2 Exploratory Common Factor Analysis

2.1. THE PARAMETERS OF THE COMMON FACTOR MODEL

Let us first bring together the central notions of Section 1.5.

In Section 1.5 we saw that the basic assumptions of the common factor model can be expressed in two distinct but equivalent ways. We can say that there exists a number of unobserved variables (common factors) that explain our observed correlations, in the sense that when these are partialled out, the partial correlations of our observed variables all become zero. Alternatively, we can say that each of our observed variables can be expressed as the sum of a (common) part that is its regression on a number of unobserved variables (common factors) and a residual about that regression and that the residuals are uncorrelated.

For the simple case where we suppose that the common factors are uncorrelated, the first of these descriptions is expressed mathematically in the statement

$$r_{jk} = f_{j1}f_{k1} + f_{j2}f_{k2} + \cdots + f_{jm}f_{km} \qquad \begin{matrix} j \neq k \\ j = 1, \ldots, n \\ k = 1, \ldots, n \end{matrix} \qquad (2.1.1)$$

where r_{jk} is the correlation (in the population) between the jth and kth variable, and f_{jp} is the regression weight of the jth variable on the pth factor (or its correlation with the pth factor, because these are the same when the factors are uncorrelated). The second of these descriptions is expressed in the statement that

$$y_j = f_{j1}x_1 + f_{j2}x_2 + \cdots + f_{jm}x_m + e_j \qquad j = 1, \ldots, n \qquad (2.1.2)$$

where y_j is the jth observed variable; x_p is the pth common factor, $p = 1, \ldots, m$; e_j is the residual of y_j about its regression on the factors (the unique factor); and f_{jp} is, again, the regression weight of y_j on x_p (the common factor loading—or coefficient—of variable j on factor p), together with the statement that the residuals are uncorrelated.

The statement in (2.1.1) is the testable numerical implication of the common factor hypothesis, as shown in Section 1.5, and is sometimes described as the *fundamental theorem of factor analysis*. The statement in (2.1.2) is the statistical common factor model itself.

The parameters of the model that we would want to estimate from a given sample are the nm factor loadings (v–f weights) f_{jp} and the n residual variances (variances of the residuals e_j), which are otherwise known as *unique variances* and which will be denoted by u_j^2. If our hypothesis specifies only the number m of common factors, this hypothesis is not sufficiently definite to *identify* the numbers in the $(n \times m)$ matrix of factor loadings (i.e., in the factor pattern). We say that these parameters are *not identified*. For example, in Tables 1.5.3 and 1.5.4, we had two sets of numbers, looking quite unlike each other so to speak, that give on computation by way of (2.1.1) exactly the same correlations. Given any one set of factor loadings that yield a set of correlations, the mathematician knows how to generate from these given numbers all the other sets of numbers that give the same correlations. These are *transformations* of the given numbers, and in factor analysis the process of transforming a given set of factor loadings into an alternative, equivalent set ("equivalent" in yielding the same correlations) is known as *rotation*. This nomenclature is correct for uncorrelated factors and incorrect for correlated factors. In what we call *exploratory factor analysis*, the user is unwilling (and just conceivably unable) to specify any more detail, so the mathematician arranges a two-step calculation for the user. In the first step, one set of fitted (estimated) factor loadings is calculated from the sample data supplied, perhaps a set that is mathematically convenient. This will give an *unrotated factor pattern* and is usually ignored by the user when it is supplied in the printed output. In the second step, the computer program performs arithmetic on the unrotated factor pattern, to obtain a "rotated" factor pattern in which the coefficients approximate simple structure (Section 2.4) while equally fitting the data.

In exploratory factor analysis, as we saw, the testable hypothesis is a hypothesis specifying the number of common factors. Using any rational best-fitting procedure for fitting the model to a sample, it seems obvious that the hypotheses that there are $1, 2, \ldots, n$ common factors form a sequence in which more factors must give a better fit. (The hypothesis that there are zero factors is the same as the hypothesis, Section 1.4, that the variables are mutually uncorrelated.) For a sufficiently large number of factors, the fit to any sample must be perfect, and the model would not constrain the data, and so be falsifiable. In

exploratory work, then, we follow Thurstone and regard the best number of common factors as the smallest number that will account for the correlations (i.e., the smallest number with which the data are consistent). The weaknesses of this approach are pointed out in Chapter 3.

We might expect that all the relevant information for the estimation of the parameters (the factor loadings and uniquenesses) of the common factor model from a sample of N subjects is contained in the sample correlation matrix and that the sample means and standard deviations are irrelevant (not to mention the scores of individual subjects). Broadly, this is true, though it is a very technical matter to prove it true. We shall describe two methods of obtaining "best" estimates: the method of maximum likelihood (ML) and the method of least squares (LS). The mathematician is able to show that ML estimates are, effectively, independent of the scale of the variables, so there is no important information in the sample standard deviations (or means). In the case of LS estimates, this is not actually true, but if we choose to use the sample correlations, we are minimizing the measure of fit (actually, of misfit), in the scale of (or, in technical language, *in the metric of*) the observed variables as standardized in the sample. This is rational behavior, so the reader is encouraged to remember from this paragraph the main point that it is technically all right to submit sample correlation matrices (rather than sample covariances) to an exploratory factor-analysis program.

2.2. ESTIMATION

Suppose now that we have drawn N subjects independently from a population, measured the n values of the variables under study on each subject, and computed the $(n \times n)$ matrix of sample correlations. It is necessary to distinguish three sets of numbers at this point and to adopt distinguishing notation for them. We shall write r_{jk} for the (unknown) correlation in the population and, correspondingly, f_{jp}, u_j^2 for the population factor loadings and unique variances. We shall write $\hat{r}_{jk}$, $\hat{f}_{jp}$, and $\hat{u}_j^2$ for our best estimates of these quantities obtained from our sample, and to avoid confusion we shall write a_{jk} for the sample correlation coefficient (i.e., the correlation between the measures of variable j and the measures of variable k obtained from our sample). (It would be nice to distinguish a fourth set of numbers—the *possible* values of the estimates out of which we are going to choose the best set—but we shall not do so.)

In least squares (LS) estimation we follow a widely used mathematical notion of "best" in fitting one set of quantities to another. We ask the mathematician to develop some computer arithmetic that will give us estimates $\hat{f}_{jp}$ and $\hat{u}_j^2$ that are better than any other numbers, in the sense that the quantity

$$Q = \sum_{j=1}^{n} \sum_{k=1}^{n} (a_{jk} - \hat{r}_{jk})^2 \tag{2.2.1}$$

where

$$\hat{r}_{jk} = \hat{f}_{j1}\hat{f}_{k1} + \hat{f}_{j2}\hat{f}_{k2} + \cdots + \hat{f}_{jm}\hat{f}_{km} \tag{2.2.2}$$

and

$$\hat{r}_{jj} = \hat{f}_{j1}^2 + \cdots + \hat{f}_{jm}^2 + \hat{u}_j^2$$

(and $\hat{r}_{jj} = 1$) is smaller than any other quantity we can get by choosing other values of f_{jp}. The expression (2.2.1) is a sum of squares and so must always be greater than or equal to zero. It could be zero only if every estimated correlation $\hat{r}_{jk}$ was exactly the same as the sample correlation a_{jk}. In that case the hypothesis would not be restrictive, and the fit would be perfect. With a small enough number of factors, we expect the fit to be imperfect in the sample, but we hope that the "residual correlations," the discrepancies between the sample a_{jk} and fitted $\hat{r}_{jk}$, will be small enough to allow the belief that the model is true of the population from which the sample was drawn. Table 2.2.1 gives a small example of estimates fitted by least squares.

The reader can try, with the aid of a calculator, varying the values of f_{jp} or u_j^2 slightly, in order to discover that any other values we choose will make the quantity Q larger and the fit worse. After the computer supplies us with a set of LS best-fitting numbers $\hat{f}_{jp}$, we can ask it to transform them (Section 2.4) to give a new set that will be equally well fitting (with the same value of Q and the same matrix of discrepancies $a_{jk} - \hat{r}_{jk}$) but will also approximate simple structure.

TABLE 2.2.1

Correlation Matrix = A

1.000	0.700	0.600	0.500
0.700	1.000	0.600	0.500
0.600	0.600	1.000	0.400
0.500	0.500	0.400	1.000

Factor Loadings (F)

f_1	-0.842
f_2	-0.842
f_3	-0.707
f_4	-0.587

FF'

0.709	0.709	0.595	0.494
0.709	0.709	0.595	0.494
0.595	0.595	0.499	0.415
0.494	0.494	0.415	0.345

Residual Matrix = $A - FF'$

0.291	-0.009	0.005	0.006
-0.009	0.291	0.005	0.006
0.005	0.005	0.501	-0.015
0.006	0.006	-0.015	0.655

Unique Variances: $u_1 = 0.291085$, $u_2 = 0.291085$, $u_3 = 0.500221$, $u_4 = 0.654269$

$Q = 0.00083152$ (Where Q = sum of squares of the off-diagonal elements of the residual matrix)

The method of least squares has the advantage that we develop a sound and simple intuition for the way in which "fit" is measured and made optimal. (It is actually better to speak of Q as measuring "misfit," or badness-of-fit, which we aim to minimize.) The method of least squares has the disadvantage, however, that with or without the stringent assumption of normality of the distribution of the population, it does not give us a test of significance of the hypothesis.[1]

The method of maximum likelihood is a very widely used method of estimation. Under normality assumptions, in many kinds of problem it gives the same mathematical expressions for estimates as the method of least squares but not in the cases considered in this book. The basic idea of the method is extremely simple, but mathematically its application here is too complicated to describe, even to the point of offering an expression like (2.2.1), without introducing some concepts from matrix algebra that are preferably omitted from an elementary book. The basic principle, quite simply, is that we ask the mathematician to give us arithmetic procedures for getting values of the population parameters that would make the probability of occurrence of our sample from that population as large as possible. (Pedants would insist that this statement assumes distributions that are not mathematically continuous. They would be right, but it doesn't matter.) When we calculate the probability of our sample as a function of various possible values of the population parameters, we refer to it as the *likelihood*, and we choose values to maximize the *likelihood* of the sample.

In applying the method of maximum likelihood to models for multivariate data, assuming normality, the mathematician finds it convenient to define a function of the likelihood that is also an algebraic measure of fit of the parameters (or, rather, of "misfit"). That is, it is positive and increases with an increase of the discrepancies between a_{jk} and $\hat{r}_{jk}$ and is zero only if the fit is perfect in the sample. This quantity is the natural logarithm of the ratio of the

[1]Instead of minimizing the *ordinary* least-squares function (2.2.1), we can minimize a weighted or generalized least-squares function

$$Q^* = \sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{m=1}^{n} w_{jk,lm}(a_{jk} - \hat{r}_{jk})(a_{lm} - \hat{r}_{lm})$$

with weights $w_{jk,lm}$ chosen to compensate for the variances and covariances of the residual covariances. A reasonable choice is

$$w_{jk,lm} = b_{jk}b_{lm}$$

where b_{jk} is the (j, k)th element of the inverse of the sample correlation matrix and approximates the expected value of the covariance of a_{jk} and a_{lm} in repeated sampling. Corresponding to (2.6.2), this generalized least-squares function can be written as

$$Q^* = \text{Tr}\ \{[\mathbf{A}^{-1}(\mathbf{A} - \mathbf{FF}' - \mathbf{U}^2)]^2\}$$

The advantage of the generalized least-squares function over the ordinary least-squares function is that it yields a chi-square test of fit in large samples, just as the likelihood function does.

likelihood of our data under the restrictive hypothesis to the likelihood of our data when we put no restrictions on the nature of the population from which it comes. It reaches a minimum when the likelihood is a maximum. Because of a very general result in the mathematical theory of statistics, we can choose to scale this measure of misfit so that its minimum is distributed as chi-square if the hypothesis of m factors is true, and if the sample size is large enough. We shall therefore assume that a computer program gives us a quantity that we shall denote by λ and just call the *likelihood ratio criterion* (LRC). (1) This quantity is a distribution-free measure of "misfit" of the estimated parameters to the sample correlations, which has been minimized by the computer program. Indeed, it is a direct measure of the departure of the correlation matrix of the residuals from an identity matrix and in this sense measures misfit of the model to the sample data. (2) It is the log-likelihood ratio for testing the hypothesis of m factors against the alternative "hypothesis" that the population is not constrained in any way. (3) It is distributed like chi-square, assuming normality, if the sample size is large enough, with degrees of freedom given by

$$\text{df} = \tfrac{1}{2}\{(n - m)^2 - (n + m)\}. \tag{2.2.3}$$

We can use ML estimates whether or not we assume normality. *If* we assume normality, we can also compare the LRC with tabulated chi-square for the given degrees of freedom. If the LRC does not exceed the chi-square value for, say, the 5% level, we have no reason to reject the postulated number of factors in favor of a larger number. If we reject the hypothesis, we can go on to test the fit with a larger number of factors.

