	6.6	6.6	5.6
Variance, S_i^2:	9.84	5.04	3.04
Std. Dev., S_i:	3.1369	2.2450	1.7436

The observations of the standardized variables are shown in the following table:

Y_1'	Y_2'	Y_3'
−1.14763	−0.26726	−0.34412
0.12751	−1.60356	−1.49117
1.08387	1.06904	1.37646
−1.14763	1.06904	0.80294
1.08387	−0.26726	−0.34412

For example, the first standardized variable is given by

$$Y_{1j}' = \frac{Y_{1j} - 6.6}{3.1369}.$$

It can be confirmed that the means of the standardized variables are equal to 0, and their variances and standard deviations equal to 1.

The covariance of Y_1 and Y_2 is

$$S_{12} = \frac{1}{n}\sum Y_1 Y_2 - \bar{Y}_1\bar{Y}_2 = (216) - (6.6)(6.6) = -0.36.$$

The correlation coefficient of Y_1 and Y_2 is

$$r_{12} = \frac{S_{12}}{S_1 S_2} = \frac{-0.36}{(3.1369)(2.245)} = -0.0511,$$

and is equal to the covariance of Y_1' and Y_2',

$$S_{12}' = \frac{1}{n}\sum Y_1' Y_2' - \bar{Y}_1'\bar{Y}_2' = (-0.2556) - (0)(0) = -0.0511.$$

In general, $Cov(Y_i', Y_j') = Cor(Y_i, Y_j)$. The correlation matrix of the original variables (equal to the covariance matrix of the standardized variables) is

$$\begin{pmatrix} 1 & -0.0511 & 0.0804 \\ -0.0511 & 1 & 0.9810 \\ 0.0804 & 0.9810 & 1 \end{pmatrix}$$

The reader is asked to confirm the remaining numbers in Problem 14.3.

Standardization, in effect, subjects the *observed correlation matrix* of the original variables—rather than the observed variance covariance matrix—to the principal component method. The principal component solution for standardized variables will not necessarily be the same as that for unstandardized ones.

In some statistical programs (e.g., SPSS, SAS), standardization and the principal component method are the default options. Figure 14.2 shows the default output of `proc factor` of program SAS for the example of this chapter and the data in Table 14.1.*

The computer output is translated and interpreted in Table 14.4.

Table 14.4
Principal component solution, data of Table 14.1, standardized variables

Standardized variable, Y_i' (1)	Observed variance, $S_i'^2$ (2)	Loadings on F_1, b_{i1} (3)	F_2, b_{i2} (4)	Communality, $b_{i1}^2 + b_{i2}^2$ (5)	Percent explained (6)=100×(5)/(2)
Finance, Y_1'	1	0.02987	0.99951	0.99991	99.991
Marketing, Y_2'	1	0.99413	−0.08153	0.99494	99.494
Policy, Y_3'	1	0.99613	0.05139	0.99492	99.492
Overall	3	1.981463[a]	1.008306[a]	2.98977	99.659

[a] Sum of squared loadings

As with the original unstandardized variables, marketing and policy grades depend on one common factor (which can be interpreted as verbal ability) but not appreciably on the other (quantitative ability); the reverse holds for finance grades. Unlike the unstandardized case, however, verbal ability contributes more than quantitative ability: F_1 accounts for 1.981463/3 or about 66%, while F_2 accounts for 1.008306/3 or about 33.6% of the sum of the observed variances. This is why the verbal ability factor is listed first in the computer output and Table 14.4 (the labels F_1 and F_2 are, of course, arbitrary). The two factors together explain 2.98977/3 or 99.659% of the sum of the observed variances of the standardized variables, slightly less than with the original variables.

* Readers familiar with linear algebra may want to know that the principal component solution involves the eigenvalues (characteristic values) and eigenvectors (characteristic vectors) of the observed variance covariance or correlation matrix. Hence the appearance of these terms in the output of computer programs. For a clear mathematical exposition of the principal component method see, for example, Johnson and Wichern, *ibid.*

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 3 Average = 1

	1	2	3
Eigenvalue	1.9815	1.0083	0.0102
Difference	0.9732	0.9981	
Proportion	0.6605	0.3361	0.0034
Cumulative	0.6605	0.9966	1.0000

2 factors will be retained by the NFACTOR criterion.