Most social scientists have been nurtured in the classical Neyman–Pearson tradition for the testing of a statistical hypothesis. In this tradition we usually set up a restrictive hypothesis that we hope to reject in favor of an alternative, less restrictive hypothesis that is really our preferred outcome of the study. For example, we seek to reject a hypothesis that the means of a control group and an experimental group are equal in order to affirm that they are different and, the important outcome, that the treatment had a "significant" effect. In contrast, in factor analysis and in almost all models of any complexity for multivariate data, our interpretation of a notion of parsimony or, to put it more straightforwardly, the need to keep our account of the data as simple as possible gives us a desire to affirm the most restrictive hypothesis that is tenable. (It should preferably be substantively reasonable and interpretable as well as statistically tenable.) However, failure to reject a restrictive hypothesis usually means only that we do not have a large enough sample to reject it.

In exploratory factor analysis, the user may not have a hypothesis as to the number of factors. Indeed, to develop such a hypothesis genuinely and not just choose an arbitrary number, the user would need to classify the tests into sets (possibly overlapping) on substantive grounds and hypothesize a number of factors equal to the number of sets. That is, logically we cannot postulate how

many factors we have without postulating what they are. In that case our hypothesis is detailed enough to permit the immediate application of the confirmatory methods described in Chapter 3 and thus avoid exploratory analysis altogether.

Suppose, nevertheless, that the user insists that nothing is known about the tests that are to be factor analyzed and that the data are to be used to determine "how many factors to extract." (The notion of *extracting* factors is analogous to extracting the roots of a polynomial and has nothing to do with dentistry.) If we start with an arbitrary small number and fit m, $m + 1$, $m + 2$, . . . common factors until the chi-square is not significant at our favorite conventional level (5% or 1% presumably), we shall not have much idea of the probability to be associated with the entire nested sequence of statistical decisions. It is, however, known that the probability that we would thereby decide to fit more than the true number of factors is less than our chosen significance level.

We can be very sure that as we increase our sample size, the number of factors needed to reach a nonsignificant chi-square will increase. One might claim that all common factor hypotheses are false, because all restrictive statistical hypotheses are false, and they will be proved false by the use of a sufficiently large sample size. It seems not unreasonable to recommend that we use the chi-square test one-sidedly. It would be a worse error to retain and interpret factors that are "not real," that is, factors that are random error masquerading as genuine structure in the data, than to omit some not-very-detectable factors that are "real." More precisely, we should not retain m factors if $m - 1$ factors do not yield a significant chi-square, for we shall be pretending that random error is genuine structure. On the other hand, it would be rational to ignore a significant chi-square that seems to be requiring at least $m + 1$ factors, if the $(m + 1)$st factor were to supply little to the fit, or to the meaning of the analysis. That is, the chi-square test, combined with the efficiency of maximum likelihood estimation, serves primarily as a protection against overfactoring in relatively small samples, a tendency to which traditional approximate methods of factor analysis are prone (see Section 2.3).

From one point of view, inspection of the entire residual covariance matrix gives us more useful information about the fit of the model to the data than we obtain from the chi-square test.

Clearly, if the residual covariances of distinct variables are all sufficiently small, then m factors have accounted sufficiently well for the correlations of the variables. Accounting for correlations is the purpose of the model, and the smallness of the residual covariances is by definition the measure of its success in doing so. It should be noted, by the way, that some computer programs still in use for exploratory factor analysis do not print out any information about the residual covariances. Such programs cannot be recommended, as it is impossible to tell from them whether the analysis fits the data well, badly, or not at all. The trouble with direct inspection of residual covariances as a basis for determining whether or not the model fits the data well enough is of course lack of the

comforting sense of objectivity that comes from choosing a statistical significance level and consistently applying it. Acknowledging, with a deliberately mixed metaphor, that rules of thumb should be taken with a grain of salt, we might get a rough guide by combining the fact that a common factor seems to need at least three tests with loadings above .3 to define it adequately (see Section 2.3) with the elementary arithmetic result that $.3 \times .3 = .09$, to find a rule that if all residual covariances are less than .1, we are unlikely to be able to fit a further common factor that would be well defined and possibly interpretable. It is also possible to examine the largest residual covariances for evidence that they cluster, indicating the constitution of the additional factor that might be fitted if the chi-square is significant and some residual covariances are too large.

Technically, the mathematics and the computer arithmetic involved in minimizing either the function Q for LS estimates or the LRC for ML estimates is quite complex, and different procedures of varying efficiency have been recommended and programmed. The important fact remains, however, that programs do exist yielding LS and ML estimates, and in the latter case we also have a chi-square test of the hypothesis. The LS and ML solutions will usually be slightly different as they are based on different measures of fit. They can disagree widely in some cases, as when one gives a Heywood case while the other does not (see Section 2.3). But we seldom find a difference that matters.

Table 2.2.2 gives a sample correlation matrix obtained by selection from a study by Thurstone, with sample size $N = 213$. We know from previous work that the variables would be classified into three measures of verbal ability, V1, V2, and V3, say; three measures of word fluency, W1, W2, and W3; and three measures of reasoning ability, R1, R2, and R3, say. But we perform an exploratory factor analysis to see what it will tell us. The ML estimation procedure under the hypothesis of three common factors gives us an estimated factor pattern as shown in Table 2.2.3, "unrotated," meaning, not yet transformed to meet the

TABLE 2.2.2

Correlation Matrix

1.000	.828	.776	.439	.432	.447	.447	.541	.380
.828	1.000	.779	.493	.464	.489	.432	.537	.358
.776	.779	1.000	.460	.425	.443	.401	.534	.359
.439	.493	.460	1.000	.674	.590	.381	.350	.424
.432	.464	.425	.674	1.000	.541	.402	.367	.446
.447	.489	.443	.590	.541	1.000	.288	.320	.325
.447	.432	.401	.381	.402	.288	1.000	.555	.598
.541	.537	.534	.350	.367	.320	.555	1.000	.452
.380	.358	.359	.424	.446	.325	.598	.452	1.000

CODE: 1 = sentences, 2 = vocabulary, 3 = sentence completion, 4 = first letters, 5 = four-letter words, 6 = suffices, 7 = letter series, 8 = pedigrees, 9 = letter grouping

TABLE 2.2.3

Unrotated Factor Pattern (ML)			Communality	Uniqueness
.867	-.269	.021	.825	.175
.881	-.237	-.057	.835	.165
.826	-.222	-.031	.732	.268
.657	.445	-.320	.732	.268
.630	.429	-.219	.628	.372
.597	.237	-.290	.496	.504
.603	.320	.502	.718	.282
.646	.053	.291	.504	.496
.540	.381	.301	.527	.473

requirement of approximating to simple structure. The coefficients in this matrix (the factor loadings) are both the regression weights of the variables on the factors and their correlations with the factors. The value of the LRC is 2.916 on

$$\text{df} = \tfrac{1}{2}\{(9 - 3)^2 - (9 + 3)\} = 12$$

which from the table of chi-square has a probability of being exceeded that is equal to .995, so we do not reject the hypothesis of three factors. The matrix in Table 2.2.4 contains the residual covariance matrix, commonly abbreviated to *residual matrix*. As mentioned already, it should be printed out by a good factor-analysis program, for we can use it to see if the discrepancies between the model

TABLE 2.2.4
Residual Matrix

.175	.001	.001	-.005	.006	-.001	-.000	-.011	.007
.001	.165	.003	.001	-.002	.003	.006	-.003	-.010
.001	-.003	.268	.006	-.006	-.006	-.010	.022	.007
-.005	.001	.006	.268	-.001	.000	.004	-.004	-.004
.006	-.002	-.006	-.001	.372	.000	-.005	.002	.008
-.001	.003	-.006	-.000	.000	.504	-.002	.007	-.000
-.000	.006	-.010	.004	-.005	-.002	.282	.003	-.000
-.011	-.003	.022	-.004	.002	.007	.003	.496	-.004
.007	-.010	.007	-.004	.008	-.000	-.000	-.004	.437

and the data are small and evenly distributed or if there is an arrangement of the worst discrepancies that suggests an additional factor that we have otherwise failed to detect. For example, Table 2.2.5 shows a reanalysis of this sample with two factors hypothesized, and apart from the fact that the chi-square is significant, we can also see a definite bunching of the worst discrepancies in the residual matrix of Table 2.2.5(b) in the last block of variables. On the other hand, the residuals from three factors in Table 2.2.4 show, apart from the nonsignificant chi-square, that a fourth factor would certainly be ill-defined and unnecessary, because none is larger than .022. It is traditional in common factor analysis to present also the estimates of the *communalities* of the variables, the row sums of squares of the loadings, which are best thought of as the squared multiple correlation of each variable with all the common factors. These are also given in Table 2.2.3. In a modern analysis we tend to emphasize instead the unique variance (residual variance) of each variable. It is the proportion of the variance of each variable that is not explained by the factors. This is a piece of information that is complementary, equivalent information to the squared multiple correlations, and it lacks the ambiguities that the term *communality* has picked up over 40 years or so. For completeness, we also present the *rotated* factor pattern, using an algorithm called VARIMAX (see Section 2.4) in Table 2.2.6. The important point to note about this now is that the fit after "rotation" to a new set of factor loadings is just as good (or poor) as before "rotation." Table 2.2.7(a) gives the (unrotated) LS estimate of the common factor pattern, Table 2.2.7(b) gives the (varimax) rotated pattern and Table 2.2.7(c) gives the resulting residual matrix.

Because of the simplicity of the example chosen, the interpretive phase of our work is very simple. We declare, on the basis of the factor pattern in Table 2.2.6(a), that the first three variables have high correlations with (and are "heavily" weighted with) the first factor, the second three with the second factor, and the last three with the third factor. We then take this to mean that the factors are

TABLE 2.2.5

(a) Two-Factor Pattern (ML)

.883	.236
.891	.174
.838	.174
.640	-.515
.620	-.519
.594	-.329
.543	-.152
.622	.007
.497	-.270

(b) Two-Factor Residual

.166	.001	-.005	-.005	.007	.000	.003	-.010	.005
.001	.177	.002	.012	.002	.017	-.026	-.018	-.038
-.005	.002	.267	.013	-.004	.003	-.028	.011	-.011
-.005	.012	.013	.325	.010	.041	-.045	-.045	-.034
.007	.002	-.004	.010	.346	.002	-.014	-.015	-.003
.000	.017	.003	.041	.002	.539	-.085	-.047	-.059
.003	-.026	-.028	-.045	-.014	-.085	.682	.218	.287
-.010	-.018	.011	-.045	-.015	-.047	.218	.613	.144
.005	-.038	-.011	-.034	-.003	-.059	.287	.144	.679

Table 2.2.6
Varimax Factor Patterns

(a) Three-Factor (ML)			(b) Two-Factor (ML)	
.833	.243	.268	.868	.283
.827	.317	.226	.841	.339
.774	.283	.230	.798	.311
.228	.792	.230	.257	.780
.213	.706	.290	.237	.773
.315	.616	.135	.319	.599
.229	.180	.796	.373	.423
.444	.166	.528	.526	.333
.151	.312	.638	.270	.498

the three generic properties that, respectively, these groups of measures indicate in common, and presumably we name these generic properties *verbal ability*, *word fluency*, and *reasoning ability*. It is desirable to remark that in worthwhile research with factor analysis, one would hope to make a detailed examination of the measures used and to use relevant substantive theory to arrive at a deeper understanding of what might be operations, processes, or theoretical concepts requiring imagination to postulate, of which the measures are joint indicators.The example, we hope, is unrepresentatively mechanical.

It should be remarked that even in this rather dull example we have genuinely gained information. It was perfectly possible, a priori, that just one common factor would account for the correlations, or that two, say word fluency and reasoning, would do so with verbal ability a complex "resultant" of those two. It is sometimes suggested that "we only get out of a factor analysis what we put into it." This statement is never put quite precisely enough for one to come to grips with it, but at least we can say that it does not mean that the results of the analysis are entirely foreordained and uninformative. We can certainly get out things that we did not think that we had put in and, occasionally, not get things out that we felt confident we had put in.

The trouble with exploratory factor analysis, however, is that we often know far more than we are pretending to know and we fail to use this knowledge. A better analysis of the present example is given in Chapter 3, where we specify in our hypothesis not only the number of factors but which variables have zero regression weights on which factors. Thereby we create an unambiguous and satisfying hypothesis and test the hypothesis precisely as it stands.

TABLE 2.2.7

(a)

.821	-.379	-.025
.838	-.350	-.104
.789	-.334	-.66
.705	.372	-.307
.679	.358	-.205
.618	.175	-.288
.650	.197	.511
.654	-.061	.272
.593	.279	.308

Three-Factor (LS) Pattern Unrotated

(b)

.828	.249	.264
.828	.320	.217
.779	.281	.229
.233	.796	.207
.213	.715	.273
.317	.616	.120
.255	.199	.796
.447	.179	.523
.152	.330	.627

Three-Factor (LS) Pattern Rotated (varimax)

(c)

.182	.005	.001	-.007	.005	-.001	.001	-.012	.007
.005	.165	-.006	.001	-.001	.003	.009	-.004	-.009
.001	-.006	.262	.008	-.005	-.005	-.012	.016	.005
-.007	.001	.008	.270	-.001	.001	.006	-.005	-.004
.005	-.001	-.005	-.001	.369	-.000	-.006	.001	.006
-.001	.003	-.005	.001	-.000	.505	-.001	.005	-.002
.001	.009	-.012	.006	-.006	-.001	.277	.003	-.000
-.012	-.004	.016	-.005	.001	.005	.003	.495	-.003
.007	-.009	.005	-.004	.006	-.002	-.000	-.003	.475

Residual Matrix

2.3. COMPONENT THEORY, IMAGE THEORY, APPROXIMATE METHODS, AND HEYWOOD CASES

There are two "theories" of multivariate data, quite logically distinct from common factor analysis, which tend to give enough numerical and conceptual similarities to it to make them be seen as competitive alternatives to it or sometimes to cause them to be confused with it and which certainly make them useful approximations to it. Some readers, on the basis of other knowledge, will in fact object to the treatment here of these two topics in one brief section, subordinated to exploratory factor analysis. Let it be emphasized that this is because it suits the overall plan of this book to do so. Principal component theory, sometimes under the guise of *optimal scaling* or *optimal weighting*, has a considerable body of literature in its own right. It is preferred by some investigators to common factor analysis. *Factor analysis* as a generic term is generally taken to include component theory and image theory. We shall continue, therefore, to use the word *common* before *factor analysis* in its narrow sense and accept, but not deliberately follow, the general usage of factor analysis as a broader, looser term.

(a) Principal Component Theory

Conceptually, the best way to understand principal components is rather different from the way favored by the mathematician. Suppose we have a set of n observed variables $y_1, \ldots, y_n$, and we make a weighted sum of them, say,

$$s = w_1 y_1 + w_2 y_2 + \cdots + w_n y_n \tag{2.3.1}$$

just as we might in regression theory, where we want to choose the weights in some "best" way. But unlike the regression case we do not have an external criterion with which to correlate the weighted "mixture." Instead, we want to consider just internal relationships. We want a simple combination of all the measures that "resembles" each individual measure as much as possible. Now one way to make this rather vague notion mathematically definite is to say that if we calculated the square of the correlation of s with each of the variables, $y_1, \ldots, y_n$, then s would on the whole resemble all of them most when we choose weights so that the sum of the squares of the n correlations of s with y_1, s with y_2, $\ldots$, s with y_n is as large as possible. Given the $(n \times n)$ matrix of correlations of $y_1, \ldots, y_n$, there is a definite mathematical answer to this question, though its arithmetic is rather unpleasant and requires a computer program. For example, Table 2.3.1 gives the correlation matrix of a set of variables and shows the weights to give them that will yield a weighted sum whose total of squared correlations is the largest possible. No other choice of weights will increase the value beyond that shown. (The effects of making small arbitrary changes in the weights are demonstrated in the table.) We note that we could, of course, multiply all the weights by a common constant without changing the sum of

TABLE 2.3.1
From Hotelling (1933)

					Weights	Correlations
y_1 = Reading Speed	1.000	.701	.266	.084	.602	.618
y_2 = Reading Power	.701	1.000	-.059	.092	.512	.695
y_3 = Arithmetic Speed	.266	-.059	1.000	.596	.448	.608
y_4 = Arithmetic Power	.084	.092	.596	1.000	.425	.578

Sum of squares of correlations = 1.846

Full set of weights

.602	-.362	-.404	.587
.512	-.512	.399	-.560
.448	.557	-.521	-.472
.425	.545	.636	.350

If we take weights all equal to .5, we get a sum of squares of correlations = 1.840.

squared correlations. What matters is the proportions in the mixture, as in the regression of a dependent variable on independent variables. Now suppose that we record, but set aside, the "best" combination that we have found and look for a "second-best" combination. To avoid confusion, we label the best combination s_1 and write

$$s_1 = w_{11} y_1 + w_{12} y_2 + \cdots + w_{1n} y_n \tag{2.3.2}$$

adding a subscript unity to the first set of weights we found. Now we look for a "second-best" combination

$$s_2 = w_{21} y_1 + w_{22} y_2 + \cdots + w_{2n} y_n \tag{2.3.2}$$

To avoid just finding s_1 again, we ask that s_2 be uncorrelated with s_1 and that subject to this condition, it should have a maximum sum of squares of its n correlations with $y_1, \ldots, y_n$. Again we obtain a set of weights $w_{21}, \ldots, w_{2n}$ that provide the weighted sum we require. We can now ask for a third-best weighted sum, with maximum squared correlations with $y_1, \ldots, y_n$ and uncorrelated with s_1 and with s_2. This process is continued. We can find n weighted sums, $s_1, \ldots, s_n$, each of which is uncorrelated with all the other sums, and each in turn has the largest sum of squares of correlations with the n variables that it can have. We shall call these sums *principal component* scores.

Now let p_{jl} be the correlation between the jth variable and the lth sum, s_l. We find that there is a converse relationship between the variables and the principal

component scores. We already have each principal component score as a weighted sum of variables

$$\begin{aligned} s_1 &= w_{11}y_1 + w_{12}y_2 + \cdots + w_{1n}y_n \\ s_2 &= w_{21}y_1 + w_{22}y_2 + \cdots + w_{2n}y_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ s_n &= w_{n1}y_1 + w_{n2}y_2 + \cdots + w_{nn}y_n \,. \end{aligned} \tag{2.3.4}$$

For example, from Table 2.3.1,

$$\begin{aligned} s_1 &= .602y_1 + .512y_2 + 448y_3 + .425y_4 \\ s_2 &= -.362y_1 - .512y_2 + .557y_3 + .545y_4 \\ s_3 &= -.404y_1 + .399y_2 - .521y_3 + .636y_4 \\ s_4 &= .587y_1 - .560y_2 - .472y_3 + .350y_4 \end{aligned}$$

It turns out that we can interchange roles and write the variables as weighted sums of the n components, with the correlations p_{jl} as the weights; that is, we have

$$\begin{aligned} y_1 &= p_{11}s_1 + p_{12}s_2 + \cdots + p_{1n}s_n \\ y_2 &= p_{21}s_1 + p_{22}s_2 + \cdots + p_{2n}s_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n &= p_{n1}s_1 + p_{n2}s_2 + \cdots + p_{nn}s_n \,. \end{aligned} \tag{2.3.5}$$

For example, from Table 2.3.1,

$$\begin{aligned} y_1 &= .818s_1 - .438s_2 - .292s_3 + .240s_4 \\ y_2 &= .695s_1 - .620s_2 + .288s_3 - .229s_4 \\ y_3 &= .608s_1 + .674s_2 - .376s_3 - .193s_4 \\ y_4 &= .578s_1 + .660s_2 + .459s_3 + .143s_4 \end{aligned}$$

We began with a regression of a sum of variables on each of those variables, and we have thence obtained a regression of each of the observed variables on those sums. Further, because the component scores are uncorrelated, the expression of each observed variable as a sum of components gives an analysis of its variance into n additive parts, one due to each component. That is, if y_j is in standard measure, then its unit variance is given by

$$\sigma_j^2 = 1 = p_{j1}^2 + p_{j2}^2 + \cdots p_{jn}^2 \tag{2.3.6}$$

for example

$$\sigma_1^2 = .818^2 + (-.438)^2 + (-.292)^2 + .240^2 = 1.00$$

As we saw initially, each component in turn explains the maximum possible proportion of the variance of all n of the variables; for example, the first component explains

$$.818^2 + .695^2 + .608^2 + .578^2 \text{ units of variance.}$$

If we wanted to substitute just one combined measurement for our n measurements $y_1, \ldots, y_n$, we could not do better than to use the first principal component score, s_1, which is maximally correlated with all of them and explains more of their variance than any other composite measurement could. If we wanted to keep some m measures, less than all n of them, we could not do better than to keep the first, second, . . . , mth principal component scores, ordered in terms of the magnitude of the sum of variance explained. We might call principal components "best approximate descriptions" of multivariate data.

As in the common factor model, we find that the correlation between any two variables can be written as the sum of the products of the correlations of the two variables with all n of the components. That is,

$$r_{jk} = p_{j1}p_{k1} + p_{j2}p_{k2} + \cdots + p_{jn}p_{kn} \tag{2.3.7}$$

For example,

$$\begin{aligned} r_{12} = {} & .818 \times .695 + .438 \times .620 - \\ & .292 \times .288 - .240 \times .229 = .701 \end{aligned}$$

Except in the special case where there are redundant variables in the set (i.e., where some variables in the set can be perfectly predicted from the rest, usually because we have included sums of part scores along with their parts in the set), we require all n of the components to explain the correlations by (2.3.7). This is in contrast to the common factor model, where we usually have m factors, where m is much less than n, explaining the correlations (but not the variances) of the variables.

It is reasonable to hope that "a few" of the principal components will explain a large part of the variance of the given variables. However, the point deserves emphasis that we cannot in general find correlation matrices of n variables that can be entirely explained by less than n components, either in respect to the variance of the variables or in respect of their correlations only. Hence, principal component theory does not yield a falsifiable hypothesis. Typically, in social science work, the output of a principal component analysis is presented as a matrix of the correlations p_{jl} between the variables and the components, usually omitting the columns of correlations that are supposed by the investigator to be "negligible" in some sense. We shall refer to these correlations as *principal*

component coefficients. (Sometimes they are known as principal component loadings, following usage in common factor analysis.) The sum of squares of the elements in each column of this matrix is the variance of all the variables that is explained by that principal component, in terms of the analysis of the variance of the variables into uncorrelated parts.

Occasionally we may feel that we can interpret the component score as a weighted sum of the given variables on the basis of the relative magnitudes and signs of the parts of the "mixture." Tables 2.3.1 and 2.3.2 give an example from Hotelling (1933). His interpretation, which is plausible enough, is that

> the chief component seems to measure general ability; the second, a difference between arithmetic and verbal ability. These two account for eightythree percent of the variance (of the four variables). An additional thirteen percent seems to be largely a matter of speed vs. deliberation. The remaining variance is trivial.

It should be clear from this example that principal component theory resembles common factor theory but with important differences. Its output gives us correlations between observed variables and components. We interpret those components, if we can, in terms of what is measured by the variables that are correlated with each component. However, the principal components are themselves known weighted sums of the given variables, chosen to explain variance in terms of multiple correlation principles; whereas common factors are unknown variables, chosen to explain correlations in terms of partial correlation principles.

To take another example, from Thomson (1934), the constructed correlation matrix in Table 2.3.3 is precisely fitted by the common factor model with one common factor. Its five principal components have the coefficients indicated and successively explain 2.683, 0.890, 0.652, 0.448, 0.328 units of variance (these sum to five as they must). The one common factor explains the correlations perfectly, although not even four of the five components explain the correlations perfectly. On the other hand, the first principal component alone explains more variance than the one common factor.