Factor Pattern

	FACTOR1	FACTOR2
FIN	0.02987	0.99951
MKT	0.99413	-0.08153
POL	0.99613	0.05139

Variance explained by each factor

FACTOR1	FACTOR2
1.981463	1.008306

Final Communality Estimates: Total = 2.989769

FIN	MKT	POL
0.999910	0.994940	0.994920

Figure 14.2
SAS output, data of Table 14.1

14.5 FACTOR ROTATION

When the first factor solution does not reveal the hypothesized structure of the loadings, it is customary to apply rotation in an effort to find another set of loadings that fit the observations equally well but can be more easily interpreted. As it is impossible to examine all such rotations, computer programs carry out rotations satisfying certain criteria.

Perhaps the most widely used of these is the *varimax criterion*. It seeks the rotated loadings that maximize the variance of the squared loadings for each factor; the goal is to make some of these loadings as large as possible,

and the rest as small as possible in absolute value. The varimax method encourages the detection of factors each of which is related to few variables. It discourages the detection of factors influencing all variables.

The *quartimax criterion*, on the other hand, seeks to maximize the variance of the squared loadings for each variable, and tends to produce factors with high loadings for all variables.

Rotation Method: Varimax

Orthogonal Transformation Matrix

	1	2
1	0.99974	0.02264
2	-0.02264	0.99974

Rotated Factor Pattern

	FACTOR1	FACTOR2
FIN	0.00723	0.99993
MKT	0.99572	-0.05900
POL	0.99471	0.07393

Variance explained by each factor

FACTOR1	FACTOR2
1.980964	1.008805

Final Communality Estimates: Total = 2.989769

FIN	MKT	POL
0.999910	0.994940	0.994920

Figure 14.3
SAS output continued, data of Table 14.1

Figure 14.3 shows the output produced by the SAS program, instructed to apply the varimax rotation to the first set of loadings shown in Figure 14.3.

This output is translated and interpreted in Table 14.5.

The estimates of the communality of each variable and of the total communality are the same as in Table 14.4, but the contributions of each factor differ slightly. In this example, rotation did not alter appreciably the first estimates of the loadings or the proportions of the sum of the observed variances explained by the two factors.

Table 14.5
Varimax rotation, data of Table 14.1, standardized variables

Standardized variable, Y_i' (1)	Observed variance, $S_i'^2$ (2)	Loadings on F_1, b_{i1} (3)	Loadings on F_2, b_{i2} (4)	Communality, $b_{i1}^2 + b_{i2}^2$ (5)	Percent explained (6)=100×(5)/(2)
Finance, Y_1'	1	0.00723	0.99993	0.99991	99.991
Marketing, Y_2'	1	0.99572	−0.05900	0.99494	99.494
Policy, Y_3'	1	0.99471	0.07393	0.99492	99.492
Overall	3	1.980964[a]	1.008805[a]	2.98977	99.659

[a]Sum of squared loadings

14.6 ON THE NUMBER AND INTERPRETATION OF FACTORS

In the preceding illustration, the number of factors and their nature were hypothesized in advance. It was reasonable to assume that verbal and quantitative ability were two factors influencing course performance and grades. In other situations, however, the number of factors involved and their interpretation may not be clear.

Some computer programs, unless instructed otherwise, only identify ("extract") factors explaining a given proportion of the sum of the variances of the variables of interest. For example, when the variables are standardized a common default is to identify factors whose contribution is greater than 1.

It is common practice in factor analysis to examine the results of assuming that one, two, three, etc. factors are involved, and to tailor the hypothesis to fit the results of these analyses.

There is, however, some subjectivity in declaring loadings to be "high" or "close to zero" in absolute value. There could thus be disagreement among investigators as to whether or not the hypothesized structure of loadings is indeed supported by the data.

It should always be borne in mind that there are several methods for obtaining first and subsequent factor solutions, and each combination of first solution and rotation method may give rise to entirely different interpretations.

Example 14.1 The file `realest.dat` contains the prices and features of 100 residential real estate properties selected at random from among those sold in a large metropolitan area over a three-month period. Four of these

Table 14.7
Results of factor analysis, Example 14.1

Standardized variable, Y_i' (1)	Observed variance, $S_i'^2$ (2)	Loadings on F_1, b_{i1} (3)	Loadings on F_2, b_{i2} (4)	Communality, $b_{i1}^2 + b_{i2}^2$ (5)	Percent explained (6)=100×(5)/(2)
AREA	1	**0.90**	−0.01	0.82	82
LOTSZ	1	0.03	**0.99**	0.98	98
ROOMS	1	**0.87**	−0.06	0.76	76
BATHS	1	**0.78**	0.17	0.64	64
Overall	4	2.18[a]	1.01[a]	3.20	80

[a] Sum of squared loadings

features are:

AREA:	Floor area, in square feet
LOTSZ:	Size of the lot, in square feet
ROOMS:	Number of rooms in the house
BATHS:	Number of separate bathrooms in the house

The variables are described in more detail in the case *City West York* in Part II of this text. Table 14.6 lists the features of the first few properties in the file.