TABLE 2.3.2
Hotelling (1933)
Principal Component Coefficients

	1st Comp.	2nd Comp.	3rd Comp.	4th Comp.
Reading Speed	.818	-.438	-.292	.240
Reading Power	.695	-.620	.288	-.229
Arithmetic Speed	.608	.674	-.376	-.193
Arithmetic Power	.578	.660	.459	.143
Sum of Squared Correlations	1.846	1.465	.521	.167

TABLE 2.3.3
Thomson (1934)

(a) The Correlation Matrix

$$\begin{bmatrix} 1.000 & .669 & .592 & .458 & .251 \\ .669 & 1.000 & .566 & .438 & .240 \\ .592 & .566 & 1.000 & .387 & .212 \\ .458 & .438 & .387 & 1.000 & .164 \\ .251 & .240 & .212 & .164 & 1.000 \end{bmatrix}$$

This matrix is explained perfectly by one factor with loadings

$$[.837 \quad .800 \quad .707 \quad .548 \quad .303]$$

(b) Principal Component Coefficients

$$\begin{bmatrix} .856 & -.092 & -.152 & -.217 & -.436 \\ .840 & -.098 & -.173 & -.346 & .365 \\ .790 & -.116 & -.294 & .523 & .060 \\ .673 & -.182 & .713 & .083 & .020 \\ .413 & .908 & .069 & .022 & .009 \end{bmatrix}$$

(c) Residuals from First Principal Component

$$\begin{bmatrix} .267 & -.050 & -.084 & -.118 & -.103 \\ -.050 & .294 & -.098 & -.127 & -.107 \\ -.084 & .098 & .376 & -.147 & -.114 \\ -.118 & -.127 & -.147 & .547 & -.114 \\ -.103 & -.107 & -.114 & -.114 & .829 \end{bmatrix}$$

A word is necessary about the relation between the account of principal component theory just given and the one that the reader is most likely to encounter elsewhere. Hotelling (1933) introduced principal component theory in his first paper, both as above and in a different fashion. The idea of finding a weighted sum that resembles *all* our variables most, by having maximum (squared) correlations with all of them, is conceptually a good way to think about principal components, but it leads to rather difficult mathematics. The most commonly preferred way to introduce principal components is to say that we are looking for a set of weights to give a score with maximum variance, subject to the condition that the sum of the squares of the weights be held constant, thus varying the proportions in the "mixture" but not the amount. Newcomers to multivariate analysis sometimes follow the mathematics of this notion, which are much easier, and yet do not understand the notion itself (i.e., *why* we want such sums). We shall simply accept the fact that the two problems have the same mathematical answers (when the variables are standardized). Further, the mathematical answers happen to be the same as certain considerably older problems in astronomy and geometry. The geometrical problem is that of finding the principal axes of ellipsoids; hence we sometimes find *principal axes* (perhaps loosely) used as a synonym for principal components. Also, because of prior

usage in writings on the fundamental problem of finding equivalents of our maximized sum of squared correlations, which is also the total variance explained by each component when the variables are in standard measure, these quantities are sometimes published as the *eigenvalues, latent roots,* or *characteristic roots* of the correlation matrix, and the corresponding sets of principal component coefficients are sometimes labeled *eigenvectors, latent vectors,* or *characteristic vectors.*

It is quite usual to find matrices of principal component coefficients presented in the literature as substitutes for common factor coefficients. It is also usual to find, under titles like "principal axes factor analysis" or "principal components, iterated once," modified principal component analyses that have made one or two arithmetic steps of unstated nature toward obtaining least squares estimates in the common factor model. These are legacies of the era from the 1930s to the 1950s when problems of estimation were not well understood. Such results in the literature are hard to evaluate. More will be said about them later.

It will be noticed that no distinction has been made so far in this section between principal components in a population and principal components in a sample. In fact, the ambiguity, which was deliberate, leaves us free to read all these remarks in terms of either, especially as there are no restrictive hypotheses to test.

(b) Image Theory

The mathematics of image theory is a simple application of regression theory. It is interesting, partly, because of a particular conception about the way in which we choose, or we should choose, which measurements to make. Generally, we know or think we know well enough how to define a population of subjects that is of interest to us and to choose subjects from it. Besides choosing our subjects, we also decide how many things and just what things to measure on them. Suppose we pretend that this decision is, like the matter of choosing subjects, one of selection from a population. In imagination, we suppose that given time we could have listed all the distinct measurable properties, or behaviors in various situations, of our subjects that can be conceived. We regard the things we choose to measure as a subset of all the distinct choosable measurements, imagined or as yet unimagined, that could ever be made on our subjects. It is not obvious for a given class of subjects whether or not this list is infinitely long. Now suppose that instead of such an *entire* list, we are imagining a list of attributes of a given, more or less definable, kind (e.g., cognitive attributes, emotional attributes, attitudes; or at a more detailed level of description, arithmetic performances or vocabulary knowledge). To the extent that we have a definite denotation of the "kind" so that we can recognize if a measurement is that kind of measurement or not, we can imagine using all distinct measures of the kind in question. Such an entire set of conceivable measures has been discussed at times under the name of a *behavior domain* or of a *universe of content.* It might be claimed that the object of factor-analytic methods is to discover what measures belong to what behavior domains. On the other hand, it can be claimed that the investigator has a duty at the beginning of a study to be as clear as possible about the definition of the behavior domain that he is about to investigate. In practice we are likely to have notions about the behavior domain whose degree of precision varies from vague to precise, depending on how little or how much we know already. It seems easy to mark off all the items requiring a subject to add numbers together and give the sum from all the items that require something else or something more. We could say, therefore, that numerical addition is a well-defined behavior domain. On the other hand we might not be able to get all psychologists and all psychiatrists to pick out just the same items as measures of "anxiety." (Do you have bad dreams? Do you perspire a lot? Do you think the world is fundamentally an evil place?) Presumably a sensitive interplay of clinical theory, measurement, and multivariate analysis should lead us from a vague conception of the behavior domain of "anxiety" to an increasingly precise denotation of it. But generally, although we can try to be precise in our conception of the behavior we wish to study, we have to be willing to work with conceptions ranging all the way from vague intuitions to precise denotations that serve as instructions for inventing all the possible measures that belong to the domain.

It is important to note that the behavior domain (the "kind" of property) we investigate need not be thought of as conceptually simple but may in fact be subdivisible, or cross-classifiable, into a number of more elementary attributes. For example, addition items form a behavior domain, but they can be further classified in terms of (a) the number of terms in the sum, (b) the number of digits in the terms, and so on. The attempt to define complex behavior domains as logical combinations of elementary attributes has been described under the title *facet theory.* In some areas of inquiry, we can indeed use logic, or substantive knowledge, to describe a behavior domain as a combination of distinct attributes or *facets.* It is then a matter of fact whether the logical analysis we produce of the kind of thing we measure will serve to predict the statistics of the measures in some population. The basic expectation of psychometricians seems to be that the more two measures resemble each other in the nature of the properties measured, the higher should be the correlation between them in any or all or most populations. This is not a logical necessity, of course. It is a postulated empirical law that is based on the "common-elements" explanation of the "why" of correlation. Why are two variables correlated? Because they (in part) measure the same thing. As we change the properties of a given item to derive less and less similar items, we usually expect their correlations with the given item to decrease. Out of a strong a priori conceptual analysis of measures, we should be able to group them into kinds or perhaps arrange them in an ordered sequence in respect of one or more attributes. For example, we could group all the one-digit addition items

together, all the two-digit addition items together, etc. Alternatively, we could *order* the items in terms of number of digits. We would then hope to find, in the correlational behavior of the measures, confirmation of our conceptual analysis.

The notion that behavior domains "exist" and that we should try to make our measures "represent" them is an interesting way to describe the notion that we should come to know what properties we are measuring and to know which alternative measures will serve as indicators of those properties.

The basic mathematical idea of image theory is independent of the conception of behavior domains. Given any set of n variables, $y_1, \ldots, y_n$, in standard measure, we can obtain the regression estimate of each variable in turn on the remaining $n - 1$ variables. We write

$$\hat{y}_j = b_1 y_1 + b_2 y_2 + \cdots + b_{j-1} y_{j-1} + b_{j+1} y_{j+1} + \cdots + b_n y_n. \tag{2.3.8}$$

(Note the way in which we indicate the omission of y_j itself from the expression on the right.) The regression weights and the multiple correlations can be calculated by the standard arithmetic procedures that were taken for granted in Chapter 1. No new arithmetic is needed. Tables 2.3.4 and 2.3.5 give the regression weights and squared multiple correlations for the Spearman case of Table 1.5.1, and the case previously factor analyzed in Tables 2.2.2 to 2.2.7.

We now think of each variable as the sum of two parts, its regression upon the remainder, $\hat{y}_j$, and its residual about that regression, e_j, say. Guttman calls the regression part $\hat{y}_j$ the *partial image* of y_j and the residual e_j the *partial antiimage* of y_j. Now suppose that there are infinitely many measures in the same behavior domain as our given measures $y_1, \ldots, y_n$ (i.e., infinitely many distinct measurable properties of the same kind). Then we define the *total image* of each y_j as its regression on all the remaining measures in the same behavior domain and its *total antiimage* as the residual about that regression.

TABLE 2.3.4
Image Analysis of
Spearman Matrix (Table 1.5.1)
Regression of Each Variable on Remainder

Variable	Squared Multiple Correlation	ω_1	ω_2	ω_3	ω_4	ω_5
1	.653	.0	.431	.266	.182	.129
2	.550	.532	.0	.154	.105	.075
3	.428	.418	.196	.0	.083	.059
4	.317	.341	.160	.099	.0	.048
5	.221	.276	.130	.080	.055	.0

TABLE 2.3.5
Image Analysis of
Thurstone Matrix (Table 2.2.2)
Regression of Each Variable on Remainder

Variable	Squared Multiple Correlation	Regression Weights ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
1	.738	.0	.529	.304	-.038	.016	.027	.051	.052	.027
2	.750	.506	.0	.286	.067	.028	.069	.028	.059	-.043
3	.675	.378	.372	.0	.065	-.003	.024	-.030	.110	.017
4	.553	-.065	.120	.089	.0	.421	.261	.039	-.027	.090
5	.520	.030	.054	-.004	.452	.0	.172	.059	.022	.124
6	.426	.058	.157	.042	.335	.206	.0	-.052	.006	.026
7	.477	.102	.058	-.048	.046	.064	-.048	.0	.293	.390
8	.450	.110	.130	.185	-.034	.025	.006	.308	.0	.114
9	.429	.059	-.009	.029	.115	.148	.026	.426	.118	.0

Unlike the residuals in common factor analysis, the partial antiimages are not mutually uncorrelated, so the partial images of the variables do not explain their correlations (in the partial-correlation sense of "explain"). Hence the partial images cannot serve as "common parts" in the factor-analytic sense. Yet there is an obvious sense in which we might feel that image theory should, like common factor theory, be about "what variables have in common," because it is, by definition, about "what each variable has in common with the remaining $n - 1$ variables." It is possible to show that the partial image variance of each variable (the squared multiple correlation of each variable with the remaining $n - 1$) is always less than its communality (squared multiple correlation with the common factors). This also means that its partial antiimage variance (residual variance about its regression on the remaining $n - 1$ variables) is always greater than its uniqueness (residual variance about its regression on the common factors). However, if the common factor model with its limited number of factors correctly describes the entire behavior domain and if the behavior domain contains infinitely many distinct measures, then the total image of each variable is the same as its common part, the two residual parts are the same, the total image variance is equal to the communality, the total antiimage variance is equal to the uniqueness, and generally the two theories coincide completely.

It is important to note the requirement, rather loosely stated, that the common factor model must correctly describe the entire (infinite) behavior domain. If, as we add more variables to our list, we add more factors, so that by the time we "have an infinity" of variables, we "have an infinity" of common factors, then we do not find total images and common parts coinciding in the limit. More precisely, the ratio of the number of factors to the number of variables should go to zero in the limit. We can and should contemplate the possibility that the common factor model is a quite inappropriate model for a given set of measures. Evidence for this is obtained if we fit, successively, m, $m + 1$, . . . factors to a sample correlation matrix and the fit successively improves but not nearly fast enough, and we have many factors of not very pleasing appearance before we can feel satisfied with the fit. Experience and representative mathematical calculation suggest that infinity, in practice in this area, is not as far away as we might think. That is, the behavior of as few as a dozen or, conservatively, 20 measures can give a very good idea of the behavior of the entire set from which we are pretending they were drawn. It may be shown that the partial image weights should all tend to zero if the common factor model "holds" in the limit and should all be "small" in observed correlation matrices of reasonable size (a dozen or more variables). Table 2.3.6 gives a correlation matrix for which the common factor model is entirely inappropriate, and gives an image analysis of it. The weights for this case may be contrasted with the weights in Tables 2.3.4 and 2.3.5.

The main contribution of image theory to common factor analysis so far has been the provision of approximate methods for exploratory analysis, recommended by Jöreskog and Kaiser, to be discussed shortly. Perhaps the more important contribution it has to offer is the way it supports the view that factor-analytic investigations should be directed at clearly conceived and defined behavior domains, using clearly representative measures. Such investigations will usually be confirmatory rather than exploratory.

TABLE 2.3.6
Image Analysis of Chain Model Data
[See section 4.2 (c)]

Correlation Matrix

1.000	.500	.333	.250	.200	.167
.500	1.000	.666	.500	.400	.333
.333	.666	1.000	.750	.600	.500
.250	.500	.750	1.000	.800	.667
.200	.400	.600	.800	1.000	.833
.167	.333	.500	.667	.833	1.000

Regression of Each Variable on Remainder

Variable	Squared Multiple Correlation	Regression Weights					
1	.250	.0	.500	.000	.000	.000	.000
2	.531	.313	.0	.561	.000	.000	.000
3	.675	.000	.388	.0	.556	.000	.000
4	.754	.000	.000	.421	.0	.545	.000
5	.801	.000	.000	.000	.440	.0	.539
6	.694	.000	.000	.000	.000	.831	.0

(c) Approximate Methods

Factor analysis has been developed in the course of 70 years of work, of which only the last 20 have been aided by electronic computers. Especially in the work from the 1930s to the 1950s, tremendous emphasis had to be placed on methods for turning the mathematics of factor analysis into feasible arithmetic for desk calculation. Unfortunately, this has led to much confusion in the literature, some of it built into the traditional language of factor analysis; confusion between sample and population, between measures of fit and methods of fitting, and between central concepts of the theory and arithmetic devices including crude approximation devices for saving work.

The modern view is that the common factor model is a statistical hypothesis that may be "true" of a population and that at least prescribes the number of common factors. Given at least this prescription, we can use any one of a number of arithmetic algorithms to get (least squares or maximum likelihood) "best"

estimates of the population factor loadings and of the population uniquenesses and use the likelihood ratio criterion or inspection of residual covariances to reject or retain the hypothesis, as described in Section 2.2.

Partly because of the pressure of arithmetic problems in multiple factor analysis and partly from an actual failure to be as precise about concepts as we, the inheritors of decades of creative work on the subject can be, the earlier workers have handed down to us a number of confused modes of thought that are still to be recognized in the vocabulary of factor analysis. The reader will find, elsewhere, the persisting notion that in factor analysis we first "guess" the communalities of the variables and then "extract" common factors one by one until the residuals are small enough to suit our taste. This process yields "obtained" communalities that are different from the "guessed" communalities. These treatments (1) tend to confuse basic conceptual aspects of the model with crude, simplified arithmetic procedures for fitting it; (2) cannot by their nature produce best fit. We can obtain best fit only when, under a hypothesis as to the number of common factors, we produce $nm + n$ estimated quantities (the factor loadings and the uniquenesses), each of which is best in the context of the others. Any method that computes the first column of factor loadings and never changes it when the others are computed cannot be choosing best numbers. Such treatments should be thought of as approximate methods, replacing "best" estimation procedures.

There would seem to be at least four valid reasons for using approximate methods. The first is, quite simply, to save computing costs. The cost of ML estimation in common factor analysis can be a considerable percentage of the cost of the entire research project. A saving here could certainly be justified if an approximate analysis or a series of analyses were used to guide decisions governing a final, "best" analysis of the data, by the methods of Section 2.2 or of Chapter 3. The second is that a study may involve so many variables that its dimensions will not fit the limitations of the available ML computer program. We could, as a matter of fact, question whether a really large-scale study should ever be carried out. That is, will an investigator ever be able to collect together, on good, rational grounds, a really large number of measures that are likely to yield a clear, interpretable analysis? But one cannot legislate against such studies. The third use for approximate methods is one that the ordinary user does not even have to know about as it is hidden away inside a computer program for an exact method. That is, it saves money to have an approximate analysis that provides a starting point, not far away from the solution, for the numerical methods that yield ML or LS estimates. The fourth reason for using approximate methods is extremely questionable. The frequency of occurrence of Heywood cases (see following) in ML estimates leads some workers to recommend approximate analyses because these do not yield Heywood results. (For the moment, think of these as "impossible" estimates.) On the other hand, when a Heywood case occurs, the data may be trying to tell us something, namely, that the study

has not been well designed in the sense that not enough variables have been included to define each factor adequately. In any case, if we are otherwise convinced that the model is appropriate, we can use the Bayesian estimation procedure described in the following.

Approximate methods that have been strongly advocated and/or are in common use include: (1) the use of principal components of the sample correlation matrix, instead of common factors; that is, we compute correlations between the variables and the m main principal component scores and regard these as factor loadings; (2) the use of "squared multiple correlations (SMCs) as communalities"; that is, we use the residual variance of each variable about its regression on the remaining $n - 1$ (its partial antiimage variance) as an approximation to the estimate of its unique variance; (3) the use of principal components of the partial images of the variables as common factors; (4) the use of "reduced" partial antiimage variances as unique variances; that is, we find a constant, less than one, by which to multiply these variances, to allow for the fact that they are strictly greater than the unique variances.[2]

Table 2.3.7 gives the results when these four methods are applied to a simple one-factor case. Table 2.3.8 lists the values of the LS fit function Q (equation 2.2.1) for representative published correlation matrices. From these cases it is fairly clear that method (4) is best, followed closely by (2), and then (3) and (1) virtually tied.

At this point we should note also that a traditional problem of exploratory factor analysis has concerned nonstatistical criteria for deciding what the number of common factors, m, might be. A criterion that can indeed be a useful guide for a first analysis has been recommended for a number of different reasons. It is often mentioned briefly in accounts of applications as "eigenvalues (or latent roots) greater than one." Interpreted, this means that the number of principal components of the observed variables that explain more than one unit of variance (whose sum of squares of correlations with the n variables is greater than unity) indicates the smallest number, or the actual number, of common factors that should account for the correlations. A discussion of the logical basis of this criterion is a highly technical matter, and the reader is invited to accept that it is sometimes a good practical guide (and occasionally very very bad). We therefore expect to find that an ML estimation program for exploratory factor analysis will allow us to prescribe the number of common factors, if we are prepared to do so, or will let us tell the computer to choose a first hypothesis on the basis of the number of eigenvalues of the correlation matrix greater than one or greater than a number that we supply (and we are expected to supply the number one). We can

[2]Of these four methods, it is difficult to attribute the first or the second to an individual investigator. The third is due to Kaiser (1970), and the fourth to Jöreskog (1962). All the last three methods are based on Guttman's classical results on image theory (see Mulaik, 1972).

TABLE 2.3.7

Correlation Matrix (Repeated)

1.00	.72	.63	.54	.45	.36
.72	1.00	.56	.48	.40	.32
.63	.56	1.00	.42	.35	.28
.54	.48	.42	1.00	.30	.24
.45	.40	.35	.30	1.00	.20
.36	.32	.28	.24	.20	1.00

Correct Solution

.9
.8
.7
.6
.5
.4

(1) Principal Component "Factor Pattern"

.8832
.8326
.7697
.6937
.6044
.5023

Residual Matrix

.220	-.015	-.050	-.073	-.083	-.084
-.015	.307	-.081	-.089	-.103	-.098
-.050	-.081	.407	-.114	-.115	-.107
-.073	-.098	-.114	.519	-.119	-.108
-.083	-.103	-.115	-.119	.635	-.104
-.084	-.098	-.107	-.108	-.104	.748

Q = .0088523

(2) SMCs Factor Pattern

.8467
.7898
.7025
.6071
.5083
.4078

Residual Matrix

.283	.051	.035	.026	.020	.014
.051	.376	.005	.000	-.001	-.002
.035	.005	.506	-.006	-.007	-.006
.026	.000	-.006	.631	-.009	-.008
.020	-.001	-.007	-.009	.742	-.007
.015	-.002	-.006	-.008	-.007	.834

Q = .0003664

(3) Partial Image Factor Pattern

.7769
.7248
.6447
.5571
.4664
.3742

Residual Matrix

.396	.157	.129	.107	.088	.069
.157	.475	.093	.076	.062	.049
.129	.093	.584	.061	.049	.039
.107	.076	.061	.690	.040	.031
.088	.062	.049	.042	.782	.025
.069	.049	.039	.031	.025	.860

Q = .0064568

(4) "Reduced" Partial Antiimage Uniquenesses

.8566
.7991
.7108
.6142
.5143
.4126

Residual Matrix

.266	.035	.021	.014	.009	.006
.035	.361	-.008	-.011	-.011	-.010
.021	-.088	.495	-.017	-.015	-.013
.014	-.011	-.017	.623	-.016	-.013
.010	-.011	-.015	-.016	.735	-.012
.006	-.010	-.013	-.013	-.012	.830

Q = .002464

TABLE 2.3.8

	Harman: 8 Physical Variables	Harman: 24 Psychological Tests	Matrix in Table 2.2.2
(1)	.002765	.003356	.003662
(2)	.000570	.001711	.000432
(3)	.002651	.002461	.004630
(4)	.000532	.001710	.000113

also expect to find that many published factor analyses, especially the approximate ones, will say that the number of factors was determined as "the number of eigenvalues greater than one." As mentioned briefly at the end of the discussion of principal components, the reader should be wary of ambiguities in published applications of factor analysis. Many applications, right up to the current decade, use an approximate method. It is sometimes difficult to tell what approximation has been used, and usually it is impossible to evaluate the goodness of fit of the model and hence to decide how seriously the results can be taken. It is hard to give advice about this situation. If the investigator has published the correlation matrix, or it is accessible, we can always reanalyze it properly if sufficiently concerned. But often editors do not permit the correlation matrix to be published, and the description of the analysis is so scant that we would not even know how to repeat it to compare our results with details of the original. It must, then, be admitted that many published factor analyses do not meet the usual criteria for respectable research reporting of presenting sufficient detail to enable independent testing.

Broadly, if enough information is given to let us believe that the sample correlation matrix is a reasonably stable estimate of the population matrix, an approximate exploratory analysis will often as a matter of fact resemble a "best-fitted" analysis enough to yield the same conclusions. But if we are given only, say, the rotated factor matrix, it is perhaps better to regard the study as never having been carried out at all.

A high percentage of published studies present something like the following information: "The correlations were factored using principal components with eigenvalues greater than one, and rotated to simple structure using varimax, yielding the rotated factor loadings shown in Table 2.3.9." Sometimes also "loadings below .3 are omitted," as though leaving out these numbers helps the reader to comprehend the results. The statement is ambiguous, for we do not know, and indeed the investigator may not know, if the program borrowed from the computer center obtained only principal components, as in approximation (1), or crudely fashioned a closer approximation to LS estimates out of these.

TABLE 2.3.9
Varimax Factor Pattern

	I	II	III	IV	V	h^2
1	0.143	.522	.197	.478	-.084	.568
2	.224	.197	.608	.225	.166	.537
3	-.049	.134	.002	-.186	.790	.678
4	.737	.303	.177	-.131	-.118	.697
5	.024	.270	.065	-.800	.148	.741
6	.746	-.267	-.161	.099	.070	.669
7	.012	-.526	.002	.004	-.329	.386
8	.261	-.271	.050	.489	.488	.621
9	.155	.081	-.819	.197	.102	.752
10	.059	.672	-.347	.081	-.151	.605

"The data were factor analyzed using principal components with eigenvalues greater than one, followed by varimax rotation."

- How to lie with factor analysis.

Documentation of package programs for "factor analysis" does not always make this clear.

If we are presented with just the result in Table 2.3.9, we are at a loss to evaluate it. In fact, it is an analysis of the correlation matrix in Table 1.4.4. The correct number of common factors is zero, not five; hence the "eigenvalues greater than one" guide fails us completely. It is disconcerting to find that the process of transformation toward approximate simple structure can yield enough large and small values of the factor loadings to give what some factor analysts quaintly refer to as a "compelling" simple structure (i.e., presumably, one that "compels" our faith in it.) It is comforting to know that a reanalysis of the correlation matrix, if this were available to us, would have given us "the truth," namely that it is drawn from a population of uncorrelated variables with no common factors (i.e., no generic properties in common).