Table 14.6
Partial listing, `realest.dat` file

Property no.	AREA	LOTSZ	ROOMS	BATHS
1	740	1854	6	1
2	914	1256	7	2
3	968	1198	7	3
⋮	⋮	⋮	⋮	⋮

It would appear that all these features are functions of a single factor, the size of the property. After all, is it not true that large houses tend to be built on large lots, and to have large floor area, more rooms and more bathrooms?

The data outlined in Table 14.6 were processed using `proc factor` of the SAS program. By default, the program used standardized variables, the built-in criterion for determining the number of factors, and the principal component method; the varimax rotation was requested. The estimated rotated loadings and other statistics are shown in Table 14.7.

Contrary to prior expectations, two factors—not one—were identified by the program. The first factor has high loadings for AREA, ROOMS, and BATHS; it can be interpreted as the size of the house. The second factor has high loadings only for LOTSZ, and can be interpreted as the size of the lot. In the metropolitan area from which the data were selected, therefore, the size of the lot can be assumed to be unrelated to the size of the house proper. The size of the house explains 2.18/4 or 54.5%, and the size of the lot 1.01/4 or 25.45% of the sum of the variances of the standardized variables. The two factors together account for 80% of the total variance.

14.7 TO SUM UP

• Factor analysis is a method for investigating whether a number of variables of interest are linearly related to a smaller number of unobservable factors.

• In the special vocabulary of factor analysis, the parameters of these linear functions are referred to as *loadings*.

• Under certain conditions (A1 and A2 in the text), the theoretical variance of each variable and the covariance of each pair of variables can be expressed in terms of the loadings and the variance of the error terms.

• The *communality* of a variable is the part of its variance that is explained by the common factors. The *specific variance* is the part of the variance of the variable that is not accounted by the common factors.

• There exist an infinite number of sets of loadings yielding the same theoretical variances and covariances.

• Factor analysis usually proceeds in two stages. In the first, one set of loadings is calculated which yields theoretical variances and covariances that fit the observed ones as closely as possible according to a certain criterion. These loadings, however, may not agree with the prior expectations, or may not lend themselves to a reasonable interpretation. Thus, in the second stage, the first loadings are "rotated" in an effort to arrive at another set of loadings that fit equally well the observed variances and covariances, but are more consistent with prior expectations or more easily interpreted.

• A method widely used for determining a first set of loadings is the *principal component method*. This method seeks values of the loadings that bring the estimate of the total communality as close as possible to the total of the observed variances.

• When the variables are not measured in the same units, it is customary to *standardize* them prior to subjecting them to the principal component method so that all have mean equal to zero and variance equal to one.

• The *varimax* rotation method encourages the detection of factors each of which is related to few variables. It discourages the detection of factors influencing all variables.

• There is considerable subjectivity in determining the number of factors and the interpretation of these factors. There are several methods for obtaining first and rotated factor solutions, and each such solution may give rise to a different interpretation.

PROBLEMS

14.1 Confirm the results presented in Tables 14.4 and 14.5 using the data of Table 14.1 and a statistical program for factor analysis.

14.2 Confirm the results given in Table 14.7 using the data in the file `realest.dat` and a statistical program for factor analysis.

14.3 Using the data of Table 14.1 and in the manner of Section 14.4, confirm that the covariances of the standardized variables Y_1', Y_2', and Y_3', are equal to the correlation coefficients of the original variables Y_1, Y_2, and Y_3.

14.4 Two observable variables, Y_1 and Y_2, are thought to be linearly related to a common unobservable factor F:

$$Y_1 = \beta_{10} + \beta_{11}F + e_1$$
$$Y_2 = \beta_{20} + \beta_{21}F + e_2$$

Assume that F, e_1, and e_2 satisfy conditions A1 and A2 of Section 14.2.

(a) Derive the variances and covariance of Y_1 and Y_2. Arrange them in the form of a theoretical variance covariance matrix.

(b) Write the form of the observed variance covariance matrix.

(c) Suppose there are four observations on variables Y_1 and Y_2:

Y_1	Y_2
10	-5
-4	2
0	1
6	-3

Calculate the observed variance covariance matrix.