(d) Heywood Cases (Improper Solutions)

With the increasing use of good methods of estimation, investigators are increasingly encountering cases where the best-fitted estimates are *improper*, because one or more estimates of uniqueness (residual variance) are negative. This

of course is unacceptable, as variances are essentially positive quantities (means of squares). Even a zero residual variance is unacceptable, as it implies exact dependence of an observed variable on the common factors. This could only be true if the variable has no measurement error. That negative uniquenesses might arise was first pointed out by Heywood (1931); hence it is commonly referred to as a *Heywood case*. Alternatively, it is known as an *improper solution*. We can summarize the situation in just six points and draw some tentative conclusions.

(i) Some investigators tend to regard the fact that Heywood cases can occur as an indication that something is wrong with the basic principles of the common factor model and that we should use some other technique of multivariate data analysis instead—perhaps principal components or image analysis.

(ii) A Heywood correlation matrix is a perfectly possible correlation matrix for a population.

(iii) On the other hand, a "non-Heywood" population can give samples, by chance, in which the *estimators* of some positive population residual variances are negative; hence a Heywood case in a sample does not *prove* that the population is a Heywood case. A second sample might yield a different conclusion.

(iv) Sometimes a Heywood case can be cured by fitting fewer factors, but often this gives unacceptably poor fit.

(v) Most modern programs for ML or LS estimation are arranged to stop the search for a minimum of the function with respect to any uniqueness before it becomes negative. Such a procedure is at best a makeshift, as we know that we have not found the required minimum, and a zero uniqueness is still really unacceptable.

(vi) A very common cause of Heywood cases seems to be a failure on the part of the investigator to represent each factor by a sufficient number of tests with large loadings on it. Consider the two simple general-factor correlation matrices in Table 2.3.10. In case (a), by simple inspection, as in the earlier discussion in Chapter 1, we deduce that the single-factor model fits the correlations with loadings .6, .5, and .4. These loadings are determined precisely by the data, as are the corresponding unique variances, .64, .75, and .84. In case (b), we find that the loadings and unique variances are not uniquely determined by the data. We can choose loadings of .6, .5, and zero or 1.2, .25, and zero or .2, 1.5, and zero or, indeed, any two numbers whose product is .30 for the first two loadings with zero for the third. And we note immediately that if we can easily find pairs of numbers for the loadings in which one or the other number is greater than unity and the corresponding unique variance is negative, so can the computer, of course. This fact seems to have been well-known to Thurstone in the 1930s as an indeterminacy of the parameters of *doublet factors*—factors with only two nonzero loadings—but its implications for the occurrence of Heywood cases have not been widely recognized. More generally, if one or more of m common factors have only two tests with nonzero loadings, then those loadings and the

TABLE 2.3.10

(a)

1.00	.30	.24
.30	1.00	.20
.24	.20	1.00

(b)

1.00	.30	.00
.30	1.00	.00
.00	.00	1.00

(c)

1.00	.30	.024
.30	1.00	.02
.024	.02	1.00

corresponding unique variances are not *identified* (uniquely determined by the correlations), and Heywood cases are likely to occur. Notice that the negative unique variance is not specifically associated with one of the variables, and the phenomenon need not be eliminated by deleting the variable with negative residual variance from the analysis.) More generally still, if one or more factors have only one or two tests with large loadings and the rest of the loadings are small (*singlet* or *doublet* factors, in Thurstone's terminology), the factor loadings on this factor may be very poorly estimated from even a large sample. Consider case (c) in Table 2.3.10. The correlations are consistent with only one set of loadings, .6, .5, and .04, but because the last loading is close to zero, it is easy to imagine that in finite samples it would behave like case (b), yielding estimates of loadings greater than unity, and estimates of residual variances that are negative. (It is not yet common practice to compute standard errors of estimate of factor loadings and uniquenesses, to put confidence bounds on them and determine how well they have, individually, been estimated. It is technically possible and desirable to do so.) Generally, experience suggests that we can hope to avoid Heywood cases and, indeed, poorly estimated common factor loadings and unique variances, if we make sure that every common factor is defined by at least three and preferably four or more variables having large loadings on it. This is reasonable in terms of the substantive aims of factor analysis, because we can hardly expect a common abstract attribute to be well defined by just two measured of it.

Negative estimates of essentially positive quantities occur in statistical problems other than those of factor analysis. One approach to these problems involves the adoption of the Bayesian philosophy of statistical estimation. Briefly, in this way of thinking, we attribute a probability distribution to the population

parameters we wish to estimate in which we may incorporate and give fairly exact expression to our beliefs about the population studied. Here, for example, we might turn our belief into the assertion that the probability of finding a negative or zero residual variance is zero, whereas the probability of finding nonzero values is greater than zero. It is possible to turn this thinking into a plausible probability distribution of uniquenesses, which leads to a simple modification of a ML estimation program that prevents it from obtaining negative or zero uniquenesses.[3] This approach to the problem is possibly better than the common device mentioned before of stopping the analysis at the point where a residual variance reduces to zero; but before either of these devices is adopted, the investigator should check whether the data yield one or more singlet or doublet factors, in which case there is a more serious problem in the form of ill-defined factors and underidentified (nonuniquely determined) parameters with the study.

2.4. DEVICES FOR APPROXIMATING SIMPLE STRUCTURE

The simple hypothesis of exploratory common factor analysis that prescribes only the number of common factors is not specific enough to determine unique estimates of the common factor loadings if the number is more than one. We say that these parameters are not "identified." When we obtain a set of estimates, we have to recognize that infinitely many alternative sets would fit the data equally well. The mathematician can tell us how to compute from a given set of factor loadings all the possible alternative values, which are *transformations* of the values we first happen to obtain.

A widely accepted goal, in transforming a given factor pattern into another, is contained in the notion of *simple structure*. Given an explanation of the intercorrelations of our n variables in terms of a minimum number, m, of common factors, the basic notion of simple structure is, further, that we explain the correlation of each variable with the others by a minimum number of those common factors. That is, broadly, a factor pattern has simple structure when each variable has nonzero loadings (regression weights) on as few of the factors as possible. Partly on the basis of experience, five rules have been given for simple structure that are supposed to legislate an unambiguous choice among alternative solutions that might be equally acceptable in terms of the fundamental definition. These are: (1) Each row of the factor pattern should have at least one zero element. (2) Each column should have at least m zero elements. (3) For every pair of columns there should be at least m variables with a zero coefficient in one column and a nonzero in the other. (4) In the case where m is greater than

[3] See Martin and McDonald (1975).

three, for every pair of columns there should be a large proportion of variables with zeros in both columns. (5) For every pair of columns, there should be a small proportion of variables with nonzeros in both columns. Notice that only the first three of these rules are objective. It is customary to illustrate simple structure by using zeros and crosses as in Table 2.4.1 to indicate the positions of zero and of nonzero elements. Table 2.4.1(a) shows a simple structure that fits all the criteria. Table 2.4.1(b) shows a more specialized structure that is certainly a simple structure and actually fits the stronger principle that every variable has a nonzero loading on only one factor. This structure is known as *independent clusters*.

Thurstone originally advocated the simple structure principle as reflecting a truth about the psychology of cognition where the concept originated (viz., that we do not seem to use all our "mental faculties" in any one cognitive task). (Thurstone, at the time of the introduction of simple structure, explicitly regarded factors as a scientific revival of an old discredited unscientific notion of mind as made up of "faculties.") The concept also carried with it a second principle of parsimony to supplement the first, by which we first explain all the correlations with as few factors as possible and then explain each correlation with as few of those few as possible.

The unquestioning acceptance of simple structure in factor patterns is possibly a reflection of human conceptual preferences rather than of anything in the subject matters we study. Essentially, a practice of factor interpretation has evolved, again without formal and explicit argument, in which we describe a factor as that which is in common to those variables that have large positive

TABLE 2.4.1

(a) Simple Structure Factor Pattern

X	0	X
X	X	0
X	0	0
0	X	X
X	X	0
0	X	0
0	0	X
0	X	X
X	0	X

(b) Independent Clusters Factor Pattern

X	0	0
X	0	0
X	0	0
0	X	0
0	X	0
0	X	0
0	0	X
0	0	X
0	0	X

regression weights on it. The factor is "most like" the variables that increase most rapidly as the factor score increases. It is unlike the variables with zero loadings, as these do not vary as the factor varies, and *least* like those variables that have large negative regression weights on it (i.e., the variables that decrease most rapidly as the factor score increases). But for large classes of data the methods of scoring employed combined with the simple structure criterion tend to eliminate negative loadings and leave a contrast between large positive loadings and zero loadings. The interpreter can then ask what the high-loading group of variables has in common in order to identify the factor. He or she does not have to ask what the zero-loading variables have in common, of course. If there are any large negative loadings, one can usually think of those variables as, in effect, scored in reverse (extraversion scored as lack of introversion or anxiety scored as lack of calmness). The task of interpretation seems easier when we just ask what a given group of variables have in common than when we have to contrast them with another group of variables. In fact, unfortunately, investigators sometimes miss the elementary precaution of making sure that variables with zero loadings on a factor do not *also* have the generic property that the high-loading variables seem to have. This would make nonsense of the proposed interpretation of the factor. It might seem, then, that zeros in the factor pattern are desired because the zero–nonzero contrast is easy to think about.

It follows from the definition of simple structure that if we have a population whose correlation matrix can be explained by a factor pattern with a definite simple structure, it is quite impossible (i.e., there is a zero probability) that a sample drawn from that population could be made to yield simple structure by transforming a given set of estimates in which no factor loadings have been set equal to zero. At best we could hope to invent devices that yield an approximation to simple structure that is "good" in some sense. We might also hope to find a way to decide when the approximation is "good" enough to enable us to believe that the variables possibly have simple structure in the population. Unfortunately, in the literature, no distinction is made between "simple structure" and "approximate simple structure" or between simple structure in the population and simple structure in the sample. A habit has grown up, with no formal or explicit basis, of regarding sample loadings less than .3 as "near-zero" in a loose definition of simple structure. Simple structure in this loose definition seems rather too easy to obtain, and investigators often interpret their results *as though* they had simple structure when they might be said not to have it at all.

Many devices have been invented for "rotation to simple structure" (i.e., for transforming a given estimated factor pattern so that it approximates the definition of simple structure as closely as it is able to, perhaps not at all closely). Just about any of these devices are adequate to indicate to us, in combination with rational thought about the objects of the study, where to postulate zero loadings in a detailed hypothesis that can be tested by the methods of Chapter 3. It seems illogical to be content with an approximate simple structure in which some of the

coefficients in the factor pattern are "small" and others are "large." We either believe the small coefficients are zero in the population, or we do not. If we do, we should not get nonzero estimates of the zero coefficients. If we do not, we should not pretend to be using simple structure.

Essentially, there are four main approaches to the problem of obtaining a transformation to approximate simple structure. These are (1) graphical methods, (2) counting methods, (3) simplicity function methods, and (4) target methods.[4]

1. The oldest method involves the drawing of graphs, pairwise plots of the columns of factor loadings against each other, by human operators and the selection by eye of new axes for the graphs. This is an extremely complicated art. Most investigators consider it satisfactorily replaced by methods that can run themselves off on a computer, thus saving human effort. It does seem, though, that the results of graphical transformations tend to be considered the standard by which the results of other methods are judged.

2. The intention behind counting methods is that we count the number of variables that have a loading less than a given size (say .3) on each factor and look for a solution that maximizes this number. Because of the geometry of the problem, originating in the graphical treatments, this count of small values is known as the *hyperplane count* (the number of points close enough to a plane in multidimensional space). However, as carried out in practice, instead of a simple count of the number of small-enough values, a weight is given to each element counted that makes the total count a function of the size of the large coefficients rather than just an integer representing the number of small-enough coefficients. This seems to constitute a departure from the original principle, according to which simple structure is a matter of the number of small factor loadings and surely should be quite independent of the size of the large ones.

3. In the *simplicity function* methods, the basic problem put to the mathematicians is to define a quantity that is computed as a function of all *nm* elements of the common factor pattern and will vary as we transform the numbers in the factor pattern, becoming a minimum (or, for some functions, a maximum) at a set of values of the factor loadings that we would regard as a reasonable approximation to simple structure. Such a function is called a *simplicity function*.

On the face of it, it looks impossible to define a usable simplicity function. In the first place, there are a number of distinct ingredients to the original recipe for simple structure. It would seem impossible to capture them all in a single mathematical function. In the second place, it would seem incorrect to have a function that depends on the values of the "large" elements in the transformed pattern, because the simple structure concept has no implication at all for the sizes of elements that are thought to be nonzero. Nevertheless, a number of simplicity functions have been defined that appear to work well in practice. No attempt will

[4]For general comments, see Hakstian (1971) and Hakstian and Abel (1974).

be made here to distinguish the different variations that have been invented. We just examine the general idea.