(d) Estimate the loadings β_{ij} according to the principal component method with the help of a statistical program.

(e) Is it possible to rotate the loadings in this case? Explain.

14.5 Refer to Model A of Section 14.3. Rotate the loadings by 90^o in an anticlockwise direction. Show that the variance covariance matrix of the variables is the same after rotation. What is the meaning of the rotation in this case?

14.6 Three observable variables are known to be related to two unobservable factors as follows:

$$Y_1 = -0.7F_1 + 0.6F_2 + e_1$$
$$Y_2 = 0.5F_1 + 0.4F_2 + e_2$$
$$Y_3 = 0.2F_1 - 0.3F_2 + e_3$$

(a) In the manner of Section 14.3, rotate the loadings by 45^o in a clockwise direction.

(b) Confirm that the variance covariance matrix of the variables is the same before and after rotation.

14.7 (a) "In order to apply factor analysis, one does not need the values of the observations for the variables of interest but only the variance covariance matrix or, when the variables are to be standardized, the correlation matrix of the variables." Comment.

(b) Some statistical programs seem to agree with the statement in (a) because they allow the user to input directly the observed variance covariance or correlation matrix for factor analysis. If possible, make use of this feature to analyze the following correlation matrix of four variables:

	Y_1	Y_1	Y_1	Y_1
Y_1	1.00	-0.01	0.81	0.02
Y_1	-0.01	1.00	0.03	-0.97
Y_1	0.81	0.03	1.00	-0.02
Y_1	0.02	-0.97	-0.02	1.00

Interpret the results.

14.8 (Due to Ms. Angela Griffiths) A simple random sample of 50 MBA students was selected from among those entering the program in a given year. The students' grades in 11 required courses were recorded. The required courses are:

FACTG	Financial Accounting for Managers
MACTG	Management Accounting
ECON	Economic Environment of Business
FINE	Managerial Finance
MKTG	Marketing Management
SKILLS	Management Skills Development
ENVIR	Business Environment
MIS	Management Information Systems
QM	Quantitative Methods
OPSM	Operations Management
OB	Organizational Behavior

The grades are expressed as integers from 9 (A+) to 1 (C-) and 0 (Fail). A partial listing of the data is given in Table 14.8.

Table 14.8
Data for Problem 14.8

ID No.	FACTG	MACTG	ECON	FIN	MKTG	SKILLS	ENVIR	MIS	QM	OPSM	OB
1	7	1	7	6	6	.	6	7	5	5	6
2	7	6	7	7	6	5	6	7	4	7	7
3	3	6	6	6	6	4	5	4	6	7	8
4	8	7	.	.	8	6	7	8	7	8	9
5	5	.	.	.	.	3	5	.	.	.	.
...	...	...	...	...	...	...	...	...	...	...	...
50	3	3	3	3	.	3	5	6	7	7	7

File `grades.dat`

```
Initial Factor Method: Principal Components

                Factor Pattern

                          FACTOR1

            FACTG          0.87942
            MACTG          0.62555
            ECON           0.27547
            FIN            0.58225
            MKTG           0.71344
            SKILLS         0.64239
            ENVIR          0.63944
            MIS            0.57534
            QM            -0.52988
            OPSM           0.78729
            OB             0.34549

        Variance explained by each factor

                        FACTOR1
                       4.261117


     Final Communality Estimates: Total = 4.261117

   FACTG      MACTG      ECON       FIN       MKTG     SKILLS
0.773382   0.391310   0.075881   0.339016   0.509002   0.412661

   ENVIR       MIS        QM       OPSM        OB
0.408879   0.331016   0.280773   0.619832   0.119366
```

Figure 14.4
Factor analysis results, Problem 14.8(a)

Many students had not completed all required courses; missing grades are indicated by a period.

(a) Figure 14.4 shows the output of a computer program for factor analysis directed to extract only one factor (program SAS with the statement `proc factor n=1`). Interpret and comment on the results.

(b) Can the analysis be improved? If so, carry out your suggestions using the file `grades.dat` and a program for factor analysis.

14.9 The file `bridge.dat` is described in Problem 4.12 and includes the following features of 45 bridges constructed by the Department of Transportation:

TIME:	Design time, in man-days
DAREA:	Deck area of bridge (000 sq.ft.)
CCOST:	Construction cost ($0000)
DWGS:	Number of structural drawings
LENGTH:	Length of bridge, in feet
SPANS:	Number of spans
DDIFF:	Degree of difficulty of bridge design (=0 easy, =1 complex)

A statistical program for factor analysis routinely processed the data according to its built-in defaults (standardization, principal component estimation, varimax rotation). It extracted two factors and produced the loadings shown in Table 14.9.