Broadly, a solution to the problem of defining a simplicity function can be based on the commonsense reflection that a factor pattern matrix that exhibits simple structure has an extreme distribution of the absolute sizes of its elements, in which there are many large (positive or negative) values and many small values with few of intermediate size. Such a spread of the values to the extremes could be measured by one of the usual measures of variability in descriptive statistics. A convenient choice would be to square the *nm* elements of the factor pattern, because we want the contrast to be between absolute values—very large versus very small—rather than signed values—large positive versus large negative. We would then try to find a transformation that maximizes the variance of the *nm squared* numbers.

Competing variants on this idea have been developed, and claims made about the general relative qualities of the results obtained. It seems impossible to find a simplicity function that is "better than" other simplicity functions in the sense that it always gives results nearer to (a) the known simple structure of artificial test data or (b) graphical solutions. Because simplicity functions depend on the irrelevant "large" values of the factor loadings, the solution given by one simplicity function will differ from the solution given by another simplicity function *and* from the "best" solution as otherwise judged, by reason of irrelevant values of factor loadings that differ from one example to another. It is doubtful, therefore, if there could be a way to show that one simplicity function is "generally best." If we use an approximate simple structure only as a guide for setting up detailed hypotheses, as in Chapter 3, this does not matter.

4. In target methods, (also rather unfortunately described as *Procrustean* methods), we suppose we know where the zeros would be in an exact version of the simple structure, and we choose a transformation to make the loadings corresponding to the "target" zeros as small as possible. (Usually we minimize the sum of squares of those numbers.) The main advantage is that the result is independent of the large loadings. The main disadvantage is that we must first choose a target. In practice, we can use a target method to improve a result obtained by one of the other methods, which also yields an automatic decision as to the location of the exact zeros.

The user of computer programs for "rotation to simple structure" could obtain some guidance from Table 2.4.2. The main choice is between "orthogonal rotation," yielding a new solution that is also in terms of uncorrelated (orthogonal) factors, with a common factor pattern in which the factor loadings are *v–f* regression weights and also *v–f* correlations, and "oblique rotation", yielding a common factor pattern (*v–f* regression weights), a common factor structure (*v–f* correlations), and the correlation matrix of the factors (*f–f* correlations). The main argument for orthogonal transformation is that factors

TABLE 2.4.2

(a) Orthogonal Transformations

(i) QUARTIMAX: Simplicity Function $s_q = \sum_{j=1}^{n} \sum_{p=1}^{m} f_{jp}^4$

the sum of the fourth power of loadings. (Maximized) Tends to "simplify" the rows but not the columns of the factor pattern--may leave a "general factor" with no near-zero loadings. (Due to Carroll, 1953.)

(ii) VARIMAX: Simplicity Function

$$s_v = \sum_{p=1}^{m} \left[\frac{n \sum_{j=1}^{n} (f_{jp}^2)^2 - (\sum_{j=1}^{n} f_{jp}^2)^2}{n^2} \right]$$

the sum across columns of the "variances" of the squared loadings in the m columns. Usually the method is applied with the loadings "normalized"--divided by the square root of the communality--to make each row sum of squares equal unity. (Maximized) Tends to avoid a "general" factor. (Due to Kaiser, 1958.)

(iii) TRANSVARIMAX: A weighted sum of s_q and s_v is used as simplicity function. (Due to Saunders, 1962.) General Comment: VARIMAX is most widely available, and most popular. In exploratory work, it seems to suffice.

(b) Oblique Transformations

(i) (DIRECT) OBLIMIN Simplicity Function

$$s_{do} = \sum_{p=1}^{m} \sum_{q \neq p}^{m} \left[\sum_{j=1}^{n} f_{jp}^2 f_{jq}^2 - \frac{1}{n} (\Sigma_j f_{jp}^2)(\Sigma_j f_{jp}^2) \right]$$

(Minimized) We minimize the "covariance" of squared loadings in distinct columns. Recommended by Hakstian (1974). (Due to Jennrich and Sampson, 1966.)

continued

continued

(ii) OBLIMAX Simplicity Function

$$s_o = \frac{\sum_{j=1}^{n} f_{ip}^{*4}}{(\sum_{j=1}^{n} f_{ip}^{*2})^2} \qquad p = 1, \ldots, m$$

For each factor in turn, the function is maximized, then the process is repeated. The quantities f_{jp}^* are not the common factor loadings but are related to them by a scale transformation. (They are known as reference-structure loadings.) Not recommended by Hakstian, 1974. (Due to Saunders, 1961.)

(iii) BIQUARTIMIN. Simplicity function resenbles s_{do}. Not recommended by Hakstian. (Due to Carroll, 1957.)

(iv) MAXPLANE. Originally intended to maximize the number of loadings whose absolute value is smaller than a given number--a counting method (i.e., to maximize the hyperplane count). In practice, weights are used as discussed in the text. Not strongly recommended. (Due to Cattell and Muerle, 1960.)

(v) PROMAX A target method. Using, say, VARIMAX, we obtain an approximate simple structure. The loadings are raised to a higher power to exaggerate the difference between the large and small loadings. Then an oblique transformation is chosen that uses the "powered" loading matrix as a target. Recommended. (Due to Hendrickson and White, 1964.)

(vi) Harris-Kaiser oblique transformations: Essentially a method for restricting the kind of transformation chosen. Cannot be described here. Certain methods suggested are recommended by Hakstian. (Due to Harris and Kaiser, 1964.)

are principles of classification that should be as independent as possible (i.e., uncorrelated). The main argument for oblique transformation is that factors that are uncorrelated in one population may well be correlated in another, and correlated factors will tend to give invariant v–f regression weights (suitably scaled—see Chapter 6) from one population to another. We would have the best of all worlds if a set of variables gave us uncorrelated factors, simply as a matter of fact, in all the populations we happen to care about, but this cannot be expected.

2.5. RELATED METHODS

In this section we briefly consider three techniques that bear some relationship to exploratory factor analysis, namely inverse factor analysis, optimal scaling, and multidimensional scaling. At least the last two of these topics are major fields of psychometric theory, and each requires no less than a book-length account to do it justice. As in the discussion of principal component theory and image theory, the treatment of these topics here is partial in both senses of the word, being both incomplete and biased toward a perspective that is essentially that of common factor analysis.

(a) Inverse Factor Analysis

An extremely confused issue in factor theory concerns the notion of "factoring persons instead of tests" in the usual context of persons taking tests as the source of our data. If we think of factor analysis as something we "do to" an $(n \times n)$ matrix of correlations between n tests measured on N persons, and if we think of sample correlations as mean products of standardized deviations of persons from their means on two tests, then it is easy to invent "inverse" or "obverse" or "converse" factor analysis as something we would "do to" an $(N \times N)$ matrix of "correlations" between N persons measured on n tests. We would immediately perceive difficulties with such an "inverse" factor analysis. If n tests are measured in n different sets of units, with n different origins and scales, we would wonder what the correlation between Smith's and Brown's sets of n scores would mean, and we would notice that the correlation would be sensitive to changes of unit. For example, measuring weight in tons versus milligrams, height in feet versus millimicrons, and length of big toe in inches versus miles would change the correlations dramatically. In spite of these difficulties, a large literature has developed on the subject of "factoring persons." Much of it has been devoted to the question whether the factor loadings obtained from correlations between persons should be in correspondence to their usual factor loadings. Much too was concerned with the effects on such correspondence of "taking out" or "leaving in" means or of rescaling the variables to comparable units before computing correlations between persons.

If we do not accept the view that factor analysis is something we do to correlation matrices and if, specifically, we regard the common factor model as a special case of latent trait theory, based on the principle of local independence (see Chapter 7), we may find it difficult to see why the notion of "factoring persons" ever arose in the first place. That is, it is fairly easy to understand a common factor as a latent trait such that in a subpopulation of persons for whom that trait is a fixed number the correlation between two tests is zero. It is hard to understand a factor, whose loadings are obtained by analyzing correlations between persons, as a latent trait such that, in a subpopulation of tests for which that trait is a fixed number, the correlation between two persons is zero. Thus, the main difficulty with the notion of applying the common factor model to a matrix of "correlations between persons" would be the logical difficulty of interpreting the residual covariances as partial covariances, the uniqueness of a person as the residual variance about the regression of the person on the factors, and so on. The position taken here is that the common factor model is a statistical model and not a device that is applicable "inversely" to "correlations" between persons. Nor, it seems, has any cogent need to apply the model in this way ever been demonstrated.

The case is somewhat different with component theory. If we have scores y_{ji} of N subjects on n measures, which for the moment we suppose to be in raw score form, we may approximate the scores by sums of products of principal component weights and principal component scores. The detailed mathematics of the problem can be presented so as to give the impression that we choose between first obtaining and operating on sums of products of the scores of pairs of variables, or sums of products of the scores of pairs of subjects. These resemble "correlations between variables" and "correlations between persons." We might very loosely describe these procedures as "factoring tests" and "factoring persons," but either is just a device to solve the entire minimization problem with convenient arithmetic. It is not to be expected that if we first transform the scores to deviation measure or to standard measure in the sample the best-fitting weights and scores will be related to the weights and scores before rescaling in any simple way. This fact, however, does not seem to be a problem of any depth or consequence for psychometric theory. Any wish to obtain a best-fitting representation of a given set of scores will presumably be in turn motivated by rational research considerations. These in turn, in most cases, should dictate whether we wish to approximate the scores or their deviations from the mean or their deviations in standard measure by principal components; hence the discussions of the effects on the relationship between "factoring tests" and "factoring persons" of "taking out means" or "standardizing" do not yet seem adequately motivated.

The cavalier attitude expressed so far in this section toward problems that have been taken very seriously by very competent investigators should not deter the reader from inquiring more deeply into these matters if the nature of his or her research data would seem to make it necessary. On the other hand, it certainly seems desirable not to become involved with such problems if it is possible to avoid them.

The use of measures in a score matrix in which each row consists of deviations of the subject's scores from his or her own mean over n tests is sometimes solemnly discussed as *ipsative* scoring, with the obvious Latin derivation. The process of converting a score matrix to this form is known as *ipsatization*. Usually, to give such a scoring scheme the semblance of rationality, the scores would first have to be put in standard measure in the sample. The effects of

ipsatization are not well understood, and it would seem very difficult to develop proper statistical theory to cover estimation problems for sample data so treated. There is need for further work, perhaps directed at the question whether we could ever have any good reason to ipsatize.

Problems of a rather different kind arise with other $n \times N$ data matrices that we might consider factor analyzing. If, for example, just one test is administered n times to N subjects, we are free to calculate the correlations between the n repeated measures and fit the common factor model to the data. If the n administrations of the test are all carried out under the same conditions, yielding N time series, one for each subject, the use of the common factor model would seem conceptually inappropriate, and we presumably would prefer to use a conventional time-series analysis. If we insist on using common factor analysis, it is unlikely that we shall be able to interpret the results in terms of common properties of times of testing, such as early versus late or middle versus early and late. If, on the other hand, the n repeated measures correspond to n distinct situations in which the test was administered, it may prove possible to interpret the analysis in terms of common properties of situations. See Chapter 6 for the analysis of data consisting of subjects by tests by occasions or situations.

(b) Optimal Scaling

Optimal scaling is one of several names (dual scaling, correspondence analysis) that have been given to certain applications of principal component analysis to multicategory data. If N subjects respond to n multicategory items, the responses can be coded in a data matrix of N rows and p columns, where p is the total number of categories in the n items. We record a unity in the column corresponding to the category of each item that each subject checks and zeros in all other columns. If the respondent is forced to choose a category in each item, each row of the data matrix must contain just n unities, one for each item. As a result, there is redundancy of information in the matrix. If we know the entries in all but one category of each item, then we know the entry in the remaining category. The object of optimal scaling is to choose weights for the item categories and scores for the subjects that are *optimal* in a mathematically well-defined sense of the word. A number of criteria have been proposed, all of which yield the same mathematical answer, which closely resembles principal component analysis. We write y_{jl} for the entry corresponding to the lth category of the jth item. If the respondent checks this category, then $y_{jl} = 1$. Consequently $y_{jk} = 0$ for every other category, k, of the item. We define a total score s, weights for each item category w_{jl}, and item scores s_j, by writing

$$s = \sum_j \sum_l w_{jl} y_{jl} \tag{2.5.1}$$

and

$$s_j = \sum_l w_{jl} y_{jl} \qquad j = 1, \ldots, n \tag{2.5.2}$$

(In fact, s_j is always the same as the weight assigned to the category checked, and s is the sum of the weights of all the categories checked.) We choose the weights to maximize the sum of the squares of correlations between the item scores s_j and the total score s. This is analogous to Hotelling's original treatment of principal components. Such alternatives as choosing the weights to maximize the ratio of the variance of the total score to the sum of the n variances of the item scores yield the same answer. These and certain other equivalent criteria are essentially designed to maximize the relationship between the total score and the item scores in some recognizable sense.

The optimal weights in (2.5.1) are regression weights of the (dependent) optimal score s on the p (independent) item categories. They are indeterminate because the independent variables contain redundant information. This means that further arbitrary restrictions need to be placed on the weights to determine them uniquely. Once they are determined, under any set of restrictions, we can, as in principal component analysis, compute a converse regression of the item categories on the optimal scores. These are invariant under arbitrary choices of the optimal weights and can be interpreted very much as in common factor analysis. In the practice of optimal scaling, the usual procedure is to obtain and interpret some set of optimal weights. From the factor-analytic point of view it seems preferable to obtain the regressions of the item categories on the optimal scores rather than the regressions of the optimal scores on the item categories.[5]

(c) Multidimensional Scaling

Multidimensional scaling is the generic term for a family of methods for representing *dissimilarities* between stimuli by distances in a multidimensional space. Because it is possible to think of a correlation coefficient as measuring the similarity of two tests, it may seem reasonable to take some function of the correlation coefficients chosen to increase as the correlation decreases to measure the dissimilarities of a set of n tests and to use multidimensional scaling as an alternative to common factor analysis to provide an account of the relations between them.

Just as the position of a point in two dimensions can be represented by its coordinates, x_1, x_2, measured on two axes at right angles, so an imagined point in m dimensions can be represented by its coordinates, $x_1, \ldots, x_m$, measured on m axes at right angles. By an extension of Pythagoras' theorem, given the coordinates of any two points in an m-dimensional space, we can calculate the square of the distance between the points as the sum of the squares of the m

[5]See Nishisato (1980) for a general account of optimal scaling. For a technical account of these remarks, see McDonald (1983).

differences between their coordinates. The converse problem is more difficult, but it can be solved.

In multidimensional scaling, given the squared distances between members of a set of n points in m-dimensional space, we can find a set of m coordinate values for each of the n points that is consistent with those squared distances. The obtained coordinates are subject to indeterminacies corresponding to both a rotation of the coordinate axes and a movement in space of the origin of the system of axes.

In metric multidimensional scaling we assume that a set of given dissimilarities measures the distances between the objects (tests, stimuli) to be mapped into a multidimensional space. The obvious difficulty with this assumption is that it can easily contradict itself. Suppose one investigator uses the quantity $\frac{1}{2}(1 - r)$ where r is the correlation coefficient between tests as a measure of their dissimilarity, ranging from zero to unity, whereas another uses $-\log\{\frac{1}{2}(1 + r)\}$, ranging from zero to infinity. The two measures of dissimilarity cannot be taken as measures of the same distance.

Nonmetric multidimensional scaling was introduced to avoid the self-contradictory assumption that distances are measured by dissimilarities. It is possible to avoid doing any arithmetic on the numbers representing dissimilarities by regressing the distances in the model on the observed dissimilarities, using a *monotone regression function*. This is a nondecreasing function of the independent variable that gives a least-squares best fit to a scatter diagram. It takes the form of a set of joined-up straight-line segments (parts of a polygon) that are either horizontal or sloping upward from left to right in the graph of the data. Arithmetic algorithms have been developed for the two steps of nonmetric multidimensional scaling, namely: (1) given a set of guessed coordinates of the objects, yielding a corresponding set of guessed distances, to regress the distances on the data using a monotone regression function; and (2) to move to a new set of coordinates chosen to reduce the residuals of the distances about their regressions on the dissimilarities. At the completion of a series of repetitions of these steps, we should have a set of coordinates for the objects that minimize the residuals of the distances about their regressions on the data. It is an unusual and interesting feature of these methods that the hypothetical quantities in the model are treated as dependent variables and regressed on the data as independent variables in order to avoid doing arithmetic on the observations.[6]

For our purposes, the important question concerns the relation between common factor analysis and nonmetric multidimensional scaling applied to quantities derived from correlations between tests. There is no direct mathematical relationship. In applications, users of nonmetric multidimensional scaling usually obtain an account of data in terms of fewer dimensions than do factor analysts.

[6]For a general account of multidimensional scaling, see Kruskal & Wish (1978).

This seems partly due to choices open to the investigator. A user wishing to avoid severe rotation problems in multidimensional space may deliberately choose a coordinate space of at most two dimensions to contain the data. It also seems partly due to the fact that multidimensional scaling allows a translation of origin that can commonly be used to eliminate one of the dimensions needed by the common factor model. Some reduction in the dimensionality of the data may also be due to the nonmetric properties of the former method. Allowing for these differences, it is possible to find a degree of consistency between these alternative analyses of the same data.[7]

Instead of using correlation coefficients with their built-in linear measurement of the relations between tests, it is possible to develop a nonmetric counterpart of principal component analysis, in which the monotone regression function is used to regress a weighted sum of components on the data.[8] Such a method recovers known parameters provided that the data contain only a small amount of unique variance. A nonmetric analog of the common factor model would presumably be able to cope with large amounts of unique variance, but it does not seem possible to develop a common factor model without doing arithmetic on the data, so such a model appears to be quite a challenge for research!

2.6. MATHEMATICAL NOTES ON CHAPTER 2[9]

(a) Notes on Section 2.2

We write $\mathbf{A}$, of order $(n \times n)$, for the usual sample correlation matrix, computed from a sample of size N. (It is actually better to think of this and the matrix fitted to it as covariance matrices.) In the unrestricted common factor model, we wish to estimate $\mathbf{F}$ and $\mathbf{U}$ under the hypothesis that

$$\mathbf{R} = \mathbf{F}\mathbf{F}' + \mathbf{U}^2$$

for $\mathbf{F}$ of order $(n \times m)$, with no further specification on the elements of $\mathbf{F}$.

[7]An unpublished study by McDonald and Chan reveals close similarities in configurations of common factor loadings and configurations of points in a nonmetric multidimensional scaling analysis of functions of the correlations, except for the loss of a dimension due to movement of the origin in multidimensional scaling. See, for example, Schlesinger and Guttman (1969) for an alternative view of these matters.

[8]Kruskal and Shepard (1974).

[9]This section may be omitted, but it may help, so try it.

General Note: Again all of the material in this chapter is very well known, and again Gorsuch (1974), Rummel (1970), and Mulaik (1972) are recommended for further reading. The analyses were done on computer programs written by the author.

In the method of least squares we choose $\mathbf{F}$ and $\mathbf{U}^2$ to minimize the quantity

$$Q = \text{Tr}\ \{(\mathbf{A} - \mathbf{R})^2\} \tag{2.6.1}$$

that is, the quantity

$$Q = \text{Tr}\ \{(\mathbf{A} - \mathbf{FF}' - \mathbf{U}^2)\ (\mathbf{A} - \mathbf{FF}' - \mathbf{U}^2)\}. \tag{2.6.2}$$

By differential calculus, omitted, we find that conditions for Q to be a minimum are

$$(\mathbf{A} - \mathbf{FF}' - \mathbf{U}^2)\mathbf{F} = \mathbf{0} \tag{2.6.3}$$

and

$$\text{Diag}\ (\mathbf{A} - \mathbf{FF}' - \mathbf{U}^2) = \mathbf{0} \tag{2.6.4}$$

This is a system of simultaneous nonlinear equations, for which a solution cannot be obtained in closed form. That is, we cannot obtain expressions for $\mathbf{F}$ and $\mathbf{U}^2$ in terms of elements of $\mathbf{A}$. However, for any given value of $\mathbf{U}^2$, we can solve (2.6.3) for $\mathbf{F}$ using the mathematics of principal component theory, rewriting it as

$$(\mathbf{A} - \mathbf{U}^2)\mathbf{F} = \mathbf{FF}'\mathbf{F} \tag{2.6.5}$$

and choosing to impose a condition that $\mathbf{F}'\mathbf{F}$ be a diagonal matrix. Conversely, for any given value of $\mathbf{F}$, we can solve (2.6.4) for $\mathbf{U}^2$, giving the "obvious" result

$$\mathbf{U}^2 = \text{Diag}\ \{\mathbf{A} - \mathbf{FF}'\} \tag{2.6.6}$$

In practice, therefore, there have been two main approaches to the numerical solution of the least-squares estimation problem. In one, we use a numerical algorithm to find values of $\mathbf{U}^2$ that successively approach nearer and nearer to the minimizing values, and for each of these we solve (2.6.5) by the methods of principal component theory. Methods for finding successively improved values of $\mathbf{U}^2$ range from ad hoc algorithms (such as one due to Thomson 1934) that "seem to work" to applications of modern Newton or quasi-Newton methods. In the other method, we use a numerical algorithm to find values of $\mathbf{F}$ that successively approach the minimizing values, and for each of these we obtain $\mathbf{U}^2$ by (2.6.6). The best known version of this method is Harman's MINRES.[10] We can also minimize Q directly with respect to both $\mathbf{F}$ and $\mathbf{U}^2$.

We turn now to maximum likelihood estimation. Not enough information has been given in Appendix A1 to enable us to derive this method from basic principles, and no attempt will be made to do so.

We shall accept as given a result obtained by Lawley (1940) that under the normal distribution assumption the quantity

[10]See Harman and Fukuda (1966).

$$\lambda = N[\text{Tr}\ \{\mathbf{AR}^{-1}\} - \log\ |\mathbf{AR}^{-1}| - n] \tag{2.6.7}$$

has its minimum at the point where the likelihood of our sample has a maximum, and λ is distributed asymptotically as N increases like chi-square with $\text{df} = \frac{1}{2}\{(n - m)^2 - (n + m)\}$. Whether the normal distribution assumption is true or not, the quantity λ is necessarily nonnegative. It is "large" when the fit of $\mathbf{R}$ to $\mathbf{A}$ is poor and "small" when the fit is good. It is zero only if we are able to obtain $\mathbf{R} = \mathbf{FF}' + \mathbf{U}^2$ that exactly equals our sample $\mathbf{A}$, for then $\mathbf{AR}^{-1} = \mathbf{I}_n$ and it is easily seen that $\text{Tr}\ \{\mathbf{I}_n\} = n$ and $|\mathbf{I}| = 1$, so $\log\ |\mathbf{I}| = 0$. (The reader may recognize that the quantity λ is of the form $x - \log x - 1$. It can be shown that such a quantity is essentially positive, becoming zero at $x = 1$.)

The conditions for a minimum of λ with respect to $\mathbf{F}$ and $\mathbf{U}^2$ may be written as

$$\mathbf{R}^{-1}(\mathbf{R} - \mathbf{A})\mathbf{R}^{-1}\mathbf{F} = \mathbf{0} \tag{2.6.8}$$

and

$$\text{Diag}\ \{\mathbf{R}^{-1}(\mathbf{R} - \mathbf{A})\mathbf{R}^{-1}\} = \mathbf{0}. \tag{2.6.9}$$

Like the corresponding equations (2.6.3) and (2.6.4), these are simultaneous nonlinear equations that require a numerical algorithm for their solution. Again we can find numerical methods that yield a sequence of improved values of $\mathbf{U}^2$, for each of which we obtain a solution to (2.6.8) in terms of the principal components of a certain matrix (usually, of $\mathbf{U}^{-1}(\mathbf{A} - \mathbf{U}^2)\mathbf{U}^{-1}$, but this has certain disadvantages). We cannot, in this case, solve (2.6.9) in closed form for $\mathbf{U}^2$, given $\mathbf{F}$. Nevertheless, methods that assume (2.6.6) to be true for $\mathbf{F}$ other than the required minimizing value do work quite well.

Very interestingly, it can be shown that the LRC l, given in (2.6.7), can also be expressed as

$$\lambda = -N\ \log|\mathbf{R}_e| \tag{2.6.10}$$

where $\mathbf{R}_e$ is the *correlation* matrix (not the covariance) of the residuals. In trying to maximize the likelihood, assuming normality, we are trying to maximize the determinant of the residual correlation matrix.