Table 14.9
Rotated factor loadings,
Problem 14.9

Variable	Factor 1	Factor 2
TIME	0.69732	0.47572
DAREA	0.74797	0.44545
CCOST	0.83123	0.35001
DWGS	0.59594	0.64808
LENGTH	0.93742	0.16039
SPANS	0.86564	0.20127
DDIFF	0.16549	0.93573

(a) Calculate the communality of each variable and the percentage of its variance that is explained by the factors. Calculate the percentage of the total variance that is explained by each factor and by both factors jointly.

(b) Interpret the results of factor analysis.

(c) Confirm the results using the data in the file `bridge.dat` and a program for factor analysis.

(d) Of what possible use is this type of analysis? Can it be improved? If so, carry out your recommendations using the data in the file `bridge.dat` and a program for factor analysis.

14.10 The file `mpg.dat` is described in Problem 5.15 and includes the following features of 116 car models:

ED:	Engine displacement (litres)
CYL:	Number of cylinders
HP:	Horsepower (bhp)
WEIGHT:	Weight (lbs.)
MPG:	Fuel mileage (miles per gallon)

The data on these variables were processed by a program for factor analysis according to its default features (standardization, principal component estimation and varimax rotation). The program extracted one factor but indicated it could not rotate the loadings shown in Table 14.10.

(a) Calculate the communality of each variable and the percentage of its variance that is explained by the factor. Calculate the percentage of the total variance explained by the factor.

(b) Interpret the results of factor analysis.

(c) Confirm the results using the data in the file `mpg.dat` and a program for factor analysis.

(d) Of what possible use is this type of analysis? Can it be improved? If so, carry out your recommendations using the data in the file `mpg.dat` and a program for factor analysis.

Table 14.10
Unrotated factor loadings,
Problem 14.10

Variable	Factor 1
ED	0.91337
CYL	0.90924
HP	0.83956
WEIGHT	0.83456
MPG	-0.92294

14.11 The file `stocks.dat`, described in Problem 10.14, contains the daily closing price of five stocks over a period of 378 consecutive trading days. A partial listing of the file can be found in Table 10.8.

The factors influencing the price of a stock are usually categorized as those that are common to all stocks (e.g., general economic conditions), those that are specific to the industry in which the firm operates (e.g., conditions in the lumber industry), and those that are specific to the firm itself (e.g., quality of its management).

Two of the five stocks in the file belong to one, and the remaining three to another industry.

Apply factor analysis to the data in the file `stocks.dat` to investigate if they are consistent with the above categorization. Explain carefully your results and any additional assumptions or special treatment you considered appropriate. Of what possible use is this type of analysis?

14.12 The file `mutfunds.dat` contains the share prices of 15 mutual funds at the end of each of 25 consecutive months. Also included in the file are the interest rate and the value of a stock market index at the end of each month. The file has the format shown in Table 14.11.

Table 14.11
Data for Problem 14.12

Month	MF1	MF2	···	MF15	IRATE	MINDEX
1	48.25	7.59	···	7.60	0.0833	3028.20
2	47.18	7.37	···	7.42	0.0835	2999.04
···	···	···	···	···	···	···
25	50.63	6.39	···	7.30	0.0985	3285.82

File `mutfunds.dat`

MF1 to MF15 are the share prices of the mutual funds, IRATE is the interest rate, and MINDEX the market index.

(a) The share price of a mutual fund is equal to the current value of its assets divided by the number of outstanding shares. If all funds carried similar portfolios of assets their share prices would vary in a similar fashion. To investigate the degree to which a single factor explains the observed variation in share prices, a factor model was estimated. The loadings obtained by the principal component method are shown in Table 14.12.

Table 14.12
Factor loadings, Problem 14.12

Variable	Factor 1
MF1	0.94156
MF2	0.85979
MF3	0.99148
MF4	0.98685
MF5	0.72733
MF6	0.76779
MF7	0.67069
MF8	0.97711
MF9	0.92559
MF10	0.95107
MF11	0.50267
MF12	0.90661
MF13	0.99303
MF14	0.98972
MF15	0.97688

Interpret the results.

(b) Are additional factors helpful in explaining the variation of mutual fund prices? Explain carefully the method used, the factors involved, and any assumptions or data transformations you considered appropriate.